ESSAYS ON ESTIMATING DISTRIBUTIONAL CONTINUOUS TREATMENT EFFECTS

by

Ying-Ying Lee

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

University of Wisconsin–Madison

2013

Date of final oral examination: 5/6/2013

The dissertation is approved by the following members of the Final Oral Committee:

Jack Porter, Professor of Economics, Chair

Bruce Hansen, Professor of Economics

Xiaoxia Shi, Assistant Professor of Economics

Christopher Tabor, Professor of Economics

Kjell Doksum, Senior Research Scientist, Statistics

Acknowledgments

I feel blessed writing this thank you note, which cannot be commensurate with what I've received. I am indebted to Jack Porter for his devotion to advise me. His trust and support encourages me to explore the unknown; his guidance leads me to rigorous and interesting research. Bruce Hansen has inspired me since the first-year econometrics course, and I am fortunate to be his assistant on teaching and research for this course. Bruce and Jack are my role models, in research and in life. Without them, this dissertation could not be competed and I cannot keep pursuing research. Xiaoxia Shi provided fresh perspective and valuable comments. Chris Taber gave me honest suggestions and sharp insight on the economic application. I am honored to have Kjell Doksum to complete my dream committee. I've also benefited from great teachers, Andres Aradillas-Lopez and Ken West. I'd like to thank Chung-Ming Kuan who brought me to the world of econometrics. I acknowledge Juan Villa for providing the data of the program Families in Action in Colombia in Chapter 1.

I've enjoyed learning with the econometrics group: Shengjie Hong, SeoJeong Lee, Jen-Che Liao, Chu-An Liu,, Nelson Ramirez-Rondan, Enrique Pinzon, Jing Tao, and Jin Yan. My office-mates, Michael Choi, YoungWook Lee, and Toshinori Onuma, create positive externality: intellectual conversation and simple laugh. Laura Dague, Cher Li, Chenyan Lu, Mai Seki, Cheng-Ying Yang, Chia-Chen Yang, and Yuan Yuan are dear girl friends. I'd like to thank them to be my family in Madison. I also thank my friends from Taiwan, Yuan Liu and Hsin-Ying Shen, for their distant care.

I dedicate this dissertation to my family for their unconditional love. My sister, Ming-Chih, always has belief in me. My dad, Chin-Tsan Li, encourages me to learn and fly as far as I can. From my mom, Man-Yun Yang, I see infinite love and strength.

DISCARD THIS PAGE

Contents

Page

Appendix

Page

DISCARD THIS PAGE

List of Tables

DISCARD THIS PAGE

List of Figures

Figure Page

Figure Page

Appendix

Figure

Abstract

The unconditional distribution of potential outcomes with continuous treatments and the quantile structural function in a nonseparable triangular model can both be expressed as a partial mean process with generated regressors. In Chapter 1, I propose a multi-step nonparametric kernel-based estimator for this partial mean process. A uniform expansion reveals the influence of estimating the generated regressors on the final estimator. In the case of continuous treatment effects, an unconfoundedness assumption leads to regression on the generalized propensity score (Hirano and Imbens, 2004), which serves as the generated regressor in the partial mean process. Analogous to the binary treatment effect case, my results suggest that the generalized propensity score reduces the dimension of nonparametric regression in estimation, but does not improve first-order asymptotic efficiency. Nonseparable triangular models commonly include a conditional independence assumption that yields a control function approach to deal with endogeneity (Imbens and Newey, 2009). In a preliminary step, the control variable is estimated nonparametrically as a generated regressor. My general partial mean process results can then be applied to provide the asymptotic distribution of the nonparametric estimator for the average and quantile structural functions. By extending my results to Hadamard-differentiable functionals of the partial mean process, I am able to provide the limit distribution for estimating common inequality measures and various distributional features of the outcome variable, such as the Gini coefficient. Monte Carlo results demonstrate the finite sample behavior of my estimator. In addition, a substantive empirical application using data from a Colombian conditional cash transfer program

illustrates the usefulness of the current findings for the estimation of continuous treatment effect models.

In the second chapter, I estimate the density-weighted Average Quantile Derivative (AQD), defined as the expectation of the partial derivative of the conditional quantile function (CQF) weighted by the density function of the covariates. The proposed estimator achieves root-n-consistency and asymptotic normality by a first-step nonparametric kernel estimation for the unknown functions and a second-step sample analogue of a full- mean. Therefore, the AQD summarizes the average marginal response of the covariates on the CQF and can be viewed as a nonparametric quantile regression coefficient. Similar to the widely studied average derivative in mean regression, the AQD identifies the coefficients up to scale in semiparametric single-index and partial linear models. For the nonparametric nonseparable structural model, the derivative of the CQF identifies the structural derivative, under the conditional independence assumption in Hoderlein and Mammen (2007).

In the third chapter, I allow for misspecification in the linear conditional quantile function (CQF) and calculate the semiparametric efficiency bound for the quantile regression (QR) parameter, the best linear predictor for a response variable under the asymmetric check loss function. As a result, the QR estimator developed by Koenker and Bassett (1978) semiparametrically efficiently estimates a pseudo-true parameter that produces parsimonious descriptive statistics for the CQF. The linear quantile projection model can be understood by the orthogonality condition of the covariates and the distribution error (i.e., the deviation of the true conditional distribution function, evaluated at the linearly approximated quantile, from the true probability). A novel observation of this article is that the QR parameter is the unique fixed point to the iterated minimization of the mean-squared distribution error, inversely weighted by the conditional density function. My result suggests that the distribution errors are larger at points with higher conditional densities, while Angrist et al. (2006) find that QR approximates the CQF more accurately at such points with more observations. These approximation features and parallel properties with ordinary least squares reinforce the scholarly understanding of QR.

Chapter 1

Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models

1.1 Introduction

I study nonparametric estimation of continuous treatment effect models when the treatments are randomly selected or assigned conditional on either observables or on unobserved control functions. [1] Two key features of these models are non-separability in the unobservables and heterogeneity in treatment intensity effects. I focus on identification and estimation of objects such as the unconditional distribution of potential outcomes and the quantile structural function in nonseparable triangular models. Such objects can be expressed as functionals of a partial mean process with generated regressors. I propose a fully nonparametric multi-step estimator for this partial mean process and show how the estimation error associated with the generated regressor affects the limit distribution of the estimator.

The proposed methods capture heterogeneous treatment intensity effects by estimating an array of distributional structural features that can be applied to a variety of economic questions. For example, when evaluating a social program, researchers might be interested in how the length of exposure to the program affects the entire distribution of wages. I extend my method to include inference on smooth functionals of the outcome distribution process. In this example, a researcher could consider how inequality responds to the length

[1] Matzkin (2007), "A control function is a function of observable variables such that conditioning on its value purges any statistical dependence that may exist between the observable and unobservable explanatory variables in an original model."

of exposure to an anti-poverty program by tracing out the Gini coefficient of the wage distribution by time in the program. In demand analysis, the Engel curve can be analyzed by the average or quantile structural functions in triangular simultaneous equations models (Imbens and Newey (2009)). I can also consider the counterfactual effects of either a change in the distribution of a set of covariates or a change in the relationship of the covariates with the outcome, as in the counterfactual analysis for a discrete treatment or policy variable in Chernozhukov et al. (2013).

Let $Y(t)$ denote the potential outcome corresponding to the level of treatment intensity t. The key causal object of interest in this paper is the unconditional[2] distribution of the potential outcome with continuous treatments. It is *unconditional* in the sense that other covariates are being integrated out, while the potential outcome $Y(t)$ framework provides the flexibility to reveal local information by fixing relevant variables at a treatment level t. The results here could be straightforwardly generalized to consider the conditional potential outcome distribution where the conditioning set consists of exogenous observables or discrete covariates. Alternatively, the object of interest is the distribution of the outcome for some fixed values of the endogenous variables of interest in the triangular simultaneous equations models in Imbens and Newey (2009). White and Chalak (2013) formally discuss the equivalence of the treatment effect models of the potential outcome framework and the structural triangular system.

The treatment variables are assumed to be exogenous conditional on observables or unobserved control functions. Together with a common-support assumption, the distributional causal effects are identified in terms of functionals of partial means of weighted conditional cumulative distribution functions (cdf) of the observed outcome $Y = Y(T)$ given treatment T and generated regressors Λ. The main contribution of this paper is a fully nonparametric

[2]The unconditional distribution of $Y(t)$ is often known as the marginal distribution. I use "unconditional" instead of "marginal" because I use the term "marginal effect" to refer to the impact of infinitesimal changes in the continuous treatment.

multi-step estimation procedure for a partial mean process with generated regressors,

$$\Big\{\ t \to \theta_t(y|\Lambda, W) \equiv E\Big[E\Big[\mathbf{1}_{\{Y\leq y\}}\Big|T=t, \Lambda=\Lambda(S)\Big]\cdot W(S_w)\Big] : y \in \mathcal{Y}\ \Big\} \tag{1.1}$$

where generated regressors $\Lambda(S)$ and a weight function $W(S_w)$ are measurable functions of sets of observables, S and S_w, and can be estimated in the first step. The observable covariates sets S and S_w are not restricted to be a subset of one another. Depending on the economic application, S_w and S usually overlap. The inner conditional expectation is simply a conditional cdf, $F_{Y|T\Lambda}(y|t,\Lambda)$, which is estimated nonparametrically in a second step. The last step of estimation then averages out over the observables (S, S_w). Because the continuous treatment variables T are fixed at level t, $F_{Y|T\Lambda}(y|t,\Lambda)$ contains more arguments than being averaged over in the third step. This partial mean structure implies that the convergence rate is slower than root-n, as typically found in discrete treatment cases. Newey (1994b) introduces the partial mean and its applications, such as consumer surplus estimation and additive nonparametric models. This paper builds on and extends the partial mean literature in two ways: First, the dependent variable is $\{\mathbf{1}_{\{Y\leq y\}} : y \in \mathcal{Y}\}$ a process indexed by the threshold y, making it possible to estimate the whole distribution of Y and also the Hadamard-differentiable functionals of that distribution. Second, the regressors can be unobserved and treated as generated regressors to be estimated parametrically or nonparametrically in the first step. Next I consider the role played by the different individual components that make up the object of interest in the expression (1.1).

Distribution Process. I obtain weak convergence of the partial sum of the weighted conditional cdf process, allowing for regressors to be nonparametrically generated. The multiplier central limit theorem is valid for uniform inference, which enables functional hypotheses tests for the whole distribution, such as tests for no effect or stochastic dominance.

By extending the results to the Hadamard-differentiable functionals of the partial mean process, I am able to provide the limit distribution and uniform inference method for estimating common inequality measures and various distributional structural features; for example, quantile functions, the Lorenz curves, and the Gini coefficients. (Bhattacharya (2007); Rothe

(2010); Firpo and Pinto (2011); Donald et al. (2012); Chernozhukov et al. (2013). This class of functionals also covers a wide class of regression functions with generated regressors, which have been of interest in the econometrics and statistics literature. For example, the unconditional mean $E[Y(t)]$ is the average structural function or the dose-response function (Blundell and Powell (2003) and Flores (2007)). The local average response or the marginal mean treatment effect on the treated $\nabla_t E\big[Y(t)\big|T=\bar{t}\big]\big|_{t=\bar{t}}$ is the average effect for those currently choosing treatment level $\bar{t}$ of an incremental increase in the treatment, holding their other observables and unobservables fixed at baseline values (Altonji and Matzkin (2005) and Florens et al. (2008)[3]). The unconditional quantile function $F^{-1}_{Y(t)}(\tau)$ defines the quantile structural function or the quantile dose-response function (Imbens and Newey (2009)). The difference between two treatment levels is the unconditional quantile treatment effect. I derive the weak convergence of estimating the entire quantile process in Section 1.6.1.2.

Generated Regressors $\Lambda(S)$. I derive a uniform stochastic expansion of the multi-step estimator characterizing the influence of estimating the generated regressors on the final estimator. The explicit stochastic expansion serves as the cornerstone to establish weak convergence for the estimated partial mean process with generated regressors of general function forms. I study two important examples for generated regressors in detail. First, a control variable can be included in the conditioning variables as in the triangular simultaneous equations models in Newey et al. (1999) and Imbens and Newey (2009). When the control variable contributes to a full mean, I show the variation from estimating this generated regressor converges at $\sqrt{n}$-rate and is first-order asymptotically ignorable. The second example is the generalized propensity score (GPS), defined as the conditional density function of treatment T given observables X. Under the unconfoundedness assumption, the GPS is known to reduce dimensionality in the second-step regression (Hirano and Imbens (2004)). I show that for estimating the overall distribution $F_{Y(t)}(y)$, the GPS does not result

[3]Florens et al. (2008) impose a stochastic polynomial assumption on the heterogeneous effects and use a control function approach to obtain identification, rather than assuming common support, as I do in this paper.

in an efficiency gain, over controlling directly for the whole set of the observables. Also, regressing on the true or parametrically estimated GPS is less efficient than regressing on the nonparametrically estimated GPS. These results parallel the binary treatment case: a recent finding for the nonparametric regression on the propensity score in Hahn and Ridder (2013) and the propensity-score-weighting estimators in Hahn (1998) and Hirano et al. (2003). See Section 1.4.2.3 for details.

A uniform expansion of my multi-step kernel-based estimator reveals the influence of estimating the generated regressors on the final estimator. The technical challenge is that the estimated generated regressor plays two roles: the regressor for the second-step regression and the argument being averaged over in the third-step partial mean. I use a stochastic equicontinuity argument from empirical process theory, following the recent literature on nonparametrically estimated generated regressors, for example, Mammen et al. (2012a), Mammen et al. (2012b), Escanciano et al. (2012), Hahn and Ridder (2013), and Song (2008). I contribute to this literature by deriving the influence of the estimation error of the generated regressor on the partial-mean process, without artificially assuming that the contribution of the first-step estimation error converges at a faster rate. I also find the trade-off between the complexity and accuracy assumptions as in Mammen et al. (2012a) and Escanciano et al. (2012): if the control function is smoother, i.e., it belongs to a less complex function space, then estimation of the generated regressor needs to converge at a faster rate.

My work is most closely related to the work of Mammen et al. (2012a), which develops a stochastic expansion based on a kernel estimation. Mammen et al. (2012a) focus on conditional mean regression with generated regressors, while I further study the partial mean of this conditional regression. Hahn and Ridder (2013) use Newey (1994a) path-derivative method to derive the asymptotic variance of multi-step semiparametric estimators for a full mean involving a generated regressor. Hahn and Ridder (2013) and Mammen et al. (2012b) derive the influence function for estimating the binary treatment effect by regressing on an estimated propensity score. For the continuous treatment case, I show the limit theory for regressing on the estimated generalized propensity score. Escanciano et al. (2012) introduce a

uniform expansion of a full mean of weighted kernel-based regression residuals. The authors' stochastic expansion is uniform with respect to the weights, bandwidth for the kernel, and generated regressors.

Weight $W(S_w)$. The counterfactual distribution of $Y(t)$ for those currently being treated or choosing $\bar{t}$, $F_{Y(t)|T}(y|\bar{t})$, can be identified by (1.1) using the weight $W(\Lambda) = f_{T|\Lambda}(\bar{t}|\Lambda)/f_T(\bar{t})$, where $f_{T|\Lambda}(\bar{t}|\Lambda)$ is the conditional density of T given Λ evaluated at $\bar{t}$. This is in the spirit of Horvitz and Thompson (1952) and DiNardo et al. (1996) to reweight the observations using the propensity score. The sampling variation resulting from estimating this weight $W(S_w) = f_{T|S_w}(\bar{t}|S_w)/f_T(\bar{t})$ and using it to estimate the weighted partial mean process in (1.1) is taken into account. [4]

I can also study decomposition and policy analysis by estimating the treatment effect on the treated under unconfoundedness. The statistical object $\theta_t\big(y\big|X, W(X) = f_{T|X}(\bar{t}|X)/f_T(\bar{t})\big)$ $= E\big[F_{Y|TX}(y|t,X)\big|T = \bar{t}\big]$ can be interpreted as a counterfactual distribution of either a change in the conditional distribution of the outcome given the characteristics or a postulated distribution of the characteristics. For example, I could assess what wage distribution would have prevailed if, conditional on the same observable characteristics X, individuals who have been participating in an anti-poverty program for $\bar{t}$ years had been paid according to the wage schedule of those who have been participating for t years ($f_{Y|XT}(y|X,t)$). This decomposition analysis might reveal discrimination or stigma based on the length of time in the program. My work is an extension of the counterfactual analysis in Chernozhukov et al. (2013), in which the treated group can be viewed as defined by a discrete treatment, such as gender, races, or time periods. In this paper, the treatment or policy variable is considered

[4]The weight will include a fixed trimming function, where the density of the conditioning variables are bounded away from zero, as in Newey (1994b). In principle, an estimated trimming function could be incorporated into my framework and considered in the asymptotic results. I do not pursue this extension in this paper. The fixed trimming choice allows me to focus on the technical issues associated with estimating the generated regressor and the whole distribution processes. In fairness, the choice of fixed trimming function can affect the interpretation of the estimands considered. The subpopulation is selected such that the observables do not take extreme values.

continuous, for example, teacher quality, subsidy, or cigarette consumption. Section 1.2.2 contains a more detailed discussion.

Treatment Intensity Effects. In contrast to the vast binary treatment effect literature that captures only the effect of participating in a program, econometrics methods for treatment intensity effects are less developed. Cattaneo (2010) studies the efficient estimation of multi-valued treatment effects, in which identification and estimation are extended from the binary case under the unconfoundedness assumption. To allow for selection on unobservables, the instrumental variable results in the binary case do not easily extend to the endogenous multi-valued treatment case; see the review paper of Imbens and Wooldridge (2009) for further discussion of this issue. In contrast, models of endogenous continuous treatment can borrow a control function approach from triangular simultaneous equations models to account for endogeneity. My estimator covers this case when the control variable is estimated in a preliminary step.

There is a growing empirical literature analyzing continuous treatment effects. Hirano and Imbens (2004) study the effect of unearned income, measured by the amount of a lottery prize, on subsequent labor earnings. With regard to program evaluation, the duration of exposure to the programs is often considered a continuous treatment. Behrman et al. (2004) use a matching-typed estimator to evaluate a Bolivian preschool program. The semiparametric estimation method developed by Hirano and Imbens (2004) has been used to analyze the Cash Transfers programs (Progresa/Oportunidades) in Mexico (Ibarraran and Villa (2010)) and the South African Child Support Grant (Agüero et al. (2010)). Flores et al. (2012) and Kluve et al. (2012) use this method to study job training programs. Although regressing on the estimated GPS is common practice in empirical analyses, to the best of my knowledge, this paper is the first presentation of a complete limit theory of nonparametric regression on the estimated GPS. Flores (2007) derives the limit theory for nonparametric estimation based on regression on the observables to estimate the dose-response function and the location and size of its maximization. I extend the continuous treatment literature by moving

beyond the mean to the whole distributions of the potential outcomes. In addition, I allow selection on unobservables by importing the control variable approach from the literature on triangular simultaneous equations models.

The rest of the paper is organized as follows: Section 2 introduces the setup and parameters of interest. Table 1.1 summarizes the identification results and asymptotic theory for estimating the causal outcome distributions. Section 3 describes three-step nonparametric kernel estimation for the weighted partial mean process with generated regressors (1.1). Section 4 outlines the main asymptotic theorems when the weights are observed. A uniform stochastic expansion characterizes the influence of estimating the generate regressor on the multi-step estimator. I focus on two economic examples for the generated regressors: the control variables in the simultaneous equations models in Newey et al. (1999) and Imbens and Newey (2009) and the generalized propensity score in Hirano and Imbens (2004). Section 1.5 presents the limit theory for estimating the weight function for the treated effect on the treated. Section 1.6 presents the limit theories for treatment/policy effects using the functional delta method for the Hadamard-differentiable policy functionals. I explicitly carry out the limit theory for estimating the mean and quantile processes. The simulation in Section 7 shows that the proposed estimators work in finite samples, comparing with the parametric method in Hirano and Imbens (2004). The empirical example is part of a joint project with Juan Villa in which we evaluate a Colombian conditional cash transfer program. The proofs are in the Appendix.

Notations. $\perp$ denotes independence. $\|\cdot\|_\infty$ is the sup-norm, i.e., $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$, where $\mathcal{X}$ is the support of X. $y_1 \wedge y_2 \equiv \min\{y_1, y_2\}$. *a.s.* is the shorthand of *almost surely*. Let d_u denote the dimension of a vector u. $(u)_{min}$ denotes the smallest element and $(u)_{max}$ denotes the biggest element of the vector u. Let $\underline{\alpha}$ be the greatest integer strictly smaller than α.

1.2 Distributional Features of Potential Outcomes

This section introduces the continuous treatment effect model and the causal objects of interest. For completeness, I present results for the regression-type and propensity-score weighting identifications of the distributions of the potential outcome for the whole population and the treated subpopulation (those who have chosen a certain treatment level). Table 1.1 summarizes the identification findings based on the weighted partial mean process with generated regressors in (1.1), specifically for the treatment effect model under unconfoundedness and the nonseparable simultaneous equations models in Newey et al. (1999) and Imbens and Newey (2009).

1.2.1 Potential Outcome Framework

Let $Y(t)$ denote the potential outcome corresponding to the level of treatment intensity t. The continuous treatment vector T takes values on a compact set $\mathcal{T} \subset \mathbb{R}^{d_t}$. The observed outcome $Y = Y(T)$ is one of the potential outcomes $\{Y(t)\}_{t\in\mathcal{T}}$. The following identification and estimation results are written to also include discrete treatments, which will be discussed in the following sections. The treatment effect model is equivalent to a nonseparable outcome with a general disturbance, where the outcome equation is $Y = \phi(T, X, \epsilon)$. The structural equation for the outcome is assumed not to change when a policy intervention determines the treatment T. No functional form assumption is imposed on the general disturbances ϵ, like monotonicity, dimensionality, or separability. Rank invariance is assumed, so that the realizations of ϵ do not change when T is counterfactually manipulated. The observed characteristics X could include endogenous pretreatment variables.

The stable unit treatment value assumption (Rubin, 1980) is inherently assumed: the outcome for one unit is independent of potential treatment status of another unit given the observed covariates. Social interaction, general equilibrium effects, and peer effects are not considered.

1.2.1.1 Treatment Effects and Treatment Effects on the Treated

The cumulative distribution function (cdf) of the potential outcome $F_{Y(t)}(y)$ is the unconditional distribution of the outcome if, hypothetically, the entire population had been assigned to the treatment level t. The cdf of $Y(t)$ for those who have chosen their treatment level $\bar{t}$ is defined by $F_{Y(t)|T}(y|\bar{t}) = E\big[\mathbf{1}_{\{Y(t)\leq y\}}\big|T=\bar{t}\big]$. I study the *overall treatment effect* based on $F_{Y(t)}(\cdot)$ and the *treatment effect on the treated* by $F_{Y(t)|T}(\cdot|\bar{t})$, for $t, \bar{t} \in \mathcal{T}$. Note the causal outcome distributions are processes indexed by threshold value $y \in \mathcal{Y}$, for some fixed treatment levels of interest $t, \bar{t} \in \mathcal{T}$. In the following Assumptions, Lemmas, and Theorems throughout this paper, I will consider some fixed treatment levels belonging to $\mathcal{T}$, without repeating this statement.

In practice, an array of estimands are often of interest based on these causal outcome distributions. I consider a general class of functionals Γ on $F_{Y(t)}(\cdot)$ and $F_{Y(t)|T}(\cdot|\bar{t})$. For example, if interest centers on the quantile treatment effect (QTE), I let Γ be the quantile operator and the QTE corresponding to a change from t to $\bar{t}$ is $\Gamma(F_{Y(\bar{t})}) - \Gamma(F_{Y(t)})$. Similarly, the QTE on the treated $\bar{t}$ is $\Gamma(F_{Y(\bar{t})|T}(y|t)) - \Gamma(F_{Y(t)|T}(y|t))$. If interest is on the mean treatment effect, then let Γ be the mean operator. Other inequality measures are also applicable, such as the coefficient of variation, the interquantile range, the Theil index, the Gini coefficient, the Lorenz curve, as discussed in Rothe (2010), Firpo and Pinto (2011), Chernozhukov et al. (2013) for a discrete treatment.

1.2.1.2 Identification

I use the conditional independence and common support assumptions to show that the partial mean process with generated regressors (1.1) identifies the causal outcome distributions $F_{Y(t)}(\cdot)$ and $F_{Y(t)|T}(\cdot|\bar{t})$. Define the control functions $\Lambda(S)$ to be a vector of measurable functions of S, a subvector of observables (T, X, Z_T), where Z_T is an excluded exogenous instrumental vector for T.

Assumption 1.1 (CIA)**.** *T and ϵ are independent conditional on $\Lambda(S)$. Or for any $t \in \mathcal{T}$, the potential outcome $Y(t)$ is independent of the treatment T, given $\Lambda(S)$.*

I focus on the following types of conditioning variables $\Lambda(S)$. Under unconfoundedness or selection on observables X, Assumption 1.1 is satisfied by $\Lambda(S) = X$ or $\Lambda(S) = f_{T|X}(t|X)$ the generalized propensity score (GPS) in Hirano and Imbens (2004). When unconfoundedness is violated, one approach to satisfying Assumption 1.1 is through control variables as in the triangular simultaneous equations model. For example, Imbens and Newey (2009) show the conditional distribution function of the endogenous variable given the instrumental variables is a control variable $V(T, Z) = F_{T|Z}(T|Z)$, where $Z \subset (X, Z_T)$. So in this case, the conditioning variables are $\Lambda(S) = (X', V'(T, Z))'$ and $S = (X', T', Z')'$.

The conditional distribution of the potential outcome $Y(t)$ given Λ is identified by Assumption 1.1,

$$F_{Y(t)|\Lambda}(y|\Lambda) \equiv E[\mathbf{1}_{\{Y(t)\leq y\}}|\Lambda] = E[\mathbf{1}_{\{Y(t)\leq y\}}|T = \bar{t}, \Lambda] = E[\mathbf{1}_{\{Y\leq y\}}|T = t, \Lambda] \equiv F_{Y|T\Lambda}(y|t, \Lambda) \tag{1.2}$$

$\forall \bar{t} \in \mathcal{T}$. That is, conditional on the control function Λ, the distribution of the potential wage for choosing treatment intensity t is invariant of the current treatment intensity $\bar{t}$, $F_{Y(t)|T\Lambda}(y|\bar{t}, \Lambda) = F_{Y|T\Lambda}(y|t, \Lambda)$.

The following common support assumption, also known as the overlapping, assumes that there is a positive probability of observing the treatment levels in some interval of interest with the same characteristics Λ.

Assumption 1.2 (Common Support)**.** *The support of Λ conditional on $T = t$ equals the support of Λ.*

For discrete treatments, the common-support assumption implies the propensity score $Pr(T = t|\Lambda)$ cannot be exactly zero or one. Although I focus on continuous treatments, the identification results generally cover discrete treatments.

By Assumptions 1.1 and 1.2, and following equation (1.2), the partial mean process with generated regressor in (1.1)

$$\theta_t(y|\Lambda, W) \equiv E\big[F_{Y|T\Lambda}(y|t,\Lambda)\cdot W(\Lambda)\big]$$

identifies $F_{Y(t)}(y)$ with $W = 1$ and identifies $F_{Y(t)|T}(y|\bar{t})$ with $W(\Lambda) = f_{T|\Lambda}(\bar{t}|\Lambda)/f_T(\bar{t})$. The identification for the overall cdf $F_{Y(t)}$ has been shown in Theorem 3 in Imbens and Newey (2009). The discrete treatment case is well-studied $E[Y(0)|T=1] = E\big[E[Y|T=0,\Lambda]\cdot Pr(T=1|\Lambda)/Pr(T=1)\big]$, where the conditional mean $E[Y|T=0,\Lambda]$ can be estimated by nonparametric regression (Heckman et al. (1998); Hahn and Ridder (2013)). When the kernel method is used with a fixed bandwidth, it is known as the matching estimator Abadie and Imbens (2006).

The nonparametric estimation in this paper is based on this regression-type identification. However, the estimation of $W(\Lambda)$ is complicated when the regressor Λ is estimated. So I propose another version of identification for $F_{Y(t)|T}(y|\bar{t})$ in the following Lemma, where the weight does not depend on the generated regressor.

Lemma 1.1. *Suppose the control function to be $\Lambda = (X', V')'$, where the control variable $V = V(T, Z)$, a measurable function of $(T', Z')'$. Suppose Assumptions 1.1 and 1.2 hold. Then the cdf of $Y(t)$ for the treated $\bar{t}$ is identified by $F_{Y(t)|T}(y|\bar{t}) = E\Big[F_{Y|TXV}\big(y\big|t, X, V(\bar{t}, Z)\big)\cdot W(X,Z)\Big]$, where the weight $W(X,Z) = f_{T|XZ}(\bar{t}|X,Z)/f_T(\bar{t})$.*

Proof:

$$\begin{aligned}
F_{Y(t)|T}(y|\bar{t}) &= E\big[F_{Y(t)|TXV}(y|\bar{t},X,V)\big|T=\bar{t}\big] = E\big[F_{Y|TXV}(y|t,X,V)\big|T=\bar{t}\big] \\
&= E\Big[E\big[F_{Y|TXV}(y|t,X,V)\big|T=\bar{t},Z\big]\Big|T=\bar{t}\Big] \\
&= E\big[F_{Y|TXV}(y|t,X,V(\bar{t},Z))\big|T=\bar{t}\big] = E\Big[F_{Y|TXV}\big(y\big|t,X,V(\bar{t},Z)\big)\frac{f_{T|Z}(\bar{t}|X,Z)}{f_T(\bar{t})}\Big].
\end{aligned}$$

The first and the third equalities are by the law of iterated expectations. The second equality is by the CIA. The first expectation of the last line is taken by the conditional distribution of (X, Z) given $T = \bar{t}$. The last expectation is taken by the marginal distribution of (X, Z). □

The discussion above has covered an array of identified objects with causal interpretations that can be included as special cases of a partial mean process with generated regressors in (1.1). Below is a table of the objects considered in this work along with a roadmap for the results on asymptotic properties of corresponding estimators.

	Identification	Asymptotics
	Selection on Observables X	
$F_{Y(t)}$	$\theta_t\big(y\big\vert X, W=1\big) = E\big[F_{Y\vert TX}(y\vert t, X)\big]$	Theorem 1.1
	$\theta_t\big(y\big\vert V, W=1\big) = E\big[F_{Y\vert TV}(y\vert t, V)\big]$, GPS $V = f_{T\vert X}(t\vert X)$	Corollary 1.2
$F_{Y(t)\vert T}(y\vert\bar{t})$	$\theta_t\big(y\big\vert X, W(X)\big) = E\big[F_{Y\vert TX}(y\vert t, X)\cdot W(X)\big]$ ‡	Theorem 1.3-1
	$\theta_t\big(y\big\vert V, W(X)\big) = E\big[F_{Y\vert TV}(y\vert t, V)\cdot W(X)\big]$ ‡	Theorem 1.3-2
	Control Variable $V = V(T, Z)$ †	
$F_{Y(t)}$	$\theta_t\big(y\big\vert (X, V), W=1\big) = E\big[F_{Y\vert TXV}(y\vert t, X, V)\big]$	Corollary 1.1
$F_{Y(t)\vert T}(y\vert\bar{t})$	$\theta_t\big(y\big\vert (X, V=V_{\bar{t}}), W(X,Z)\big) = E\big[F_{Y\vert TXV}(y\vert t, X, V(\bar{t}, Z))W(X,Z)\big]$ ‡	Theorem 1.3-3

Table 1.1 Summary of Results

† The control variables are constructed for $V(T,Z) = T - E[T\vert Z]$ in Newey et al. (1999) and $V(T,Z) = F_{T\vert Z}(T\vert Z)$ in Imbens and Newey (2009) in Section 1.4.2.2.

‡ $W(S_w) = f_{T\vert S_w}(\bar{t}\vert S_w)/f_T(\bar{t})$

Remark (Propensity-Score Weighting)

I generalize identification by propensity-score weighting from the binary treatment effect literature to the case of continuous treatments. Propensity-score weighting identification for continuous treatment variables relies on the introduction of a kernel function $K_h(T-t) \equiv \frac{1}{h^{d_t}}\Pi_{l=1}^{d_t} k(\frac{T_l - t_l}{h})$, where k is any conventional kernel. The identification argument will depend on the smoothness in $F_{Y\vert T\Lambda}(y\vert t, \Lambda)$ and $f_{T\vert\Lambda}(t\vert\Lambda)$ matching the order of the kernel k. Let r

denote the order of the kernel k, and assume $F_{Y|T\Lambda}(y|t,\Lambda)$ and $f_{T|\Lambda}(t|\Lambda)$ are rth-order continuously differentiable in t with uniformly bounded derivatives. By a calculation involving a Taylor expansion of the kernel,

$$F_{Y|T\Lambda}(y|t,\Lambda) = \lim_{h\to 0} \frac{E[\mathbf{1}_{\{Y\le y\}} K_h(T-t)|\Lambda]}{E[K_h(T-t)|\Lambda]} = \lim_{h\to 0} \frac{E[\mathbf{1}_{\{Y\le y\}} K_h(T-t)|\Lambda]}{f_{T|\Lambda}(t|\Lambda)}.$$

Together with (1.2) and the law of iterated expectations, the propensity-score weighting identification is

$$E[F_{Y|T\Lambda}(y|t,\Lambda)W(\Lambda)] = \lim_{h\to 0} E\Big[\frac{\mathbf{1}_{\{Y\le y\}} K_h(T-t)}{E[K_h(T-t)|\Lambda]} W(\Lambda)\Big] = \lim_{h\to 0} E\Big[\frac{\mathbf{1}_{\{Y\le y\}} K_h(T-t)}{f_{T|\Lambda}(t|\Lambda)} W(\Lambda)\Big]. \tag{1.3}$$

Flores et al. (2012) estimate the continuous treatment effect by (1.1) and (1.3) nonparametrically with a parametric generalized propensity score without providing a limit theory. I do not exploit estimation based on this propensity-score weighting identification. For a discrete treatment, the kernel function degenerates to an indicator function, $E\Big[\frac{\mathbf{1}_{\{Y\le y\}}\mathbf{1}_{\{T=t\}}}{P(T=t|\Lambda)} W(\Lambda)\Big]$. Propensity-score weighting estimation for the discrete treatment is well-studied in the treatment effect literature, e.g., Cattaneo (2010), Hirano et al. (2003), Firpo and Pinto (2011).

1.2.2 Counterfactual Effects and Treatment Effects on the Treated

In the conventional treatment effect literature, the treatment effect on the treated $E[Y(t)-Y(\bar{t})|T=\bar{t}]$ is the effect of hypothetically assigning a different treatment level t to the subpopulation whose current treatment is $\bar{t}$. Assuming unconfoundedness, the treatment effects on the treated can be interpreted as the *counterfactual effects* of a policy intervention by shifting exogenously to a conditional distribution of another potential outcomes given the covariates or a postulated distribution of the covariates. Chernozhukov et al. (2013) study counterfactual effects for the multivalued policy intervention or a counterfactual change in economic conditions. They divide the population into subpopulations by a multivalued variable, e.g., gender, races, time periods. I consider the economic conditions or policy generated from a continuous variable.

Consider an example of program evaluation when the outcome of interest Y is the wage and the continuous treatment T is the length of exposure to the program. Define the $\bar{t}$-group by those individuals who have been in the program for $\bar{t}$ years, i.e., the sub-population of $T = \bar{t}$. The observed wage distribution of the $\bar{t}$-group is

$$F_{Y|T}(y|\bar{t}) = \int F_{Y|XT}(y|x,\bar{t})\ dF_{X|T}(x|\bar{t}).$$

Consider a *wage schedule* to be a map from characteristics X to a wage distribution function. Then $\{x \mapsto F_{Y|XT}(y|x,\bar{t}) : y \in \mathcal{Y}\}$ can be seen as the observed wage schedule for the $\bar{t}$-group given characteristics X. $F_{X|T}(x|\bar{t})$ is the status-quo characteristics distribution for the $\bar{t}$-group. Following Chernozhukov et al. (2013) , the differences in wage distributions among sub-populations with different length of exposure $F_{Y|T}(y|t) - F_{Y|T}(y|\bar{t})$ can be decomposed into a *structure effect* and a *composition effect* in the following sense,

$$\begin{aligned}
&F_{Y|T}(y|t) - F_{Y|T}(y|\bar{t}) \\
&= -\int \Big(\boldsymbol{F_{Y|XT}(y|x,\bar{t})} - F_{Y|XT}(y|x,t)\Big)\ dF_{X|T}(x|t) \qquad (1.4)\\
&\quad + \int F_{Y|XT}(y|x,\bar{t})\ d\Big(\boldsymbol{F_{X|T}(x|t)} - F_{X|T}(x|\bar{t})\Big). \qquad (1.5)
\end{aligned}$$

The first term in (1.4) $\int \boldsymbol{F_{Y|XT}(y|x,\bar{t})}\ dF_{X|T}(x|t) = \theta_{\bar{t}}\big(y\big|X, W(X) = f_{T|X}(\bar{t}|X)/f_T(\bar{t})\big)$ is the counterfactual wage distribution that would have prevailed for the t-group if they faced the $\bar{t}$-group's wage schedule $\{x \mapsto F_{Y|XT}(y|x,\bar{t}) : y \in \mathcal{Y}\}$. Therefore, (1.4) represents the structure effect or the *discrimination effect* in Chernozhukov et al. (2013), arising due to pay difference among these sub-populations with the same characteristics. This can be a measure of discrimination based on the length of exposure to the program.

On the other hand, $\theta_{\bar{t}}\big(y\big|X, W(X) = f_{T|X}(\bar{t}|X)/f_T(\bar{t})\big)$ also in the first term in (1.5) can be interpreted as the counterfactual wage distribution of the $\bar{t}$-group if they had the t-group's characteristics distribution $F_{X|T}(x|t)$. So (1.5) is the composition effect, arising due to differences in characteristics among these sub-populations. Taking teacher quality as a continuous treatment, the decomposition effect might answer the question: if Teacher-t's

students had the same characteristics distribution as Teacher-$\bar{t}$'s ($F_{X|T}(X|\bar{t})$), how would the wage distribution change for Teacher-t's students?

These two counterfactual effects are well defined statistical parameters. Assuming unconfoundedness, this descriptive decomposition analysis has causal interpretation in the sense that the structure effect in (1.4), which is the counterfactual effect of changing the conditional distribution, is the treatment effect on the treated $F_{Y(\bar{t})|T}(y|t) - F_{Y(t)|T}(y|t)$. And the composition effect in (1.5), which is the counterfactual effect of changing the covariate distribution, equals $F_{Y(\bar{t})|T}(y|t) - F_{Y(\bar{t})|T}(y|\bar{t})$. Therefore, the estimation procedure and limit theory in this paper apply to the decomposition analysis. Even when the unconfoundedness assumption does not hold, the decomposition analysis is still valid.

1.3 Estimation

This section introduces a general procedure to estimate the process in (1.1) $\theta_t(y|\Lambda, W) = E\Big[E\Big[\mathbf{1}_{\{Y\leq y\}}\Big|T = t, \Lambda = \Lambda(S)\Big] \cdot W(S_w)\Big]$. I will discuss specific estimators for each economic examples in the later sections. I find that the estimation approach outlined below has different properties depending on the details of implementation corresponding to each estimand. As a result, different asymptotic distribution results are proceeded for the different versions of this general estimation approach described below. My estimator is straightforward and involves three steps:

1. (Generated Regressors)

   The generated regressor $\Lambda(S)$ can be estimated parametrically or nonparametrically, as long as its uniform convergence rate satisfies certain conditions, specified in the next section. If $\Lambda(S)$ is estimated as a nonparametric regression by a kernel method, let the bandwidth be h_1, the order of the kernel be r_1, and the dimension of the regressors be d_1.

2. (Regression)

The second step is the nonparametric regression of the indicator function $\mathbf{1}_{\{Y\leq y\}}$ on $\hat{\Lambda}(S)$ and evaluated at λ, i.e.,

$$\hat{F}_{Y|T\hat{\Lambda}}(y|t,\lambda)=\frac{\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{Y_j\leq y\}}K_h(T_j-t)K_h(\hat{\Lambda}(S_{vj})-\lambda)}{\frac{1}{n}\sum_{j=1}^{n}K_h(T_j-t)K_h(\hat{\Lambda}(S_{vj})-\lambda)}\equiv\frac{\hat{g}_{YT\hat{\Lambda}}(y,t,\lambda)}{\hat{f}_{T\hat{\Lambda}}(t,\lambda)}.$$

The product kernel is defined as $K_h(u)\equiv h^{-d_u}\Pi_{l=1}^{d_u}k\big(\frac{u_l}{h}\big)$, where $h=h_2$ is the bandwidth assumed the same for all the elements of the vector u for simplicity, and k is the r_2-ordered kernel function satisfying the following Assumption 1.4. Let the dimension of the regressors at this step be $d_2=d_t+d_\lambda$.

3. (Partial Sum)

   The third step is the partial sum, fixing the treatment variable T at level t, i.e., $\hat{\theta}_t(y|\hat{\Lambda},W)=\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y|T\hat{\Lambda}}(y|t,\hat{\Lambda}(S_{vi}))\cdot W(S_{wi})$.

- (Weight)

  The weight is a measurable function of the observables S_w and can be estimated by $\hat{W}(S_{wi})$, then $\hat{\theta}_t(y|\hat{\Lambda},\hat{W})=\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y|T\hat{\Lambda}}(y|t,\hat{\Lambda}(S_{vi}))\cdot\hat{W}(S_{wi})$.

When the generated regressor $\Lambda(S)$ is observed, the estimator is simplified to the second and third steps as the partial mean in Newey (1994b). [5]

In Sections 1.4 and 1.5, I will present results for estimators using the approach outlined above. The following assumptions will be maintained on the data generating process and the kernel used in the nonparametric regression step described above.

Assumption 1.3 (Smoothness). *(i) The data $\{Y_i,T_i,X_i,Z_{Ti}\}$, $i=1,...,n$, is i.i.d.*

(ii) $\Lambda(S)$ is a vector of measurable functions of S, a subverter of $\{T,X,Z_T\}$. The support of $\Lambda(S)$, $\mathbf{\Lambda}$, is a compact and convex subset of $\mathbb{R}^{d_\lambda}$. The support of T, $\mathcal{T}$, is a compact and convex subset of $\mathbb{R}^{d_t}$.

[5] Behrman et al. (2004) propose a similar estimator, called the generalized matching estimator for a continuous treatment variable. They use a local linear estimator, without developing its asymptotic theory.

(T, Λ) *has a probability density function* $f_{T\Lambda}(t, \lambda)$*, which is bounded away from zero and is* Δ*th-order continuously differentiable with respect to both* t *and* v*, with uniformly bounded derivatives.*

(iii) Suppose the support of Y*,* $\mathcal{Y} \equiv [y_l, y_u] \subset \mathbb{R}$*, where* y_l, y_u *are bounded.*
$F_{Y|T\Lambda}(y|t, \lambda)$ *is* Δ*th-order continuously differentiable with respect to both* t *and* λ*, with uniformly bounded derivatives.*

(iv) The unconditional distribution $F_Y(y)$ *is continuous.*
$F_{Y|T\Lambda}(y|t, \lambda) f_{T\Lambda}(t, \lambda)$ *is uniformly locally Lipschitz of order* α*, i.e., for* $0 < \alpha \leq 1$*,* $\delta_\alpha > 0$ *and* $M^{(\alpha)} < \infty$*,*

$$\sup_{y \in \mathcal{Y}, \|(t,\lambda)-(t',\lambda')\| \leq \delta_\alpha} \left| F_{Y|T\Lambda}(y|t, \lambda) f_{T\Lambda}(t, \lambda) - F_{Y|T\Lambda}(y|t', \lambda') f_{T\Lambda}(t', \lambda') \right|$$
$$\leq M^{(\alpha)} \|(t, \lambda) - (t', \lambda')\|^\alpha.$$

(iv) is from Haerdle et al. (1988) for uniform convergence in estimating a cdf. The compact support assumption (iii) implies the moments of Y exist, which is stronger than the moment conditions for the partial mean in Newey (1994b). This stronger assumption is used for inference on the empirical process. The treatments T or covariates Λ could contain discrete variables and the kernel is replaced by an indicator function, known as the frequency method. For notational convenience, discrete covariates are not allowed for. The smoothness Assumption 1.3 (ii) ensures that the treatment variables cannot have point masses, i.e., $Pr(T = t) = 0$ for $t \in \mathcal{T}$.

Assumption 1.4 (Kernel)**.** *The kernel function* $k(u) : \mathbb{R} \to \mathbb{R}$ *satisfies the following conditions: (i)* $\int k(u)du = 1$*,* $\int u^l k(u)du = 0$ *for* $0 < l < r$*, and* $\int |u^r k(u)|du < \infty$ *for some* $r \geq 2$*; (ii)* k *is of bounded support such that for some* $L < \infty$*,* $k(u) = 0$ *for* $|u| > L$*; (iii)* $k(u)$ *is* r*-times continuously differentiable and the derivatives are uniformly continuous and bounded; (iv) For an integer* Δ_k*, the derivatives of the kernel up to order* Δ_k *exist and are Lipschitz.*

Assumption 1.4 (iv) ensures that the estimator takes values in a function space not too complex for the stochastic equicontinuity argument.

Uniform convergence of the first- and second-step estimators over the range of integration suffices for deriving the properties of the third-step estimator. However, it is known that kernel estimation is biased at the boundary of the support. Following Newey (1994b), I include a fixed trimming function in the third step (Partial Sum), i.e., $\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y|T\hat{\Lambda}}(y|t,\hat{\Lambda}(S_{vi}))\cdot W(S_{wi})\cdot\pi(T_i,S_{vi})$, where the fixed trimming function π chooses a compact, interior subsupport of (T,Λ) such that the estimator $\hat{F}_{Y|T\Lambda}(y|t,\lambda)$ satisfies the uniform convergence rate in Proposition 1. In this case, the supports $\mathcal{T}$ and $\mathbf{\Lambda}$ are not restricted to be bounded.[6]

When the first step (Generated Regressors) uses a nonparametric kernel estimation, I adopt another trimming function to trim the boundary of $\mathcal{S}$ such that the estimator $\hat{\Lambda}(S)$ satisfies the convergence rate uniformly over the interior of the support of S, $\mathcal{S}_0$ in Proposition 1. That is, in the second step (Regression),

$$\begin{aligned}\hat{F}_{Y|T\hat{\Lambda}}(y|t,\lambda) &= \frac{\sum_{j=1}^{n}\mathbf{1}_{\{Y_j\leq y\}}K_h(T_j-t)K_h(\hat{\Lambda}(S_{vj})-\lambda)\cdot\mathbf{1}_{\{S_{vj}\in\mathcal{S}_0\}}}{\sum_{j=1}^{n}K_h(T_j-t)K_h(\hat{\Lambda}(S_{vj})-\lambda)\cdot\mathbf{1}_{\{S_{vj}\in\mathcal{S}_0\}}}\\ &\equiv \frac{1}{n^\dagger}\sum_{i=1}^{n^\dagger}\mathbf{1}_{\{Y_i^\dagger\leq y\}}K_h(T_i^\dagger-t)K_h(\hat{\Lambda}_i^\dagger-\lambda)\Big/\hat{f}_{T\hat{\Lambda}}^\dagger(t,\lambda)\end{aligned}$$

where $n^\dagger\equiv\sum_{j=1}^{n}\mathbf{1}_{\{S_{vj}\in\mathcal{S}_0\}}$ and the subsample is selected and relabeled $\{(Y_i^\dagger,S_{vi}^\dagger):S_{vi}^\dagger\in\mathcal{S}_0,i=1,2,..,n^\dagger\}\subset\big\{(Y_j,S_{vj}),j=1,..,n\big\}$. Then $\hat{\Lambda}_i^\dagger\equiv\hat{\Lambda}(S_{vi}^\dagger)$ and assume $\Lambda^\dagger=\Lambda(S_v^\dagger)$ satisfies Assumption 1.3. Therefore, the trimmed estimator consistently estimates $F_{Y|T\Lambda}^\dagger(y|t,\lambda)$ for the subpopulation whose observables S do not take extreme values. For the two examples I consider, the control variables in Section 1.4.2.2 and the generalized propensity score

[6]There are two alternative approaches in order to estimate for the whole support $\mathcal{S}$. The first approach assumes a compact support. A generalized kernel or boundary kernel might be used to attain the uniform convergence over the whole compact support, as suggested in Rothe (2010) and Darolles et al. (2011). The asymptotic theories and proofs derived in this paper are yet shown to be unchanged. The second approach supposes the support to be unbounded or the density to be zero at the boundary of the support. A random or data-driven trimming is needed for the denominator problem and uniform consistency over the whole support. This approach will complicate the proofs. Escanciano et al. (2012) estimate a full mean with generated regressor and random trimming.

in Section 1.4.2.3, $F^{\dagger}_{Y|T\Lambda}(y|t,\lambda)$ identifies the causal distribution $F^{\dagger}_{Y(t)|\Lambda}(y(t)|\lambda)$ for the subpopulation. Then the third step uses this subsample with the second trimming function π, $\frac{1}{n^{\dagger}}\sum_{i=1}^{n^{\dagger}} \hat{F}_{Y|T\hat{\Lambda}}(y|t,\hat{\Lambda}(S^{\dagger}_{vi}))\cdot W(S^{\dagger}_{wi})\cdot\pi(T^{\dagger}_{i},S^{\dagger}_{vi})$.

In the following, I suppress the two fixed trimming functions for notational ease, without loss of clarity. That is, I work on a compact subsupport where the density functions are bounded away from zero, as in Assumption 1.3 (ii). And the uniform convergence results in Proposition 1 hold over these compact integration ranges. But keep in mind that the identified object is for a subgroup of the population, which might be determined by the researcher's specific interest.

The following Proposition is from Lemma B.3 in Newey (1994b) and Theorem 3.2 in Haerdle et al. (1988).

Proposition 1. *Suppose the bandwidth $h \to 0$ and $\log n/(nh^d) \to 0$, where the dimension of the regressors is $d = d_t + d_\lambda$. Suppose Assumptions 1.4 and 1.3 hold. For the first four results below, assume $\Delta \geq r$ and $\Delta_k \geq 0$.*

1. $\sup_{(y,\lambda,t)\in\mathcal{Y}\times\mathbf{\Lambda}\times\mathcal{T}}\left|\hat{g}_{YT\Lambda}(y,t,\lambda)-g_{YT\Lambda}(y,t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2}+h^r\Big)$
2. $\sup_{(\lambda,t)\in\mathbf{\Lambda}\times\mathcal{T}}\left|\hat{f}_{T\Lambda}(t,\lambda)-f_{T\Lambda}(t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2}+h^r\Big)$
3. $\sup_{(\lambda,t)\in\mathbf{\Lambda}\times\mathcal{T}}\left|\hat{f}_{T|\Lambda}(t|\lambda)-f_{T|\Lambda}(t|\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2}+h^r\Big)$
4. $\sup_{(y,\lambda,t)\in\mathcal{Y}\times\mathbf{\Lambda}\times\mathcal{T}}\left|\hat{F}_{Y|T\Lambda}(y|t,\lambda)-F_{Y|T\Lambda}(y|t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^d}\big)^{1/2}+h^r\Big)$
5. *Now assume $\Delta \geq r+q$ and $\Delta_k \geq q$. Then*
   $\sup_{(y,\lambda,t)\in\mathcal{Y}\times\mathbf{\Lambda}\times\mathcal{T}}\left|\frac{\partial^q}{\partial_t^q}\hat{F}_{Y|T\Lambda}(y|t,\lambda)-\frac{\partial^q}{\partial_t^q}F_{Y|T\Lambda}(y|t,\lambda)\right| = O_p\Big(\big(\frac{\log n}{nh^{d+2q}}\big)^{1/2}+h^r\Big)$.

1.4 Estimation with Known Weight Function

This section focuses on estimation of the partial mean process with a known weight function, i.e., for each individual $i = 1,..,n$, $W(S_{wi}) = W_i$. The first subsection below presents the limit theory for estimating the partial mean process when all the regressors are observed. This is a nontrivial extension of the partial mean in Newey (1994b) to the entire

distribution process. The second subsection considers estimation of the control functions $\Lambda(S)$ as generated regressors.

1.4.1 Observable Regressors

In this subsection, I estimate the partial mean process with observed regressors Λ_i and observed weight W_i for $i = 1, ..., n$, by

$$\hat{\theta}_t(y|\Lambda, W) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_{Y|T\Lambda}(y|t, \Lambda_i) \cdot W(S_{wi})$$

processes of $y \in \mathcal{Y}$, for any $t \in \mathcal{T}$. This is for the case to identify the causal distribution of the potential outcome under unconfoundedness, where the Conditional Independence Assumption 1.1 is satisfied by $\Lambda = X$. The estimator is constructed by the second and third steps in Section 2.3. I begin by stating conditions on the object of estimation that will be used in showing the behavior of the estimator. To employ empirical process theory as part of the estimator behavior argument, I need to restrict the smoothness and complexity of the conditional cdf of outcomes. The smoothness class that I will use is defined next. In words, the partial derivatives of these functions are uniformly bounded up to some specified orders.

Definition 1.1 ($\mathcal{C}_M^\alpha(\mathcal{S})$, van der Vaart and Wellner (1996) (P. 154))**.** *$\mathcal{C}_M^\alpha(\mathcal{S})$ is defined on a bounded set $\mathcal{S}$ in $\mathbb{R}^{d_s}$ as follows: For any vector $q = (q_1, ..., q_d)$ of q_d integers, let D^q denote the differential operator $D^q = \frac{\partial^{q.}}{\partial s_1^{q_1} ... \partial s_d^{q_d}}$. Denote $q. = \sum_{l=1}^{d} q_l$. Let*

$$\|g\|_\alpha = \max_{q. \leq \underline{\alpha}} \sup_s |D^q g(s)| + \max_{q. \leq \underline{\alpha}} \sup_{s \neq s'} \frac{|D^q g(s) - D^q g(s')|}{\|s - s'\|^{\alpha - \underline{\alpha}}}$$

where $\max_{q. \leq \underline{\alpha}}$ denotes the maximum over $(q_1, ..., q_d)$ such that $q. \leq \underline{\alpha}$ and the suprema are taken over the interior of $\mathcal{S}$. Then $\mathcal{C}_M^\alpha(\mathcal{S})$ is the set of all continuous functions $g : \mathcal{S} \subset \mathbb{R}^d \mapsto \mathbb{R}$ with $\|g\|_\alpha \leq M$.

Assumption 1.5 (Complexity)**.** *(i) For any $t \in \mathcal{T}$ and for each fixed $y \in \mathcal{Y}$, $F_{Y|T\Lambda}(y|t, \cdot) \in \mathcal{C}_{\mathcal{M}}^\alpha(\mathbf{\Lambda})$, where $\underline{\alpha} = d_\lambda/2$ for even d_λ and $\underline{\alpha} = (d_\lambda - 1)/2$ for odd d_λ.*

(ii) There exists an universal constant C satisfying a Hölder continuity condition: for any $y_1, y_2 \in \mathcal{Y}$,

$$\left\| F_{Y|T\Lambda}(y_1|t,\cdot) - F_{Y|T\Lambda}(y_2|t,\cdot) \right\|_\infty \leq C|y_1 - y_2|^{1/2}.$$

Assumption 1.5 is for the stochastic equicontinuity argument in empirical process theory. It implies $\{F_{Y|T\Lambda}(y|t,\Lambda) : y \in \mathcal{Y}\}$ is Donsker by Example 19.9 in van der Vaart (2000). The smoothness condition (i) is implied by Assumption 1.3 (iii). Because the object of interest is a process indexed by y, (ii) is required to specify the complexity of the function space in y. By the assumptions on my estimators, Remark A.2 in Appendix shows that for any $y \in \mathcal{Y}$, the estimator $\hat{F}_{Y|T\Lambda}(y|t,\Lambda)$ satisfies Assumption 1.5 with probability approaching one, i.e., belongs to $\mathcal{C}^\alpha_{\mathcal{M}}(\mathbf{\Lambda})$ and satisfies the Hölder continuity condition with probability approaching one.

Assumption 1.6 (Bandwidth)**.** *The bandwidth h satisfies (i) $h \to 0$, (ii) $nh^{2r+d_t} \to 0$, and (iii) $nh^{2d-d_t}/(\log(n))^2 \to \infty$, as $n \to \infty$.*

(ii) is under-smoothing the second-step regression by reducing the bias $h^r = o((nh^{d_t})^{-1/2})$, so that the limiting distribution is centered around zero. When choosing $h \sim n^{-\eta}$, Assumption 1.6 implies $\frac{1}{2r+d_t} < \eta < \frac{1}{2d-d_t}$. So $r > d_\lambda$ implies a higher-order kernel is needed when the dimension of the regressors is large.

The following theorem presents the asymptotic linear representation and weak convergence of my estimator.

Theorem 1.1 (Weak Convergence)**.** *Suppose Assumptions 1.2, 1.3, 1.4, 1.5, and 1.6 hold, where $\Delta_k \geq \underline{\alpha}$ and $\Delta \geq \underline{\alpha} + r$. Suppose the weight W is uniformly bounded. Suppose the derivatives of $E[W|\Lambda]$ up to order r exist and are uniformly bounded and continuous. Then*

$$\sqrt{nh^{d_t}}\Big(\hat{\theta}_t(\cdot|\Lambda,W) - \theta_t(\cdot|\Lambda,W)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{tin}(\cdot|\Lambda,W) + o_p(1) \Rightarrow \mathbb{G}_t(\cdot|\Lambda,W)$$

where the influence function

$$\psi_{tin}(y|\Lambda,W) \equiv \sqrt{h^{d_t}}\ \frac{K_h\big(T_i - t\big)}{f_{T|\Lambda}(t|\Lambda_i)} \cdot \Big(\mathbf{1}_{\{Y_i \leq y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big) \cdot E\big[W(S_w)\big|\Lambda = \Lambda_i\big] \qquad (1.6)$$

[7] *and the empirical process converges weakly to a Gaussian process* $\mathbb{G}_t(\cdot|\Lambda, W)$ *with mean zero and the covariance kernel*

$$Cov\big(\mathbb{G}_t(y_1|\Lambda,W),\mathbb{G}_t(y_2|\Lambda,W)\big) = \lim_{h\to 0} E\big[\psi_{tin}(y_1|\Lambda,W)\psi_{tin}(y_2|\Lambda,W)\big]$$
$$= E\bigg[\Big(F_{Y|T\Lambda}(y_1\wedge y_2|t,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t,\Lambda)\Big)\frac{E[W|\Lambda]^2}{f_{T|\Lambda}(t|\Lambda)}\bigg]\int K^2(v)dv$$

for any $y_1, y_2 \in \mathcal{Y}$.

For the unconfoundedness case where $\Lambda = X$ and the weight is a function of the observable regressors $W = W(X)$, $E[W|\Lambda] = W(X)$ in the influence function. When the generated regressor is the generalized propensity score $\Lambda = f_{T|X}(t|X)$ and the weight function $W(X) = f_{T|X}(\bar{t}|X)/f_T(\bar{t})$, the projection $E[W|\Lambda]$ is not known in general.

Remark 1.1 (Bias)**.** The bias of the estimator is made of smaller order by the bias-reducing kernel $\sqrt{nh^{d_t}}h^r = o(1)$. Consider one continuous treatment variable $d_t = 1$ for simplicity. The bias is dominated by the bias of the influence function, which is $O(h^r)\frac{\partial^r}{\partial t^r}E\big[F_{Y|T\Lambda}(y|t,\Lambda)\cdot E[W|\Lambda]\big]$ by the standard kernel calculation in Appendix (A.11). For the finite-sample estimation, the bias is larger at the points when the counterfactual distribution has more curvature. The Monte-Carlo simulations in Section 2.5 reflect this point.

1.4.2 Generated Regressors

This section presents the asymptotic theory for nonparametric estimation of the partial mean process with generated regressors $\Lambda(S) = (X', V(T,S_v)')'$ in (1.1), where $S = (X', T', S_v')'$, $V(T,S_v)$ is a vector of measurable functions of observables $S_v \subset (X, Z_T)$ and it could contain the treatment T or not. The control function $V(T,S_v)$ is estimated in the first step. Then

$$\theta_t\big(y\big|(X,V),W\big) = E\big[F_{Y|TXV}(y|t,X,V(T,S_v))\cdot W(S_w)\big]$$
$$\hat{\theta}_t\big(y\big|(X,\hat{V}),W\big) = \frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y|TX\hat{V}}\big(y|t,X_i,\hat{V}(T_i,S_{vi})\big)\cdot W(S_{wi})$$

[7]The influence function is analogous to the influence function for the binary treatment literature, for example, Firpo and Pinto (2011), where the kernel function degenerates to an indicator function $\mathbf{1}_{\{T=t\}}$.

processes of $y \in \mathcal{Y}$, for any $t \in \mathcal{T}$. I first present a uniform stochastic expansion of the multi-step estimator, revealing the influence of estimating the generated regressors of general function form $V(T, S_v)$ on the final estimator. Then I apply this expansion to derive weak convergence of the estimated partial mean process for two economic examples from Section 1.2: the control variable and the generalized propensity score in Sections 1.4.2.2 and 1.4.2.3, respectively.

The estimator $\hat{\theta}_t\big(y\big|(X, \hat{V}), W\big)$ is constructed by the general procedure in Section 2.3. The following assumptions require the first-step estimator $\hat{V}$ to converge fast enough and to take values in a function space that is not too complex, with probability approaching one. The key for the asymptotic theory is a stochastic equicontinuity argument from empirical process theory, modified from Lemma 1 in Mammen et al. (2012a). The following high-level assumptions are borrowed from Mammen et al. (2012a). Primitive sufficient conditions are given in Section 1.4.2.1. The complexity of the function space is measured by the cardinality of the covering sets or the packing number, which can be achieved by assuming smoothness of the functions.

Assumption 1.7 (Accuracy)**.** *Let the second-step bandwidth $h_{2j} \sim n^{-\eta_j}$ for $j = 1, ..., d_2$. The j-th components $\hat{V}_j$ and V_j of vectors $\hat{V}$ and V, respectively, satisfy $\|\hat{V}_j - V_j\|_\infty = o_p(n^{-\delta_j})$, for some $\delta_j > \eta_j$ and for all $j = 1, ..., d_v$.*

Assumption 1.8 (Complexity)**.** *There exist sequences of sets of functions $\mathcal{M}_n$ such that*

1. *$V \in \mathcal{M}_n = \mathcal{M}_{n,1} \times ... \times \mathcal{M}_{n,d_v}$. $Pr(\hat{V}_j \in \mathcal{M}_{n,j}) \to 1$ as $n \to \infty$ for all $j = 1, ..., d_v$.*
2. *For a constant $C_M > 0$ and a function V_{nj} with $\|V_{nj} - V_j\|_\infty = o(n^{-\delta_j})$, the set $\bar{\mathcal{M}}_{n,j} = \mathcal{M}_j \bigcap \{V_j : \|V_{nj} - V_j\|_\infty \leq n^{-\delta_j}\}$ can be covered by at most $C_M \exp(\upsilon^{-\beta_j} n^{\xi_j})$ balls with $\|\cdot\|_\infty$-radius υ for all $\upsilon \leq n^{-\delta_j}$, where $0 < \beta_j < 2$ and $\xi_j \in \mathbb{R}$.*

The influence function of the oracle or infeasible estimator with the true regressor V is (1.6) derived in Theorem 1.1,

$$\psi_{tin}\big(y\big|(X,V),W\big) \equiv \frac{\sqrt{h_2^{d_t}}\ K_h\big(T_i - t\big)}{f_{T|XV}(t|X_i,V_i)}\Big(\mathbf{1}_{\{Y_i \le y\}} - F_{Y|TXV}(y|t,X_i,V_i)\Big) \\ E\big[W(S_w)\big|X = X_i, V = V_i\big].$$

Given these assumptions, I can now state my main results:

Theorem 1.2 (Stochastic Expansion)**.** *Suppose the conditions in Theorem 1.1 hold. Suppose Assumptions 1.7 and 1.8 hold. Then uniformly in $y \in \mathcal{Y}$, (i) when the generated regressors are not functions of the treatment T, $V = V(S_v)$,*

$$\sqrt{nh_2^{d_t}}\Big(\hat{\theta}_t\big(y\big|(X,\hat{V}),W\big) - \theta_t\big(y\big|(X,V),W\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}\big(y\big|(X,V),W\big) \\ + \sqrt{nh_2^{d_t}}E\Big[\big(\hat{V}(S_v) - V(S_v)\big)'\Big(ARG\big(y,X,S_v,V(S_v)\big) + REG\big(y,X,S_v,V(S_v)\big)\Big)\Big] + \sqrt{nh_2^{d_t}}R_n$$

where

$$ARG(y,X,S_v,V(S_v)) = \nabla_v F_{Y|TXV}(y|t,X,V(S_v)) \cdot E\big[W(S_w)\big|X,S_v\big]$$

$$REG(y,X,S_v,V(S_v)) = \Bigg[- \nabla_v F_{Y|TXV}(y|t,X,V(S_v)) \cdot E\big[W(S_w)\big|X,V = V(S_v)\big] \\ + \Big(F_{Y|TXV}(y|t,X,V(S_v)) - F_{Y|TXS_v}(y|t,X,S_v)\Big) \cdot \Bigg(- \nabla_v E\big[W(S_w)\big|X,V = V(S_v)\big] \\ + \frac{\nabla_v f_{T|XV}(t|X,V(S_v))}{f_{T|XV}(t|X,V(S_v))} \cdot E\big[W(S_w)\big|X,V = V(S_v)\big]\Bigg)\Bigg] \cdot \frac{f_{T|XS_v}(t|X,S_v)}{f_{T|XV}(t|X,V(S_v))}$$

and $R_n = O_p\big(n^{-\kappa_1} + n^{-\kappa_2} + n^{-r_2(\eta)_{min}}\big)$, $\kappa_2 < \min\big\{1 - \eta_+, 2(\delta - \eta)_{min}, (\delta - \eta)_{min} + \frac{1}{2}(1 - \sum_{j=1}^{d_2}\eta_j)\big\}$, $0 < (\delta - \eta)_{min} < \kappa_1 < \frac{1}{2}(1 - \sum_{j=1}^{d_2}\eta_j) + (\delta - \eta)_{min} - \frac{1}{2}(\delta\beta + \xi)_{max}$.

(ii) When the generated regressors are functions of the treatment T, $V = V(T, S_v)$,

$$\sqrt{nh_2^{d_t}}\Big(\hat{\theta}_t\big(y|(X,\hat{V}),W\big) - \theta_t\big(y|(X,V),W\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}\big(y|(X,V),W\big)$$

$$+ \sqrt{nh_2^{d_t}}E\Big[\big(\hat{V}(T,S_v) - V(T,S_v)\big)'ARG\big(y,X,(T,S_v),V(T,S_v)\big)\Big] \tag{1.7}$$

$$+ \sqrt{nh_2^{d_t}}E\Big[\big(\hat{V}(t,S_v) - V(t,S_v)\big)'REG\big(y,X,S_v,V(t,S_v)\big)\Big] \tag{1.8}$$

$$+ \sqrt{nh_2^{d_t}}R_n.$$

The influence of estimating the generated regressor is characterized by $E\Big[\big(\hat{V}(\cdot) - V(\cdot)\big)'\big(ARG + REG\big)\Big]$, where the first-step estimator for the generated regressor $\hat{V}(\cdot)$ is taken as a fixed function and the expectation is taken over the underlying variables $\{T, X, S_v\}$. The partial mean/full mean structure provides useful insight on the influence of estimating the generated regressors, discussed in detail in the following Remark 1.2. ARG is from estimating V as an argument. REG comes from estimating V for the regressors. The term $\nabla_v F_{Y|TXV}$ captures the magnitude of the influence of estimating V for its dual roles, regressor and argument. That is, these terms are zero when the regression function is flat in V. If I imposed the index assumption on $V(\cdot)$ such that $F_{Y|TXV} = F_{Y|TXS_v}$, then the influence of estimating the generated regressors is reduced to a similar structure as the results derived in Escanciano et al. (2012) for a full mean. In the remaining terms R_n, $n^{-\kappa_2}$ controls the smaller-order terms from linearization. Stochastic equicontinuity contributes a term of order $n^{-\kappa_1}$ to R_n.

When the generated regressors are functions of the treatment T, such as the control variable in the triangular models in Imbens and Newey (2009), the treatment variable T is fixed at t in the influence of estimating V as a regressor in (1.8). Intuitively, this is because $E[\mathbf{1}_{\{Y\leq y\}}|T = t, V(T,S_v) = v] = E[\mathbf{1}_{\{Y\leq y\}}|T = t, V(t,S_v) = v]$. On the other hand, for estimating the argument $V(T,S_v)$, the expectation in the influence (1.7) averages over T, capturing variation of T in V. This is a distinct feature of the partial mean with estimated generated regressors which contain the treatment variables.

Since my stochastic expansion is for the final partial mean estimator instead of the second-step regression function, I do not assume the Lipschitz condition on the regression function with respect to the regressors as in Mammen et al. (2012a) and Escanciano et al. (2012). That is, $\sup_v \left|E[Y|V_1 = v] - E[Y|V_2 = v]\right| \leq C\|V_1 - V_2\|_\infty$ for $V_1, V_2 \in \mathcal{V}$, some constant $C < \infty$, which requires the regression function to be very smooth, as discussed in Song (2012a).

Remark 1.2 (Full Mean/Partial Mean)**.** Newey (1994b) introduces terminology for the *partial mean* and *full mean*, which has the following structure: Let $m(X_1, X_2)$ be a conditional expectation or density.

- *Full Mean* $E_{X_1,X_2}[m(X_1, X_2)]$: All the conditioning variables or regressors (X_1, X_2) are averaged out by the outer expectation, for example, average derivative in Powell et al. (1989).
- *Partial Mean* $E_{X_1}[m(X_1, X_2 = x_2)]$: The outer expectation averages over a strict subset of the conditioning variables or regressors X_1 in the inner regression, while X_2 is fixed at value x_2. When X_2 contains continuous variables, the *Partial Mean* is infinite-dimensional and hence estimated at a nonparametric convergence rate, for example, Theorem 1.1.

Then important insight for the influence of estimating the generated regressor $E\big[\big(\hat{V}(\cdot) - V(\cdot)\big)'A(y,\cdot)\big]$ can be learned by the partial mean/full mean structure.

1. When the generated regressor is estimated parametrically or contributes as a full mean, the estimation error converges at root-n rate, i.e., $E\big[\big(\hat{V}(\cdot)-V(\cdot)\big)'A(y,\cdot)\big] = O_p(n^{-1/2})$. So the estimation error of the generated regressors is first-order ignorable, for example, the control variables in Section 1.4.2.2.
2. When the generated regressor is estimated nonparametrically and contributes as a partial mean, then $E\big[\big(\hat{V}(\cdot)-V(\cdot)\big)'A(y,\cdot)\big] = O_p(n^{-\gamma})$, where $\gamma < 1/2$. If the generated regressor is estimated by a kernel method and $E\big[\big(\hat{V}(\cdot)-V(\cdot)\big)'A(y,\cdot)\big] = O_p((nh_1^{d_t})^{-1/2})$,

then choosing $h_2 = o(h_1)$ could artificially make the estimation error of smaller order. Theorem 1.2 characterizes the complete first-order influence of the estimation error, instead of ignoring it by bandwidth choice. An example of this case is the GPS $V(X) = f_{T|X}(t|X)$ in Section 1.4.2.3, where T is fixed at t.

1.4.2.1 Primitive Conditions

When the generated regressors are specified and estimated parametrically, $\delta = 1/2$ and Assumption 1.8 is satisfied by Example 19.7 in van der Vaart (2000) for a Donsker parametric function. The following primitive conditions are sufficient for the complexity Assumption 1.8.

Assumption 1.9 (Complexity)**.** *For any $j = 1, ..., d_v$,*

1. *Let $\mathcal{M}_{n,j}$ be the set of functions defined on some compact and convex sets $\mathcal{S} \subset \mathbb{R}^{d_s}$. For any $r \in \mathcal{M}_{n,j}$, $r/n^{\xi_j^*} \in \mathcal{C}_M^\alpha(\mathcal{S})$, for some $\xi_j^* \geq 0$, $\alpha > d_s/2$, and $M > 0$.*
2. *$V_j \in \mathcal{C}_M^\alpha(\mathcal{S})$*
3. *$\|D^{\underline{\alpha}}\hat{V}_j - D^{\underline{\alpha}}V_j\|_\infty = o_p(n^{\xi_j^*})$*

Assumption 1.9-1 assumes $\mathcal{M}_{n,j}$ to be the set of functions whose partial derivatives up to order $\underline{\alpha}$ exists and are uniformly bounded by some multiple of $n^{\xi_j^*}$. By Corollary 2.7.2 in van der Vaart and Wellner (1996), Assumption 1.9-1 implies Assumption 1.8-2 by letting $\beta_j \equiv d_s/\underline{\alpha}$ and $\xi_j \equiv \xi_j^* d_s/\underline{\alpha}$. Then the complexity of the function space is controlled by the uniform bound ξ^* and the differentiability α. Assumption 1.9-2 and -3 are sufficient for Assumption 1.8-1, as discussed in Ichimura and Lee (2010). Escanciano et al. (2012) also derive a similar primitive condition in their Appendix C.

The following assumption for the second-step bandwidth is sufficient for the conditions in Theorem 1.2 and makes the remaining terms of smaller order, i.e., $\sqrt{nh_2^{d_t}}R_n = o_p(1)$.

Assumption 1.10 (2nd-step Bandwidth)**.** *The bandwidth for the second-step regression* $h_2 \sim n^{-\eta}$ *satisfies*

$$\frac{1}{2r_2 + d_t} < \eta < \min\Big\{\frac{1}{2d_2 - d_t}, \frac{\delta(2-\beta) - \xi}{d_2 - d_t + 2}, \frac{1}{d_2}(1 - \delta\beta - \xi)\Big\} \tag{1.9}$$

where the last term in the minimization is dropped if $d_t \leq 2$*. When* $d_t \leq 4$*, assume* $\delta > 1/4$*. When* $d_t > 4$ *and* $\delta < 1/4$*, assume* $\eta > \frac{1-4\delta}{d_t - 4}$*.*

I assume the convergence rate of the first-step estimator $\delta > 1/4$ in the following for simplicity. To make the upper bounds in (1.9) positive, δ must satisfy

$$\delta > \max\Big\{\frac{\xi}{2-\beta}, \frac{1}{4}\Big\}. \tag{1.10}$$

The smoothness parameters ξ and β satisfy $2 - \beta - 2\xi > 0$. The order of kernel r_2 is chosen accordingly such that (1.9) is a valid condition.

To see Assumption 1.10 is not restrictive, take an example where the first-step generated regressor is the parametric estimated GPS in Section 1.4.2.3. Then $\delta = 1/2$, $d_t = 1$, and $d_2 = 2$. Assumption 1.10 implies $\frac{1}{2r_2+1} < \eta < \frac{1}{3}\big(1 - \beta/2 - \xi\big)$. A standard second-order kernel ($r_2 = 2$) satisfies this condition by choosing the smoothness parameters β and ξ.

Now consider the first step to be a nonparametric kernel regression estimation, where the r_1-order kernel satisfies Assumption 1.4 with the bandwidth $h_1 \sim n^{-g} \to 0$. Assumptions 1.9 and 1.11 provide primitive conditions for Assumptions 1.7 (Accuracy), 1.8 (Complexity), and $\sqrt{nh_2^{d_t}}R_n = o_p(1)$.

Assumption 1.11 (Bandwidths — NP 1st-step)**.** *The first-step bandwidth* $h_1 \sim n^{-g}$ *satisfies*

$$\frac{1}{2r_1 + d_t} < g < \min\Big\{\frac{1}{2d_1},\ \frac{1}{d_1}\Big(1 - \frac{2\xi}{2-\beta}\Big), \frac{2\xi + \beta}{d_1\beta + 2d_s}\Big\}. \tag{1.11}$$

The second-step bandwidth $h_2 \sim n^{-h}$ *satisfies*

$$\frac{1}{2r_2 + d_t} < \eta < \min\Big\{\frac{1}{2d_2 - d_t}, \frac{1 - \frac{\beta}{2}(1 - d_1 g) - \xi - d_1 g}{d_2 - d_t + 2}, \frac{1}{d_2}\Big(1 - \frac{\beta}{2}(1 - d_1 g) - \xi\Big)\Big\} \tag{1.12}$$

where the last term is dropped for $d_t \leq 2$*. The smoothness parameters* β *and* ξ *satisfy* $2 - \beta - 2\xi > 0$*.*

The first two terms in the upper bound of (1.11) are from (1.10). The third term in (1.11) ensures the first-step estimator $\hat{V}$ converging to $\mathcal{M}_n$ with probability approaching one, by Assumption 1.9-3 and Proposition 1. Setting $\beta = d_s/\underline{\alpha}$ and $\xi = \xi^* d_s/\underline{\alpha}$ in Assumption 1.9, this upper bound is smaller for smaller ξ^* or larger α, i.e., if the function space is more restrictive/less complex, the first-step of estimation needs to be more accurate to ensure $\hat{V}$ belongs to $\mathcal{M}_n$ with probability approaching one. This is the additional cost of assuming a smoother function space. The trade-off between the complexity and accuracy assumptions is also discussed in Mammen et al. (2012a) and Escanciano et al. (2012).

It is feasible to choose the same bandwidth for the estimations in the first two steps. Combining (1.11) and (1.12) yields

$$\frac{1}{2\min\{r_1, r_2\} + d_t} < \eta = g < \min\left\{\frac{1}{2d_2 - d_t}, \frac{1}{2d_1}, \frac{1 - \xi - \beta/2}{d_1(1-\beta/2) + d_2 + 2 - d_t}, \frac{2\xi + \beta}{d_1\beta + 2d_s}\right\}. \tag{1.13}$$

1.4.2.2 Example I: Control Variables

To relax the unconfoundedness assumption to account for endogeneity, a well-known approach is to include control variables in the covariates. I apply my estimator to the triangular simultaneous equations models in Newey et al. (1999), Imbens and Newey (2009), and Kasy (2013). In these examples, the influence of the estimation error of the control variables is a full mean discussed in Remark 1.2 and hence is first-order ignorable.

Consider the nonseparable outcome equation $Y = \phi(T, X, \epsilon)$, where the treatment vector of interest $T = (T_1, T_2')'$ contains a single endogenous variable T_1 failing the unconfoundedness assumption. The remaining treatment subverter T_2 satisfies the selection on observables assumption, i.e., $T_2 \perp \epsilon | X$. Assume a valid control variable V for T_1 that satisfies the Conditional Independence Assumption 1.1 in the sense that $(T_1, T_2')' \perp \epsilon | (X, V)$. So the second-step regression $F_{Y|TXV}(y|t, X_i, V(S_{vi}))$ identifies $F_{Y(t)|XV}(y|X_i, V_i)$.

Assume a nonseparable first stage equation for the treatment variable:

$$T_1 = g(Z, e)$$

where the function g is strictly monotonic in the second argument. The instrumental vector Z is independent of (ϵ, e). The disturbance e is a continuously distributed scalar with cdf strictly increasing on the support of e. In general, there can be endogenous observables $X_1 \subset X = (X_1', X_2')'$ in the outcome equation ϕ, but are not in the first stage equation of T_1. And the exogenous X_2 is a subvector of $Z = (X_2', Z_T')'$, where Z_T is the excluded exogenous instrumental vector for T_1. I consider two models for control variables:

1. Imbens and Newey (2009) construct a control variable by $F_{T_1|Z}(T_1|Z)$ in their Theorem 1. Because the generated regressor is a function of both T_1 and Z, the stochastic equicontinuity argument requires the cdf estimator to be smooth in both T_1 and Z,

$$\hat{V} = \hat{V}(T_1, Z) = \hat{F}_{T_1|Z}(T_1|Z) = \frac{\frac{1}{n}\sum_{i=1}^{n} G_{h_1}(T_1 - T_{1i}) K_{h_1}(Z - Z_i)}{\frac{1}{n}\sum_{i=1}^{n} K_{h_1}(Z - Z_i)}$$

   where $G_{h_1}(u) \equiv \int^u K_{h_1}(v)dv$, $S_v = (T_1, Z')'$, and $d_s = d_z + 1$. [8]

2. Newey et al. (1999) specify a separable first stage equation,

$$T_1 = g(Z) + e, \text{where } E[e|Z] = 0.$$

   Then a valid control variable is the residual $V = T_1 - E[T_1|Z]$ with the estimator $\hat{V}(S_i) = T_{1i} - \hat{E}[T_1|Z_i]$. [9] [10]

Corollary 1.1 (Control Variable)**.** *Assume the conditions in Theorem 1.2 and Assumptions 1.9 and 1.10 hold. Consider the cases when $\hat{V}$ is a (i) parametric estimator; or (ii) a*

[8] A different model proposed by Kasy (2013) assumes that the first stage equation g is strictly monotonic in the scalar instrumental variable Z and allows multi-dimensional unobservables e. Kasy (2013) shows that although the control variable $V = F_{T_1|Z}(T_1|Z)$ fails the Conditional Independence Assumption 1.1, the control variable still identify $F_{Y(t)}(y)$ by $E\big[F_{Y|TXV}(y|t, X, V)\big] = \theta_t(y|(X, V), W = 1)$. However, $F_{Y(t)|T}(y|\bar{t})$ cannot be identified in this model.

[9] Because the first stage equation is additive, $d_s = d_z$ in (1.11). So the condition is weaker than the nonseparable first stage equation in Imbens and Newey (2009).

[10] As discussed in Section 2.3, the first trimming function in the second step (Regression) is based on the compact subsupport of Z. The second trimming function in the third step (Parital Sum) is based on the compact subsupport of (T, X, V). So the subpopulation is selected so that their values of instrumental variables, treatments, characteristics, and the unobservable in the first stage equation do not take extreme values. Using the two fixed trimming functions, the identification argument is still valid by the fact that $Z^\dagger \perp \epsilon^\dagger$ for the subpopulation.

nonparametric kernel estimator with h_1 satisfying Assumption 1.11. Then

$$\sqrt{nh^{d_t}}\Big(\hat{\theta}_t\big(\cdot\,\big|(X,\hat{V}),W\big)-\theta_t\big(\cdot\,\big|(X,V),W\big)\Big)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}\big(\cdot\,\big|(X,V),W\big)+o_p(1)\Rightarrow\mathbb{G}_t\big(\cdot\,\big|(X,V),W\big).$$

Weak convergence follows Theorem 1.1 to a Gaussian process $\mathbb{G}_t\big(\cdot\,\big|(X,V),W\big)$ with mean zero and covariance kernel

$$Cov\big(\mathbb{G}_t\big(y_1\big|(X,V),W\big),\mathbb{G}_t\big(y_2\big|(X,V),W\big)\big)=E\Bigg[\Big(F_{Y|TXV}(y_1\wedge y_2|t,X,V(T,S_v))$$
$$-F_{Y|TXV}(y_1|t,X,V(T,S_v))F_{Y|TXV}(y_2|t,X,V(T,S_v))\Big)\frac{E\big[W(S_w)\big|X,V(T,S_v)\big]^2}{f_{T|XV}(t|X,V(T,S_v))}\Bigg]\int K^2(u)du$$

for any $y_1,y_2\in\mathcal{Y}$.

The influence from estimating the control variable is first-order asymptotic ignorable. Mammen et al. (2012a) study the case $V(S)=T_1-E[T_1|Z]$ for the average structural function $E[Y(t)]$ in their Corollary 6 by using the U-statistic theory for the partial mean. Because of the full mean structure, I do not assume $h_2=o(h_1)$ as Mammen et al. (2012a) do to make smaller order of the estimation error from the second-step regression.

1.4.2.3 Example II: Treatment Effects - Generalized Propensity Score

Under the unconfoundedness assumption, Hirano and Imbens (2004) show that regressing on the generalized propensity score (GPS) $f_{T|X}(t|X)$ is sufficient for estimating continuous treatment effects. The propensity score is often used for dimension-reduction to avoid the need to match units on the values of all covariates. In practice, it is typically easier to check the common support assumption by projection on the GPS than on the support of the covariates X, as discussed in Flores et al. (2012)[11] and in the empirical example in Section 1.7.2. Consider one continuous treatment variable $d_t=1$. Define $V(t,X)=f_{T|X}(t|X)$ and

[11]Flores et al. (2012) use the parametric first step for the GPS and local polynomial estimator for the second-step regression. But they do not provide an asymptotic theory and bootstrap is used for the inference.

$V(T,X) = f_{T|X}(T|X)$. Theorem 3.1 in Hirano and Imbens (2004) implies $\forall y \in \mathcal{Y}$,

$$(i)\ E[\mathbf{1}_{\{Y(t)\le y\}}|V(t,X)=r] = E[\mathbf{1}_{\{Y\le y\}}|T=t,V(T,X)=r] = E[\mathbf{1}_{\{Y\le y\}}|T=t,V(t,X)=r]$$

$$\begin{aligned}(ii)\ E[\mathbf{1}_{\{Y(t)\le y\}}] &= E\big[E[\mathbf{1}_{\{Y\le y\}}|T=t,V(T,X)=f_{T|X}(t|X)]\big] \\ &= E\big[E[\mathbf{1}_{\{Y\le y\}}|T=t,V(t,X)=f_{T|X}(t|X)]\big]\end{aligned}$$

$$(iii)\ E[\mathbf{1}_{\{Y(t)\le y\}}|T=\bar{t}] = E\Big[E\big[\mathbf{1}_{\{Y\le y\}}\big|T=t,V(t,X)\big]\frac{f_{T|X}(\bar{t}|X)}{f_T(\bar{t})}\Big]$$

[12] (ii) implies that regressing on $V(T,X)$ or $V(t,X)$ both identify the causal object $E[\mathbf{1}_{\{Y(t)\le y\}}]$, but motivate different estimators. I use $V(t,X)=\Lambda$ for the 2nd-step regressors. [13] Then the influence from estimating the GPS contributes to a partial mean in the first order expansion, as discussed in Remark 1.2. [14]

The following corollary first presents the limit property when the nonparametric estimation of the GPS is not first-order ignorable. Second, when the GPS is estimated parametrically or nonparametrically with a faster convergence rate, the first-order asymptotic property is the same as if the true GPS was observed.

Corollary 1.2 (Generalized Propensity Score)**.** *Suppose the conditions in Theorem 1.2 and Assumptions 1.9 and 1.10 hold. For $V(X) = f_{T|X}(t|X)$, let $d_s \equiv d_x$, $d_1 = d_x+1$, and $d_2 = 2$.*

1. *Consider $\hat{V}(X) = \hat{f}_{T|X}(t|X)$ to be a nonparametric kernel estimator with order $r_1 = r_2 = r$ and $h_1 = h_2 \sim n^{-\eta}$ such that (1.13) implies*

$$\frac{1}{2r+1} < \eta < \min\Big\{\frac{1-\xi-\beta/2}{(d_x+1)(1-\beta/2)+3}, \frac{1}{2(d_x+1)}, \frac{2\xi+\beta}{(d_x+1)\beta+2d_x}\Big\}.$$

[12] Hirano and Imbens (2004) do not show the identification for the cdf for the treated $\bar{t}$, $E[\mathbf{1}_{\{Y(t)\le y\}}|T=\bar{t}]$ in (iii). I derive (iii) by modifying the proof of Theorem 3.1 in Hirano and Imbens (2004).

[13] Hirano and Imbens (2004) estimate the GPS by a normal model in the first step, i.e., $T|X \sim \mathcal{N}(X'\beta,\sigma^2)$. Their second step is a linear regression on the estimated GPS ($\hat{V}(T,X)$), the treatment variable, and their quadratic terms. The third step is a partial sum over $\hat{V}(t,X_i) = \hat{f}_{T|X}(t|X_i)$ fixing the treatment value at t. Most recent empirical research for a continuous treatment follows this semiparametric approach.

[14] As discussed in Section 2.3, the first trimming function in the second step (Regression) selects the interior compact subsupport of X. The second trimming function in the third step (Partial Sum) selects the interior compact subsupport of the treatments and the GPS. The identification argument is still valid by Theorem 1 in Hirano and Imbens (2004), making use of the facts that $f^{\dagger}_{T|X}(t|x)$ for the trimmed subpopulation is proportional to $f_{T|X}(t|x)$ by a normalization constant. So the unconfoundedness assumption holds for the subpopulation.

Then (i) for the overall distribution $F_{Y(t)}(y)$*,*

$$\sqrt{nh}\Big(\hat{\theta}_t\big(\cdot\,\big|\hat{V}, W=1\big) - F_{Y(t)}(\cdot)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(\cdot|X, W=1) + o_p(1) \Rightarrow \mathbb{G}_t(\cdot|X, W).$$

(ii) For $W \neq 1$*,*

$$\begin{aligned}
&\sqrt{nh}\Big(\hat{\theta}_t\big(\cdot\,\big|\hat{V}, W\big) - \theta_t\big(\cdot\,\big|V, W\big)\Big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}\big(\cdot\,|X, W\big) + \sqrt{h}K_h(T_i - t)\nabla_v F_{Y|TV}(\cdot|t, V_i)\Big(W_i - E[W|V_i]\Big) \\
&- \sqrt{h}K_h(T_i - t)\nabla_v E[W|V_i]\Big(F_{Y|TV}(\cdot|t, V_i) - F_{Y|TX}(\cdot|t, X_i)\Big) + o_p(1)
\end{aligned}$$

2. *Consider the cases* $\hat{V}(X) = \hat{f}_{T|X}(t|X)$ *is (i) a parametric estimator; or (ii) a nonparametric kernel estimator with* h_1 *satisfying* $h_2 = o(h_1)$ *and Assumption 1.11. Then*

$$\sqrt{nh_2}\Big(\hat{\theta}_t\big(\cdot\,\big|\hat{V}, W\big) - \theta_t\big(\cdot\,\big|V, W\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(\cdot|V, W) + o_p(1) \Rightarrow \mathbb{G}_t(\cdot|V, W).$$

Weak convergence to a Gaussian process follows as in Corollary 1.1.

The following Lemma 1.2 provides a formal and general result comparing expectations of the conditional variances given the whole set of observables X and given the index $V(X)$, respectively. Because the whole set of observables X provides finer condoning variables than its index $V(X)$, Lemma 1.2 implies the estimator for the overall distribution $F_{Y_(t)}(y)$ based on the regression on X is more efficient than the estimator using the true $V(X)$; Heckman et al. (1998) discuss the binary treatment case.

Lemma 1.2. *Suppose* $A(V(X)) = A(X)$ *is a function of* $V(X)$ *and* $A(X) \geq 0$ *a.s. Let* B *be any measurable function of* Y *such that the following moments exist. Then*

$$E\Big[var\big(B(Y)\big|T=t, X\big)\cdot A(X)\Big] \leq E\Big[var\big(B(Y)\big|T=t, V(X)\big)\cdot A(X)\Big].$$

Equality holds if and only if $E[B(Y)|T=t, V(X)] = E[B(Y)|T=t, X]$ *a.s., when* $A(X) > 0$ *a.s.*

Remark 1.3. Corollary 1.2 implies the following efficiency results for estimating the overall distribution $F_{Y(t)}(y)$ using the GPS:

1. Regression on the nonparametrically estimated GPS is first-order asymptotically equivalent to regressing on X. So there is no efficiency gain in using the GPS.
2. Corollary 1.2 and Lemma 1.2 show that the estimator based on the regression on the true GPS (the oracle estimator) or the parametric estimated GPS is less efficient than the estimator using the nonparametrically estimated GPS or the whole set of covariates X. When there exists x such that $F_{Y|TV}(y|t, V(x)) \neq F_{Y|TX}(y|t, x)$ with positive probability, the inequality between the corresponding asymptotic variances is strict.

For estimating discrete treatment effects, the convergence rate is parametric, and the semiparametric efficient bound has been calculated in Hahn (1998), Hirano et al. (2003), Firpo (2007), Chen et al. (2008), Cattaneo (2010), among others. For estimating continuous treatment effects, there are two dimensions of the efficiency to consider: the nonparametric convergence rate and the first-order asymptotic variance. In the second result of Corollary 1.2, the estimation error of the GPS converges faster by a parametric estimation or a larger bandwidth. Remark 1.3 discusses the first result when the convergence rate from the error of estimating the GPS is set the same as the first part of the influence function. The key to these results is the property of the GPS $V(X) = f_{T|X}(t|X) = f_{T|V}(t|V(X))$ and hence $\partial_v f_{T|V}(t|v) = 1$. And $\hat{V}(X)$ is estimated nonparametrically using the same kernel and bandwidth as the second-step regression. As a result, the influence from $F_{Y|TV}$ is offset in the first-order asymptotics. Parallel results for the mean effect of a binary treatment have been shown in Hahn and Ridder (2013) for the nonparametric regression estimator proposed by Heckman et al. (1998). Mammen et al. (2012b) provide an estimator for this average binary treatment effect and its regularity conditions. They derive the same influence function as Hahn and Ridder (2013).

The second point in Remark 1.3 is parallel to the result in the propensity score weighting estimator for discrete treatment effects in Hirano et al. (2003), Cattaneo (2010), and Graham (2011). Hirano et al. (2003) interpret the efficiency loss of using the true propensity score (PS) by the empirical likelihood estimation. Nonparametrically estimating the PS captures the information content of a conditional moment restriction of the PS ($E[T - V(X)|X] = 0$) by a sequence of unconditional moment restrictions. While a parametric estimate of the PS will only satisfy a finite number of the moment conditions, using the true PS makes no use of any information contained in the auxiliary moment. So the efficiency is improved in the same way as adding moment restrictions in a GMM framework. Graham (2011) calculates the efficiency bound incorporating the conditional moment of the PS as the auxiliary moment. Cattaneo (2010) claims the nonparametrically estimated GPS could approximate the correction term in the efficient influence function, which is missed in a GMM estimator using only the identifying moment. Imai and van Dyk (2004) also discuss this result for subclassification using the GPS by an application with randomized treatment assignment. The estimated GPS accounts for the sample-specific relationship of the treatment and the covariates, which is lost in the true GPS. The above discussion gives intuition to the finding that the information of the GPS will not improve efficiency for estimating the overall $F_{Y(t)}(y)$.

On the other hand, Corollary 1.2 1 (ii) implies estimating $F_{Y(t)|T}(y|\bar{t})$ for the treated group with the weight $W(X) = f_{T|X}(\bar{t}|X)/f_T(\bar{t})$ does not carry the above efficiency properties as in the binary case. This is because $E[W|V] \neq W$ in general. And the regression estimator for a binary treatment in Heckman et al. (1998) is calculated only for the treated subgroup instead of using a weight.

My estimator is $\sqrt{nh}$-consistent due to the partial mean, so there is no curse of dimensionality. When the dimension of regressors is larger, Assumption 1.11 requires smoother distribution functions and higher-order kernels. Then it follows that the bandwidth converges to zero slower, which results in a faster convergence rate ($\sqrt{nh}$). Therefore, the advantage of using the GPS (over regressing on the whole set of covariates) is on dimension-reduction at the second-step regression, but not on the convergence rate. The first-step GPS

estimation has higher-dimensional regressors, so we could choose $h_2 = o(h_1)$. It then allows weaker assumption on the smoothness of $F_{Y|TV}$ and a lower-order kernel at the cost of a slower convergence rate. Note that the above results are for the first-order asymptotics. The finite-sample performance is to be investigated. Song (2012b) also discusses this dimension-reduction issue for the single-index nuisance parameters in the semiparametric models.

1.5 Estimation with Unknown Weight Function

The section considers estimating the partial mean process (1.1) when the weight function $W(S_w) \equiv f_{T|S_w}(\bar{t}|S_w)/f_T(\bar{t})$ is unobserved and estimated. This weight function uncovers the distribution of $Y(t)$ for the treated $\bar{t}$, $F_{Y(t)|T}(y|\bar{t})$, which is identified by the partial mean of weighted distribution process $\theta_t(y|\Lambda, W)$ for different models summarized in Table 1.1. The estimator $\hat{\theta}_t(y|\hat{\Lambda}, \hat{W})$ involving estimating the generated regressor Λ and the weight W follows the procedure described in Section 2.3. I derive an additional term in the influence function contributed by the estimation error of the weight function. Together with the limit theorems in the previous section when the weight function is known, I obtain weak convergence of the estimator $\hat{\theta}_t(y|\hat{\Lambda}, \hat{W})$ for each of the following objects:

1. (Observable Regressors)

$$\theta_t\big(y\big|X, W\big) = E\big[F_{Y|TX}(y|t, X) \cdot f_{T|X}(\bar{t}|X)/f_T(\bar{t})\big]$$

2. (Generalized Propensity Score)

$$\theta_t\big(y\big|V, W\big) = E\big[F_{Y|TV}(y|t, V) \cdot f_{T|X}(\bar{t}|X)/f_T(\bar{t})\big], \text{where } V = f_{T|X}(t|X)$$

3. (Control Variables)

$$\theta_t\big(y\big|(X, V = V_{\bar{t}}), W(X, Z)\big) = E\big[F_{Y|TXV}(y|t, X, V(T = \bar{t}, Z)) \cdot f_{T|X}(\bar{t}|X, Z)/f_T(\bar{t})\big]$$

To estimate $F_{Y(t)|T}(y|\bar{t})$, the weight function is unobserved and needs to be estimated, except in a randomized experiment. The first two objects are under the unconfoundedness assumption. The third object allows selection on unobservables by Lemma 1.1 and the control

variable $V = V(T, Z)$ is constructed in the models of Imbens and Newey (2009) and Newey et al. (1999) in Section 1.4.2.2.

The key to the following asymptotic theorem is stochastic equicontinuity in the weight function. In Lemma A.7 in Appendix, I show that $\sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n} F_{Y|T\Lambda}(y|t,\Lambda_i)\big(\hat{W}(S_{wi}) - W(S_{wi})\big) = \sqrt{nh^{d_t}}E\Big[F_{Y|T\Lambda}(y|t,\Lambda)\big(\hat{W}(S_w) - W(S_w)\big)\Big] + o_p(1)$, uniformly in $y \in \mathcal{Y}$. The estimation error of the parametric weight, such as a normal model, is of smaller order by $\sqrt{nh^{d_t}} = o(\sqrt{n})$. So the first-order asymptotic distribution of the estimator is the same as if the weight was known. When the weight is estimated by a nonparametric kernel method, the estimation error from the weight contributes to a partial mean by fixing T at $\bar{t}$ and hence is not first-order ignorable. The following theorem presents the limit theory. Let b denote the bandwidth when the weight $W(S_w) = f_{T|S_w}(\bar{t}|S_w)/f_T(\bar{t})$ is estimated nonparametrically. Denote the influence function for $\sqrt{nb^{d_t}}E\Big[F_{Y|T\Lambda}(y|t,\Lambda)\big(\hat{W}(S_w) - W(S_w)\big)\Big]$ by

$$\psi^{\bar{t}}_{tin}(y|\Lambda) \equiv \sqrt{b^{d_t}}K_b\big(T_i - \bar{t}\big)\Big(E\big[F_{Y|T\Lambda}(y|t,\Lambda(S))\big|S_w = S_{wi}\big] - \theta_0(y|\Lambda, W)\Big)\Big/ f_T(\bar{t}). \quad (1.14)$$

Theorem 1.3 (Treatment Effects on the Treated)**.** *Assume $f_{T|S}(\bar{t}|\cdot) \in \mathcal{C}^{\alpha}_{M}(\mathcal{S})$. Suppose Assumption 1.4 holds for the r_w-order kernel in $\hat{W}(X)$. The bandwidth $h = O(b)$, $b \to 0$, $nh^{d_2-d_t}b^{d_w+d_t}/\log(n) \to \infty$, $nb^{2d_w+d_t}/(\log n)^2 \to \infty$,* [15] *and $nb^{2r_w+d_t} \to 0$,* [16] *where $d_w \equiv dim(S_w)$. When $h = b$, then*

1. *(Observable Regressors) Suppose the conditions for Theorem 1.1 hold.*

$$\sqrt{nh^{d_t}}\Big(\hat{\theta}_t\big(\cdot\,\big|X, \hat{W}(X)\big) - F_{Y(t)|T}(\cdot|\bar{t})\Big)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\psi_{tin}(\cdot|X, W(X)) + \psi^{\bar{t}}_{tin}(\cdot|X)\Big) + o_p(1) \Rightarrow \mathbb{G}_{t|\bar{t}}(\cdot)$$

[15] This is from the stochastic equicontinuity argument to make $\hat{W} \in \mathcal{C}^{\alpha}_{M}$ with probability approaching one by Remark A.2 in Appendix.

[16] The bias-reducing kernel is used for $\hat{W}$ for simplicity.

where the influence function $\psi_{tin}(y|X,W)$ is defined in (1.6) and the Gaussian process $\mathbb{G}_{t|\bar{t}}$ is mean zero with covariance kernel $Cov(\mathbb{G}_{t|\bar{t}}(y_1), \mathbb{G}_{t|\bar{t}}(y_2)) \equiv$

$$\frac{\int K^2(v)dv}{f_T^2(\bar{t})} E\Bigg[\Bigg[\frac{f_{T|X}(\bar{t}|X)}{f_{T|X}(t|X)}\Big(F_{Y|TX}(y_1 \wedge y_2|t,X) - F_{Y|TX}(y_1|t,X)F_{Y|TX}(y_2|t,X)\Big) + \Big(F_{Y|TX}(y_1|t,X) - F_{Y(t)|T}(y_1|\bar{t})\Big)\Big(F_{Y|TX}(y_2|t,X) - F_{Y(t)|T}(y_2|\bar{t})\Big)\Bigg] f_{T|X}(\bar{t}|X)\Bigg].$$

When the generated regressors are estimated, further assume $n^{\delta} h_2^{1-d_t} b^{d_w+d_t} / \log n \to \infty$.

2. *(Generalized Propensity Score) Suppose the conditions for Corollary 1.2 hold. $V = f_{T|X}(t|X)$. Then uniformly in $y \in \mathcal{Y}$,*

$$\sqrt{nh^{d_t}}\Big(\hat{\theta}_t\big(y\big|\hat{V}, \hat{W}(X)\big) - F_{Y(t)|T}(y|\bar{t})\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\psi_{ti}^{GR}(y) + \psi_{tin}^{\bar{t}}(y|V)\Big) + o_p(1)$$

where ψ_{ti}^{GR} denotes the influence function derived in Corollary 1.2.

3. *(Control Variables) Suppose the conditions for Corollary 1.1 hold. Then uniformly in $y \in \mathcal{Y}$,*

$$\begin{aligned}
&\sqrt{nh^{d_t}}\Big(\hat{\theta}_t\big(y\big|(X, \hat{V} = \hat{V}_{\bar{t}}), \hat{W}(X,Z)\big) - F_{Y(t)|T}(y|\bar{t})\Big) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg\{ K_h\big(T_i - t\big)\Big(\mathbf{1}_{\{Y_i \le y\}} - F_{Y|TXV}(y|t, X_i, V(\bar{t}, Z_i))\Big) \frac{f_{T|XZ}(\bar{t}|X_i, Z_i)}{f_{T|XV}(t|X_i, V(\bar{t}, Z_i))} \\
&\quad + K_h\big(T_i - \bar{t}\big)\Big(F_{Y|TXV}(y|t, X_i, V(\bar{t}, Z_i)) - F_{Y(t)|T}(y|\bar{t})\Big)\Bigg\} \frac{\sqrt{h^{d_t}}}{f_T(\bar{t})} + o_p(1).
\end{aligned}$$

When $h = o(b)$ or $\hat{W}(S)$ is parametric, the first-order asymptotic property is described by Theorems in Section 1.4 as if the weight was observed. For estimating the distribution for the treated with estimated control variables in Theorem 1.3-3, the generated regressor is $V = V(T, Z)$ and the argument summed out in the third step is $V(T = \bar{t}, Z)$ fixing T at $\bar{t}$.

1.6 Inference for the Treatment Effects

Often the objects of ultimate interest are policy effects or inequality measures. Such objects can be expressed as functionals of the potential outcome distributions identified by

$\theta_t(y|\Lambda, W)$ in Table 1.1 and estimated in previous sections. In this section, I provide the distribution theory for a class of smooth functionals of the three step outcome distribution estimator above. The key to this result is the functional delta method for Hadamard-differentiable functionals (Theorem 20.8 in van der Vaart (2000)). I illustrate the results by the mean and quantile operators.

Assumption 1.12. *The functional* Γ *defined over the distribution functions of potential outcomes is Hadamard differentiable.* [17]

These Hadamard-differentiable functionals can be highly nonlinear functionals of the cdf, but admit a linear functional derivative. Weak convergence of the estimators will be implied by the functional delta method in empirical process theory. Assumption 1.12 is a high-level assumption that could impose restrictions or smoothness on the distribution functions of potential outcomes. In particular, when Γ is the τ-quantile operator on $\theta_t(y) = \theta_t(y|\Lambda, W)$, Γ is a generalized inverse $\theta_t^{-1} : (0,1) \to \mathcal{Y}$ given by $\theta_t^{-1}(\tau) = \inf\{y : \theta_t(y) \geq \tau\}$. Then Assumption 1.12 means $\theta_t(y)$ is continuously differentiable at the τth-quantile, with the derivative being strictly positive and bounded over a compact neighborhood. Additional assumptions might be needed for different policy functionals. For instance, Bhattacharya (2007) gives regularity conditions for Hadamard-differentiability of Lorenz and Gini functionals.

I consider each of the identification functions for the causal objects, $F_{Y(t)}(\cdot)$ and $F_{Y(t)|T}(\cdot|\bar{t})$, listed in Table 1.1. The corresponding asymptotic theorem derived in previous sections provides the influence function and weak convergence: denoting as $\sqrt{nh^{d_t}}\big(\hat{\theta}_t - \theta_t\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{tin} + o_p(1)$ and converges weakly to a Gaussian process $\mathbb{G}_t$.

Theorem 1.4 (Functional Delta Method)**.** *Assume the conditions in the asymptotic theorem for* $\hat{\theta}_t$ *hold.*

[17] Let Γ be a Hadamard-differentiable functional mapping from $\mathcal{F}$ to some normed space E, with derivative Γ'_f, a continuous linear map $\mathcal{F} \mapsto E$. For every $h_n \to h$ and $f \in \mathcal{F}$,

$$\lim_{v \to 0} \frac{1}{v}\Big(\Gamma(f + vh_n) - \Gamma(f)\Big) = \Gamma'_f(h).$$

Consider the parameter θ as an element of a parameter space $D_\theta \subset l^\infty(\mathcal{Y})$ with D_θ containing the true value θ_t. Suppose a functional $\Gamma(\theta)$ mapping D_θ to $l^\infty(\mathcal{W})$ is Hadamard differentiable in θ at θ_t with derivative Γ'_θ. Then

$$\left|\sqrt{nh^{d_t}}\big(\Gamma(\hat{\theta}_t)(w) - \Gamma(\theta_t)(w)\big) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma'_\theta(\psi_{tin})(w)\right| = o_p(1)$$

$$\sqrt{nh^{d_t}}\big(\Gamma(\hat{\theta}_t)(w) - \Gamma(\theta_t)(w)\big) \Rightarrow \Gamma'_\theta(\mathbb{G}_t)(w) \equiv G(w)$$

where G is a Gaussian process indexed by $w \in \mathcal{W}$ in $l^\infty(\mathcal{W})$, with mean zero and covariance kernel defined by the limit of the second moment of $\Gamma'_\theta(\psi_{tin})$.

The following corollary gives the policy/inequality treatment effects of shifting the treatment from $\bar{t}$ to t, $\Gamma\big(\theta_t\big) - \Gamma\big(\theta_{\bar{t}}\big)$. The estimations of the distributional features at different treatment levels t and $\bar{t}$, $\Gamma\big(\theta_t\big)$ and $\Gamma\big(\theta_{\bar{t}}\big)$, are asymptotically uncorrelated.

Corollary 1.3 (Causal Effects)**.** *Assume the conditions in Theorem 1.4. Then*

$$\sqrt{nh^{d_t}}\begin{pmatrix}\hat{\theta}_t(\cdot) - \theta_t(\cdot)\\ \hat{\theta}_{\bar{t}}(\cdot) - \theta_{\bar{t}}(\cdot)\end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\psi_{tin}(\cdot)\\ \psi_{\bar{t}i}(\cdot)\end{pmatrix} + o_p(1) \Rightarrow \mathbb{G}_{t\bar{t}}(\cdot)$$

a Gaussian process with zero mean. The diagonal elements of the covariance matrix are the covariance matrix of $\mathbb{G}_t$ and $\mathbb{G}_{\bar{t}}$. And the off-diagonal terms are zero. Theorem 1.4 implies

$$\sqrt{nh^{d_t}}\Big(\Gamma(\hat{\theta}_t) - \Gamma(\hat{\theta}_{\bar{t}}) - \big(\Gamma(\theta_t) - \Gamma(\theta_{\bar{t}})\big)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\Gamma'_\theta(\psi_{tin}) - \Gamma'_\theta(\psi_{\bar{t}i})\Big) + o_p(1) \Rightarrow \mathbb{G}^{\Gamma}_{t\bar{t}}$$

a mean-zero Gaussian process with the covariance kernel $Cov = \lim_{h\to 0} E\big[\big(\Gamma'_\theta(\psi_{tin}) - \Gamma'_\theta(\psi_{\bar{t}i})\big)^2\big]$, the summation of the covariance of $\Gamma'_\theta(\mathbb{G}_t)$ and the covariance of $\Gamma'_\theta(\mathbb{G}_{\bar{t}})$.

1.6.1 Examples: Mean and Quantile

1.6.1.1 Mean

The following corollary presents the asymptotic theory of estimating the means, $E[Y(t)]$ and $E[Y(t)|T = \bar{t}]$. The first result applies to the overall mean of the potential outcome $E[Y(t)]$ that is the partial mean in Newey (1994b), the average structural function in Blundell

and Powell (2003), and the dose response function in Flores (2007). My results allow to relax the unconfoundedness assumption by estimating the control variables as in Section 1.4.2.2.

The mean for the cdf θ_t is $\Gamma(\theta_t) = \int_{\mathcal{Y}} u d\theta_t(u)$, which has the Hadamard derivative $\Gamma'(\theta) = \int u d\theta(u)$. Then the estimator is $\int y \, d\hat{\theta}_t(y)$, i.e., replace the dependent variable $\mathbf{1}_{\{Y \leq y\}}$ with Y in the estimation procedure described in Section 2.3. Denote the mean operator on the influence function (1.6) by

$$\psi^{\mu}_{tin}(\Lambda, W) \equiv \Gamma'(\psi_{tin}(\cdot|\Lambda, W)) = \frac{\sqrt{h^{d_t}} K_h\Big(T_i - t\Big)}{f_{T|\Lambda}(t|\Lambda_i)} \Big(Y_i - E[Y|T = t, \Lambda = \Lambda_i]\Big) \cdot E[W|\Lambda_i]. \tag{1.15}$$

Denote the mean operator on the influence from estimating the weight (1.14) by

$$\psi^{\mu\bar{t}}_{tin}(\Lambda) \equiv \Gamma'\big(\psi^{\bar{t}}_{tin}(\cdot|\Lambda)\big) = \frac{\sqrt{h^{d_t}} K_h\big(T_i - \bar{t}\big)}{f_T(\bar{t})} \Big(E\big[E[Y|T = t, \Lambda = \Lambda_i]\big|S_w = S_{wi}\big] - E[Y(t)|T = \bar{t}]\Big). \tag{1.16}$$

Corollary 1.4 (Mean)**.** *Assume the conditions in Theorem 1.4.*

- *(Known Weight) Consider the case when $\Lambda = X$ and $W(S) = W$ are observable. Theorem 1.1 implies*

$$\sqrt{nh^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{E}[Y|T = t, X_i]W_i - E\Big[E[Y|T = t, X]W\Big]\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi^{\mu}_{tin}(X, W) + o_p(1)$$
$$\xrightarrow{d} \mathcal{N}(0, V_\mu), \quad \textit{where } V_\mu = E\Big[var\big(Y\big|T = t, X\big)\frac{E[W(S_w)|X]^2}{f_{T|X}(t|X)}\Big] \cdot \int K^2(u)du.$$

- *(Unknown Weight - Treatment Effects on the Treated $E[Y(t)|T = \bar{t}]$)*

  1. *(Selection on Observables) Consider the case when $\Lambda = X$. Theorem 1.3-1 implies*

$$\sqrt{nh^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{E}\big[Y\big|T = t, X_i\big] \cdot \frac{\hat{f}_{T|X}(\bar{t}|X_i)}{\hat{f}_T(\bar{t})} - E\big[Y(t)\big|T = \bar{t}\big]\Big)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\psi^{\mu}_{tin}(X, W) + \psi^{\mu\bar{t}}_{tin}(X)\big) + o_p(1) \xrightarrow{d} \mathcal{N}(0, V_\mu + V^{\bar{t}}_\mu)$$

where $V^{\bar{t}}_\mu = E\Big[\Big(E\big[Y\big|T = t, X\big] - E\big[Y(t)\big|T = \bar{t}\big]\Big)^2 f_{T|X}(\bar{t}|X)\Big] \cdot \frac{\int K^2(u)du}{f^2_T(\bar{t})}$.

2. *(Generalized Propensity Score) Consider the case when* $\Lambda = f_{T|X}(t|X)$ *in Section 1.4.2.3. Theorem 1.3-2 implies*

$$\sqrt{nh^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{E}\big[Y\big|T=t,\hat{V}=\hat{V}_i\big]\cdot\frac{\hat{f}_{T|X}(\bar{t}|X_i)}{\hat{f}_T(\bar{t})}-E\big[Y(t)\big|T=\bar{t}\big]\Big)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\Gamma'\big(\psi_{tin}^{GR}\big)+\psi_{tin}^{\mu\bar{t}}(V)\Big\}+o_p(1)$$

where ψ_{tin}^{GR} *denotes the influence function derived in Corollary 1.2 and* $\Gamma'\big(\psi_{tin}^{GR}\big)$ *is calculated by (1.15).*

3. *(Control Variable) Consider the case when* $\Lambda=(X',V')'$ *and* $V=V(T,Z)$ *in Section 1.4.2.2. Theorem 1.3-3 implies*

$$\sqrt{nh^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{E}\big[Y\big|T=t,X=X_i,\hat{V}=\hat{V}(\bar{t},Z_i)\big]\cdot\frac{\hat{f}_{T|XZ}(\bar{t}|X_i,Z_i)}{\hat{f}_T(\bar{t})}-E\big[Y(t)\big|T=\bar{t}\big]\Big)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{K_h\big(T_i-t\big)\Big(Y_i-E\big[Y\big|T=t,X=X_i,V=V(\bar{t},Z_i)\big]\Big)\frac{f_{T|XZ}(\bar{t}|X_i,Z_i)}{f_{T|XZ}(t|X_i,Z_i)}$$
$$+K_h\big(T_i-\bar{t}\big)\Big(E\big[Y\big|T=t,X=X_i,V=V(\bar{t},Z_i)\big]-E\big[Y(t)\big|T=\bar{t}\big]\Big)\Big\}\frac{\sqrt{h^{d_t}}}{f_T(\bar{t})}$$
$$+o_p(1).$$

I can modify the first result for the known weight in Corollary 1.4 to the case with generated regressors. If we suppose the conditions of Corollary 1.1 hold where the generated regressors V are control variables, then the influence function in the conclusion would be $\psi_{tin}^{\mu}(X,V)$ in place of $\psi_{tin}^{\mu}(X)$. Similarly, the asymptotic results with the generalized propensity score in Corollary 1.2 are implied by replacing the corresponding influence functions for the known weight in Corollary 1.4 with $\psi_{tin}^{\mu}(X)$ or $\psi_{tin}^{\mu}(V)$.

1.6.1.2 Quantile Processes

The unconditional quantile function is inverted directly from the unconditional cdf. For the quantile process $\{Q_\tau:\tau\in(0,1)\}$ of the cdf θ_t, $Q_\tau\equiv\inf\{y:\theta_t(y)\geq\tau\}$. The following corollary gives the asymptotic theory of estimating unconditional quantile function of $Y(t)$

for the whole population and the treated group $\bar{t}$, assuming unconfoundedness and using control variables respectively, as listed in Table 1.1. I illustrate the results by two examples: the quantile structural function in Imbens and Newey (2009) and the quantile treatment effects on the treated assuming unconfoundedness.

Corollary 1.5 (Quantile Process)**.** *Assume the conditions in Theorem 1.4. Suppose* $\sqrt{nh^{d_t}}\big(\hat{\theta}_t(\cdot) - \theta_t(\cdot)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(\cdot) + o_p(1) \Rightarrow \mathbb{G}_t(\cdot)$*. Assume* θ_t *is continuously differentiable with strictly positive derivative* $\frac{\partial}{\partial y}\theta_t(y)\Big|_{y=Q_\tau} \equiv \theta_t'(Q_\tau)$*. Then the influence function for estimating the quantile process is*

$$\psi_{tin}^{Q}(\tau) \equiv -\psi_{tin}\big(Q_\tau\big)\Big/\theta_t'(Q_\tau).$$

Therefore,

$$\sqrt{nh^{d_t}}\big(\hat{Q}_\cdot - Q_\cdot\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}^{Q}(\cdot) + o_p(1) \Rightarrow -\mathbb{G}_t\big(Q_\cdot\big)\Big/\theta_t'(Q_\cdot) \equiv \mathbb{G}_t^{Q}(\cdot)$$

where $\mathbb{G}_t^{Q}$ *is a Gaussian process indexed by* $\tau \in [a,b] \subset (0,1)$ *in the metric space* $l^{\infty}([a,b])$*. The Gaussian process* $\mathbb{G}_t^{Q}$ *has zero mean and covariance kernel, for any* $\tau_1 < \tau_2 \in [a,b]$*,* $Cov\big(\mathbb{G}_t^{Q}(\tau_1), \mathbb{G}_t^{Q}(\tau_2)\big) = \lim_{h\to 0} E\big[\psi_{tin}^{Q}(\tau_1)\psi_{tin}^{Q}(\tau_2)\big]$*.*

Remark

1. (Quantile Structural Function in Imbens and Newey (2009))

   Consider the τth-quantile function of $Y(t)$, $Q_\tau = Q_\tau(Y(t)) = F_{Y(t)}^{-1}(\tau)$. The conditioning variables are $\Lambda(S) = (X', V')'$, where the control variables V is estimated in Section 1.4.2.2. Corollaries 1.1 and 1.5 imply

   $$\sqrt{nh^{d_t}}\big(\hat{Q}_\cdot(Y(t)) - Q_\cdot(Y(t))\big) \Rightarrow \mathbb{G}_t^{Q}(\cdot)$$

   a Gaussian process with mean zero and covariance

   $$Cov(\mathbb{G}_t^{Q}(\tau_1),\mathbb{G}_t^{Q}(\tau_2)) \equiv E\Bigg[\frac{1}{f_{T|XV}(t|X,V)}\Big(F_{Y|TXV}(Q_{\tau_1}|t,X,V) - F_{Y|TXV}(Q_{\tau_1}|t,X,V)F_{Y|TXV}(Q_{\tau_2}|t,X)\Big)\Bigg]\frac{\int K^2(v)dv}{f_{Y(t)}(Q_{\tau_1})\ f_{Y(t)}(Q_{\tau_2})}.$$

2. (Quantile Treatment Effects on the Treated)

   Consider the τth-quantile function of $Y(t)$ for the treated $\bar{t}$, $Q_\tau = Q_\tau(Y(t)|T=\bar{t}) = F^{-1}_{Y(t)|T}(\tau|\bar{t})$. Assuming unconfoundedness, consider the estimator with $\Lambda = X$ and $\hat{W}(X) = \hat{f}_{T|X}(\bar{t}|X)/\hat{f}_T(\bar{t})$ as in Theorem 1.3-1. Then

$$\sqrt{nh^{d_t}}\Big(\hat{Q}_\cdot(Y(t)|T=\bar{t}) - Q_\cdot(Y(t)|T=\bar{t})\Big) = \frac{1}{n}\sum_{i=1}^{n}\Big(\psi^{Q}_{tin}(\cdot|X,W) + \psi^{Q\bar{t}}_{tin}(\cdot|X)\Big) + o_p(1)$$

   converges weakly to a Gaussian process indexed by $\tau \in [a,b]$ with mean zero and covariance matrix $\lim_{h\to 0} E\big[\big(\psi^{Q}_{ti}(\tau_1|X,W) + \psi^{Q\bar{t}}_{tin}(\tau_1|X)\big)\big(\psi^{Q}_{ti}(\tau_2|X,W) + \psi^{Q\bar{t}}_{tin}(\tau_2|X)\big)\big]$ for any $\tau_1, \tau_2 \in [a,b]$. By Theorem 1.3-1 and Corollary 1.5, the influence functions are

$$\psi^{Q}_{tin}(\tau|X,W) \equiv \frac{-\sqrt{h^{d_t}}K_h\big(T_i - t\big)}{f_T(\bar{t})f_{Y(t)|T}(Q_\tau|\bar{t})}\,\frac{f_{T|X}(\bar{t}|X_i)}{f_{T|X}(t|X_i)}\Big(\mathbf{1}_{\{Y_i \le Q_\tau\}} - F_{Y|TX}(Q_\tau|t,X_i)\Big)$$

$$\psi^{Q\bar{t}}_{tin}(\tau|X) \equiv \frac{-\sqrt{h^{d_t}}K_h\big(T_i - \bar{t}\big)}{f_T(\bar{t})f_{Y(t)T}(Q_\tau|\bar{t})}\Big(F_{Y|TX}(Q_\tau|t,X_i) - F_{Y(t)|T}(Q_\tau|\bar{t})\Big).$$

   I do not state the limit theory for the estimators of regressing on the GPS and the control variables, which could be derived similarly, based on Theorem 1.3-2 andTheorem 1.3-3.

1.6.2 Inference

The pointwise influence function can be estimated by replacing unknown functions with consistent estimators. Then the covariance matrix can be estimated by the sample variance of the estimated influence functions. Alternatively, the covariance matrix can be estimated by a plug-in method that is a sample analogue with consistently estimated unknown functions.

Besides pointwise inference, we might be interested in testing a hypothesis involving a policy on the whole distribution: constant effect or stochastic dominance. I suggest using a multiplier method to simulate the empirical processes defined in Theorem 1.1, Corollary 1.1, and Corollary 1.2. The multiplier method has been used in Donald et al. (2012) to simulate a conditional distribution process. It is easy to perform asymptotically valid inference on distributional features defined by the Hadamard-differentiable functionals. Let $\{U_i\}_{i=1}^n$ be a

sequence of i.i.d. random variables with mean zero and variance one, for example, $\mathcal{N}(0,1)$, independent of the data. The influence function ψ_t for the estimator $\hat{\theta}_t$ is estimated consistently by some estimator $\hat{\psi}_t$. The following theorem shows that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\hat{\psi}_{tin}(\cdot)$ simulates the asymptotic distribution of the estimator.

Theorem 1.5 (Multiplier CLT - Known Weight Function)**.** *Assume the conditions in Theorem 1.1 or Corollary 1.1 or 1.2 which gives* $\sqrt{nh^{d_t}}\big(\hat{\theta}_t(\cdot)-\theta_t(\cdot)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(\cdot)+o_p(1) \Rightarrow \mathbb{G}_t(\cdot)$. *Then*

$$\mathbb{G}^M_{tin}(\cdot) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\hat{\psi}_{tin}(\cdot) \Rightarrow \mathbb{G}_t(\cdot)$$

conditional on sample path with probability approaching 1. For the Hadamard-differentiable functional Γ*,*

$$\Gamma'\Big(\mathbb{G}^M_{tin}(\cdot)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\Gamma'\big(\hat{\psi}_{tin}(\cdot)\big) \Rightarrow \Gamma'\big(\mathbb{G}_t(\cdot)\big).$$

The multiplier CLT for estimating the unknown weight function should be modified straightforward by estimating the additional influence function $\psi^{\bar{t}}_{tin}$ consistently, although I do not show the proof here.

1.7 Numerical Examples

1.7.1 Monte Carlo Simulation

The finite-sample performances of the proposed semi- and nonparametric estimators are compared with the parametric estimator in Hirano and Imbens (2004). I consider two data generating processes with varying degrees of nonlinearity. Perhaps not surprisingly when the true function $E[Y(t)]$ is more non-linear, my semi- and nonparametric estimators perform relatively better, even in small sample ($n = 100$).

Four estimators are examined. The first three estimators regress on the estimated generalized propensity score (GPS). The fourth estimator regresses directly on covariates X. For the 1st-step GPS estimators, I consider two methods, the parametric normal model, i.e.,

$T|X \sim \mathcal{N}(X\beta, \sigma^2)$, and the nonparametric kernel method. The first estimator is proposed by Hirano and Imbens (2004), who use a quadratic linear regression on T and the estimated GPS for the 2nd-step. The second and third estimators use a nonparametric kernel method for the 2nd-step regression. A detailed procedure to implement the proposed estimators can be found in Section 3. These four estimators are summarized and labeled as follows:

1. *HI* (Hirano and Imbens, 2004): 1st-Normal GPS + 2nd-quadratic linear regression.
2. *SP* (semiparametric): 1st-Normal GPS + 2nd-kernel regression.
3. *NP-GPS* (nonparametric): 1st-kernel estimated GPS + 2nd-kernel regression.
4. *NP-X* (nonparametric): kernel regression on the covariates X.

I consider two data generating processes: (DGP2) is more non-linear than (DGP1).

$$Y = 3\sin(0.5T) + X_2T^2 + (1 - X_1)X_2 + U \qquad \text{(DGP1)}$$

$$Y = 3\sin(2T) + X_2T^2 + (1 - X_1)X_2 + U \qquad \text{(DGP2)}$$

where X_1 and X_2 are i.i.d. $Unif(0, 1)$. The conditional distribution of U is $\mathcal{N}(0, X_1^2)$. The treatment variable is $T = \cos(2\pi X_1) + Z + e$, where Z and e are i.i.d. $\mathcal{N}(0, 1)$. So the GPS is $f_{T|X}(t|X) = \exp(-(t - \cos(2\pi X_1))^2/4)/\sqrt{4\pi}$.

The trimming function trims the 1% lower and upper empirical quantiles of each covariate (X_1, X_2 and T), ending up trimming around 5% of the observations. The bandwidth is $C\sigma n^{-\eta}$, where σ is the standard deviation of the variable and $\eta = 0.12$ satisfies the conditions for the asymptotic theorems. The bandwidths are chosen by varying the constants C between 0.5 and 2 to minimize the RMSE. A fourth-order Epanechnikov kernel is used.

Figure 1.1 and Figure 1.2 are for (DGP1) for sample sizes 100 and 1000, respectively. Figure 1.3 and Figure 1.4 are for the more nonlinear (DGP2). The left two panels are the true function $E[Y(t)]$ and the average estimations over 1000 replications. The differences show the biases. For any finite sample, the estimator is biased with the order of $O(h^r)\frac{\partial^r}{\partial t^r}E\big[Y(t)\big]$ for the rth-order kernel, as implied by Remark 1.1. The figures support that the proposed

semi- and nonparametric estimators are more biased at the points when the functions have more curvature. When the sample size increases ($n = 1000$), the bias is improved for my semi- and nonparametric methods. But the bias for the parametric method *HI* remains, especially for the nonlinear (DGP2).

The right panel in the figures are the RMSEs for the four estimators. The nonparametric methods have more variation in small samples, especially at the tails of the treatment variable. For the nonlinear (DGP2), *SP* outperforms *HI* even in small sample size. The proposed estimators work for a rather linear data generating process (DGP1) as well.

The estimation for the median of $Y(t)$ are shown in Figure 1.5 using the semiparametric estimator *SP*. This demonstrates that the semiparametric estimator work well for estimating the distribution of $Y(t)$.

1.7.1.1 Coverage Rate

I consider four methods of constructing the pointwise confidence intervals for $E[Y(t)]$, using the semiparametric estimator *SP*. The first is bootstrap and the second is the multiplier method by Theorem 1.5. The third and fourth are the standard methods by estimating the asymptotic variance using the plug-in sample analogue and Newey's (1994a) Delta method.[18] Then the 95% confidence interval is constructed by $\left[\hat{\theta} - \alpha\sqrt{\widehat{Var}/n}, \hat{\theta} + \alpha\sqrt{\widehat{Var}/n}\right]$, where α is the 97.5% quantile from a standard normal distribution.

[18] Newey (1994b) proposes a "delta-method" variance for the partial-mean kernel estimators. The estimator takes the form,

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\delta}_i - \frac{1}{n}\sum_{j=1}^{n}\hat{\delta}_j + \hat{\phi}_i\right)^2,$$

where $\hat{\delta}_i = \frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{1}_{\{Y_i \le y\}} - \hat{F}_{Y|TX}(y|t, X_j)\right)K_h(t - T_i)K_h(X_i - X_j)\frac{\hat{W}(X_j)}{\hat{f}_{TX}(t,X_j)}$ and $\hat{\phi}_i = \hat{F}_{Y|TX}(y|t, X_i)\hat{W}(X_i) - \frac{1}{n}\sum_{j=1}^{n}\hat{F}_{Y|TX}(y|t, X_j)\hat{W}(X_j)$. Then asymptotic variance is estimated by $\widehat{Var}(\mathbb{G}(y)) = h\hat{V}$. Theorem 4.1 in Newey (1994b) shows $h\hat{V} \to Var(\mathbb{G}(y))$ with the additional assumption $nh_2^{3d_2-1} \to \infty$. The delta-method variance estimator takes into account of the small order terms to achieve more accurate finite-sample approximation than the plug-in estimator. Newey (1994b) interprets δ_i as the first-order effect of the ith observation in the nonparametric second-step regression $\hat{F}_{Y|TX}(y|T, X)$ on the final estimator. Alternatively, in my case, δ_i is estimating (A.5) in the Appendix, where the dominating term is the influence function ψ_{tin}. The smaller-order $\hat{\phi}_i$ is from (A.4) for the Donsker property of the true function, which converges at $\sqrt{n}$-rate.

Figures 1.6 and 1.7 show that the proposed multiplier method and Newey's Delta method outperform the standard plug-in estimator. The proposed multiplier method works reasonably and no additional estimation is needed. So it can be an alternative to the bootstrap which is more computationally intensive, especially for large datasets.

The coverage rates are lower at the points where the finite-sample biases are large. At some points, the coverage rates are not improved as the sample size increases. Similar results are also shown in Flores (2007) by the plug-in method. Constructing confidence intervals with correct coverage rates for the nonparametric estimator is outside the scope of this paper.

1.7.2 Empirical illustration

The following empirical work on the program Families in Action in Colombia is based on joint work with Juan Villa. I illustrate the application of my estimation procedure through this ongoing project and present some preliminary results in this section. Colombia's Families in Action is a conditional cash transfer (CCT) program that gives money to poor families conditional on school attendance and health check-ups of children under 18.[19] CCT programs were first introduced in Brazil and Mexico more than a decade ago. The main objective of CCT programs is to increase the human capital formation of minors and to alleviate current poverty. There is evidence of the effectiveness of CCT programs on a variety of dimensions. We summarize some related literature in Section 1.7.2.1.

We use administrative data and household survey data that includes all beneficiaries since the program started in 2001 in Colombia. By 2009, the program covered approximately 2.8 million households at a cost of around 0.27 percent of GDP. Using this dataset from Families in Action and the continuous treatment estimator proposed in this paper, we contribute to the literature on CCT programs by accounting for the heterogeneity of treatment effects that may arise from variation in the duration of treatment. The treatment variable T is the number of days of participation in the program, converted to years in what follows. The goal

[19] According to the program's operational manual, a household becomes ineligible if: it does not comply with the co-responsibilities of the program for two consecutive periods, it does not withdraw the money for three consecutive periods, or all the children in the household turn 18 years old.

of our analysis is to learn how the length of exposure to the program affects the distributions of household income and education. Our analysis sheds light on foundational questions such as "how long should a household participate in an anti-poverty program?" and "what should be the exit criterion?"

We present the estimation results for household income and also discuss some estimation issues, such as bandwidth choice, in Section 1.7.2.2. We find that the length of exposure (up to 5.5 years) does not significantly affect the distribution of the household income. For those currently staying in the program for $\bar{t}$ years, where $\bar{t} = 1.5, 3, 4.5$ (the $\bar{t}$-year treated groups), we do not find a significant effect either. However, mean income for the one-year treated group is higher than that for the five-year treated group. This suggests there is a difference in the distributions of the characteristics among these treated groups. Since the five-year treated group is mainly composed of those who enroll in 2005 and the one-year treated group is mostly from the 2009-wave, this result might be interesting from policy perspective.

In Section 1.7.2.3, we study two educational outcomes, high-school completion rate and years of education. We focus on ineligible children in a household who are 18 to 28 years old and therefore not required to attend schools by the CCT program. They are exposed to this program because there are eligible children in their households. The school enrollment of one specific child might lead parents to reallocate child work away from the eligible children and to the adult children in the household. We study the spillover effect of the length in the program on the education outcomes of these ineligible children. We find that the displacement effect on high-school completion is more severe for longer exposure up to 3 years. There are no significant effects by extending the program from 3 years to 4.5 years. The high-school completion rate decreases about 2.5% when program exposure is extended from 1.5 years to 3 years. For years of education, we find 4.5-year exposure reduces interquantile range of educational attainment for women, comparing with 1.5-year exposure.

The following estimation results are base on a random sample of 5% of the data. The outcome variables are measured in 2010 for all observations. The pretreatment variables are measured before an individual enters the program, so they are measured in different years for

different individuals. These pretreatment variables include demographics, education, past employment, household, and location characteristics. We assume the unconfoundedness assumption holds based on this large set of observable characteristics. We do not have the non-treated comparison group.

1.7.2.1 Literature Review

Baez and Camacho (2011) discuss short-term (within two years) evaluations of Colombia's Families in Action, which indicate that the program leads to higher consumption, higher spending on nutritious food, more children sent to school, more time devoted to studying, and infants growing taller while having fewer health issues. Most evaluations of CCT programs focus on short-term impacts because of data availability. Baez and Camacho (2011) study the long-term impacts of Colombia's Families in Action on human capital. They find the program helps children, particularly girls and beneficiaries in rural municipalities, accumulate more years of education. For the well-known CCT program Progresa/Oportunidades in Mexico, Behrman et al. (2011) find positive long-run (five to six years) impacts on schooling and work. However, Rodrguez-Oreggia and Freije (2012) do not find significant long-term effects (at most nine years) on labor outcomes, such as employment, wages, and inter-generational occupational mobility.

Agüero et al. (2010) and Ibarraran and Villa (2010) take the length of exposure to the program as a continuous treatment and use the generalized propensity score methodology by Hirano and Imbens (2004). Agüero et al. (2010) note that continuous treatment estimation is important to study long-term impacts which cannot be reliably estimated by simply projecting out short-term rates of impact. (Agüero et al., 2010) find positive impacts of the South African Child Support Grant on the child nutritional outcome height-for-age. Ibarraran and Villa (2010) find that the probability of employment for 14-year-old kids is greatest if they participated at least three years in Mexico's Oportunidades.

The model in Ferreira and Schady (Ferreira and Schady) predicts child-specific CCT programs will unambiguously increase school enrollment among eligible children, because

of the positive income, substitution, and displacement effects of the cash transfer. The displacement effect comes from the reallocation of labor or child care duties away from eligible children to their ineligible siblings. The net spillover effects of the CCT program on the ineligible siblings depend on the magnitude of the positive income effect and the negative displacement effect. Ferreira and Schady (Ferreira and Schady) find no evidence of such spillovers in the CESSP Scholarship Program in Cambodia. Evaluating the Conditional Subsidies for School Attendance in the city of Bogota in Colombia, Barrera-Osorio et al. (2008) find negative spillovers of the program on the education of ineligible children. They find evidence of lower school attendance and more labor market work for an ineligible child living with an eligible sibling, compared to a child with a similarly sibling in an untreated household. Behrman et al. (2011) also find increases in work for older girls in Mexico's Oportunidades. Baez and Camacho (2011) find no spillover effect on the probability of graduating from high school in Colombia's Families in Action. In contrast to the above findings, which use a binary treatment methodology, we investigate the spillover effects of the length of exposure to the program on high-school completion rates and education levels in Section 1.7.2.3.

1.7.2.2 Household Income

We first study the potential average household income $Y(t)$ with respect to the length of exposure in the program t. We focus on the potential treatment values t smaller than 5.5 years, which is at the 87th percentile of the observed length of exposure T. Table 1.8 shows the descriptive statistics. After selecting the common support[20] and trimming the boundary of the continuous covariates, we end up with 8,851 households (around 12% are dropped).

[20]We follow Flores et al. (2012) to select the sub-sample satisfying the common support assumption. Consider a set of potential treatment levels to be estimated, $\mathcal{T}$. For each potential treatment level $t \in \mathcal{T}$, find the maximum and minimum of the GPS $V_i \equiv \hat{f}_{T|X}(t|X_i)$ among the sample $i = 1, ..., n$. Denote the maximum and minimum by V_{max} and V_{min}, respectively, for each $t \in \mathcal{T}$. Then select the sub-sample by $\{i : \max_{t\in\mathcal{T}}\{V_{min}\} \leq V_i \leq \min_{t\in\mathcal{T}}\{V_{max}\}\}$.

We consider two estimators, the *SP* estimator developed in this paper and the *HI* estimator of Hirano and Imbens (2004), which are described above.[21] The second-step kernel estimation for *SP* uses the second-order Epanechnikov kernel satisfying Assumption 1.10. Figure 1.8 presents the estimation of the mean of potential income ($E[Y(t)]$). The left panel compares the *SP* estimator with *HI poly-2* from Hirano and Imbens (2004). In the right panel, we use a more flexible *HI poly-6* with a sixth-degree polynomial in the second-step regression.[22] It shows *HI poly-6* is closer to our semi-parametric estimation. This implies the linear model using *HI poly-2* might be misspecified and overly parsimonious.

Figure 1.9 shows our estimation is robust to different bandwidth choices of $C\sigma n^{-0.3}$, where C is some constant and σ is the standard deviation. Although a theoretical method for choosing the bandwidth is outside of the scope of this paper, an eye-ball metric could suggest a reasonable choice. The simulation and theory results above suggest a smaller bandwidth leads to a smaller bias and larger variance. We choose the bandwidth 0.7 years for the treatment variable throughout this estimation.

Figure 1.10 presents the mean potential income $E[Y(t)|T = \bar{t}]$ for those currently staying in the program for $\bar{t}$ years. The treatment effects on the treated have similar patterns as the treatment effects for the whole population, i.e., $E[Y(t)|T = \bar{t}]$ does not significantly change over t, for $\bar{t} = 1, 3, 5$. Now consider the counterfactual experiment of changing the distribution of the characteristics as discussed in Section 6. The difference between the mean 1-year potential income for the one-year and five-year treated group $E[Y(1)|T = 1] - E[Y(1)|T = 5] \approx 22,000$ (pesos) is the average income loss if the one-year treated group had the same distribution of characteristics as the five-year group. On the other hand, $E[Y(5)|T = 1] - E[Y(5)|T = 5] \approx 25,000$ (pesos) is the average income gain if the five-year group had the same distribution of characteristics as the one-year group.

[21] In the first-step estimation, the generalized propensity score is modeled by a normal distribution. We've tried a log-normal model, but the results are not much different. A balancing test as in Kluve et al. (2012) for specifying the GPS will be performed in a separated paper.

[22] For an empirical application on evaluating a German job-training program, Kluve et al. (2012) use a third-degree polynomial for a flexible specification.

In addition to the mean effects, Figure 1.11 shows the distributions for the one-year and five-year treated groups. In the left panel, $F_{Y(1)|T}(y|5) - F_{Y(1)|T}(y|1)$ is the change in the distribution of the one-year potential income if the one-year group had the same distribution of observable characteristics as the five-year group. It suggests that the distribution of one-year potential income for the one-year treated groups $F_{Y(1)|T}(y|1)$ first-order stochastically dominates that for the five-year treated group $F_{Y(1)|T}(y|5)$, although a stochastic dominance test is not yet performed. In the right panel, the distributions for the five-year potential outcome $Y(5)$ share a similar result.

As a caveat, we note that studying income as the potential outcome might suffer from the reverse causality problem so that the unconfoundedness assumption might not hold. The household might withdraw from the program, because their income is so low that they need their children to drop out of school and work.

1.7.2.3 Education for Ineligible Children

To evaluate the high-school completion rate, we focus on the ineligible children who did not finish high school before their households enrolled in the program. This group is 83.35% (out of 7,675) in the population of the 18-28 year-old children for females and 82.94% (out of 9,518) for males. Table 2 shows the descriptive statistics. The outcome variable is an indicator for high school graduation, i.e., years of education larger than or equal to 11. The estimation results in Table 1.4 imply that the high school completion rate for 1.5-year exposure to the program is about 2.5% higher than that for 3-year exposure, for both female and male. This suggests that the displacement effect is larger by extending the program from 1.5 years to 3 years. However, there is no significant effect of increasing the length of exposure from 3 years to 4.5 years. The completion rates for the treated groups share similar patterns with the overall population.

Figure 1.13 presents the distribution of the potential education levels for all the ineligible children including high-school graduates. The distribution of education for 1.5-year exposure $F_{Y(1.5)}$ appears to stochastically dominate the distribution of education for 3-year exposure

$F_{Y(3)}$, for males. However, we do not perform a test for stochastic dominance. These results imply there is a negative spillover effect of extending the program from 1.5 years to 3 years. For females, this negative spillover effect is only for the upper quantiles (larger than 80%). It is interesting to observe that the displacement effect is alleviated by extending the program from 3 years to 4.5 years. For females, the low end of the distribution is significantly lower for the 4.5-year exposure, compared with the 1.5-year exposure. It suggests that the inter-quantile difference for potential education level with 4.5-year exposure is smaller than that with 1.5-year potential education (comparing $F_{Y(4.5)}$ and $F_{Y(1.5)}$). This implies that longer exposure to the program improves the inequality of education attainment for female ineligible children. The mean of potential years of education does not vary significantly with the length of exposure. The treated groups have similar patterns to the whole population, so the results are omitted.

1.8 Conclusion

I derive a stochastic expansion showing how the presence of generated regressors affects the limiting behavior of the three-step nonparametric estimator of a partial mean process (1.1). I explicitly estimate the mean and quantile structural functions for the overall population and the treated group. The uniform expansion and weak convergence theorems derived in this paper are readily applied to many inequality measures, such as the Theil index and coefficient of variation in Firpo and Pinto (2011). My results can also be extended to test for stochastic dominance, such as the Kolmogorov-Smirnov-type test in Rothe (2010).

I adopt fixed trimming functions in the estimation procedure in order to focus on the influence of estimating the generated regressors. For future work, random trimming functions are desirable to estimate for the whole population, instead of the subpopulation chosen by fixed trimming functions of the observables. Escanciano et al. (2012) introduce a stochastic expansion, that is uniform in the weights, the generated regressors, and a random bandwidth, for sample means of weighted semiparametric regression residuals. Their methods could be modified and used in my setup.

Another extension is to estimate the location and size of the optimal dose for the distributional features of the potential outcome, defined by Hadamard-differentiable functionals on the counterfactual distribution. A policy maker might be interested in the treatment level that minimizes some inequality measure, for instance, the interquantile range. Flores (2007) estimates the optimal dose of the mean dose response function and derives the limit theory.

Counterfactual Distributions. I derive the limit theory when the weight function is estimated for the distribution function on the treated $F_{Y(t)|T}(y|\bar{t})$ for those currently treated or choosing the treatment level $\bar{t}$. In Section 1.2.2, I discuss the interpretation of the treated effect on the treated as the counterfactual effect of a policy intervention by either changing the conditional distribution of outcome given characteristics or changing the characteristics distribution. It is interesting to note that a more general usage of this weight function would allow me to consider a wider variety of counterfactual objects than what was discuss in Section 1.2.2. The counterfactual distribution of the potential outcome can be characterized by the weight function, which is a ratio of the *counterfactual density* and the status-quo density of the observable characteristics, f_{X^*}/f_X. More explicitly, define the counterfactual cdf of the potential outcome $Y(t)$ for the population whose characteristics are distributed as the counterfactual distribution F_{X^*} by

$$F^*_{Y(t)}(y) \equiv \int F_{Y(t)|X^*}(y|x)\ dF_{X^*}(x) = E\Big[F_{Y(t)|X}(y|x)\frac{f_{X^*}(x)}{f_X(x)}\Big]$$

assuming $F_{Y(t)|X} = F_{Y(t)|X^*}$. The policy effect of changing the distribution of the observables from F_X to F_{X^*} is $F^*_{Y(t)} - F_{Y(t)}$.

The counterfactual density of the characteristics f_{X^*} can be deterministic by a policy intervention or based on the treatment variable. Rothe (2010) and Chernozhukov et al. (2013) discuss various choices for the counterfactual distribution F_{X^*}: a different subpopulation corresponding to a different demographic group, geographic region or time period. Or $X^* = \pi(X)$ is a deterministic function of X. Comparing with Rothe (2010) who studies the unconditional effects by averaging over all the covariates including treatments, the causal

effects by the potential outcome $Y(t)$ reveal more local information with respect to a fixed value of the endogenous treatment variables of interest.

By the conditional independence and common support assumptions, $F^*_{Y(t)}(y)$ is identified by $E\Big[F_{Y|TX\Lambda}(y|t,X,\Lambda)\cdot W(X)\Big]$, where $W(X)=\frac{f_{X^*}(X)}{f_X(X)}$. This regression is a partial mean process with generated regressors, which is estimated by the procedure proposed in this paper assuming the weight function is known. This re-weighting functions of relative densities is seen in DiNardo et al. (1996) to estimate counterfactual densities.

Empirics. I illustrate the usefulness of my proposed estimator by evaluating a conditional cash transfer program in Colombia, which is an ongoing project with Juan Villa. We analyze how the distributions of income and education outcomes respond to the length of exposure to the program, which is taken as a continuous treatment variable.

As richer and more detailed data is available, the proposed nonparametric estimator could be useful to analyze the continuous treatment effects for various economic outcomes. The proposed estimator is also applicable to nonseparable triangular models, as in Engel curve analysis.

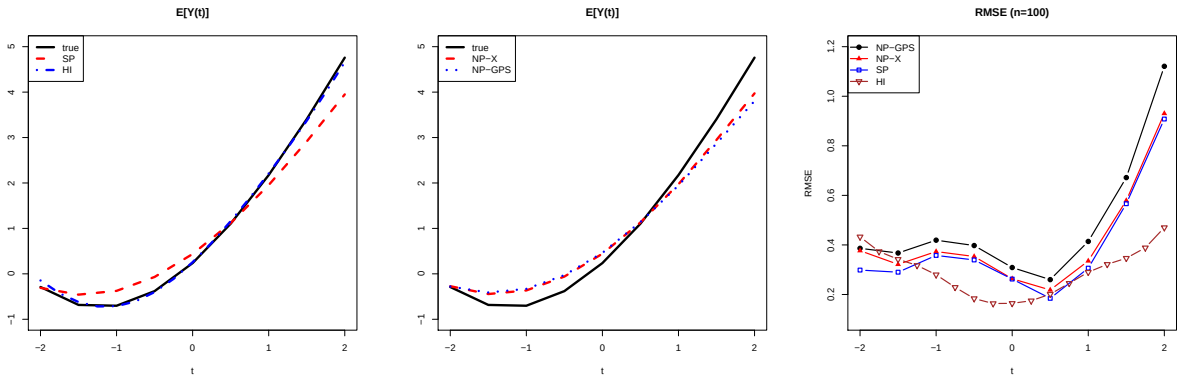

Figure 1.1 (DGP1) $n = 100$. The left two panels show the average estimation over 1000 replications and the true $E[Y(t)]$. The difference indicates the bias. The right panel shows the root-mean-square errors (RMSE).

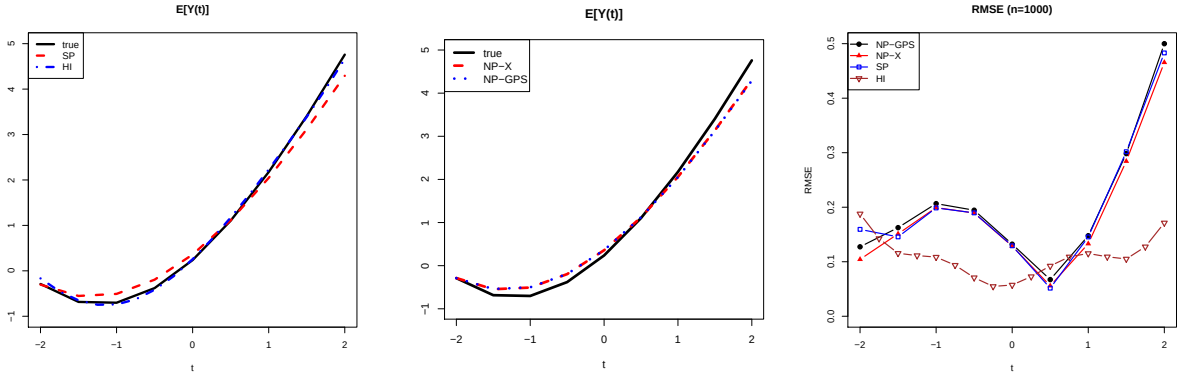

Figure 1.2 (DGP1) $n = 1000$. The left two panels show the average estimation over 1000 replications and the true $E[Y(t)]$. The right panel shows the root-mean-square errors (RMSE).

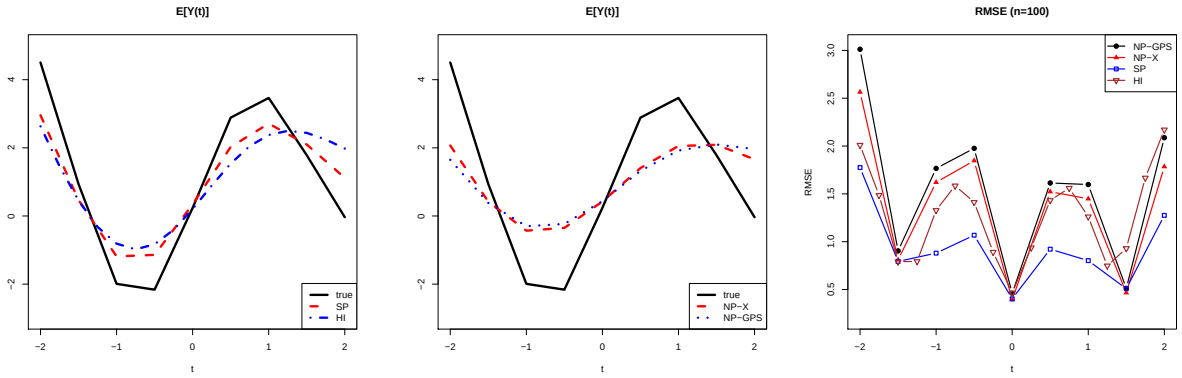

Figure 1.3 (DGP2) $n = 100$. The left two panels show the average estimation over 1000 replications and the true $E[Y(t)]$. The difference indicates the bias. The right panel shows the root-mean-square errors (RMSE).

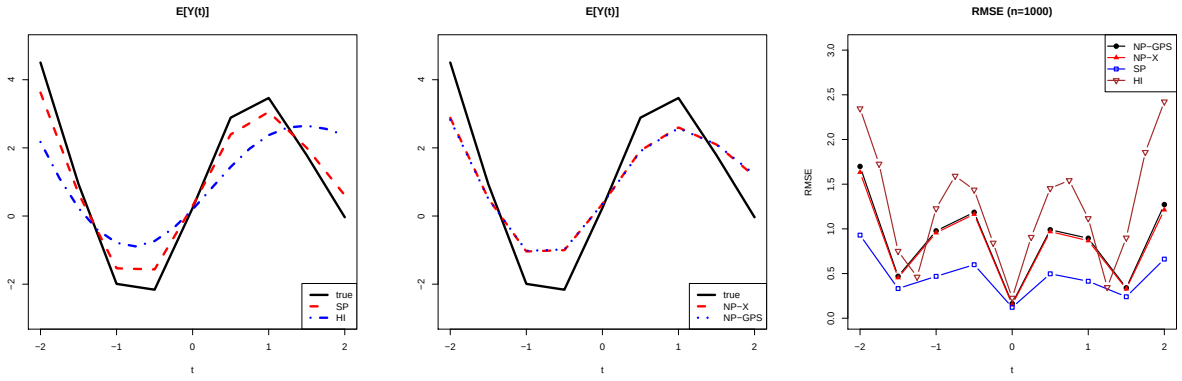

Figure 1.4 (DGP2) $n = 1000$. The left two panels show the average estimation over 1000 replications and the true $E[Y(t)]$. The right panel shows the root-mean-square errors (RMSE).

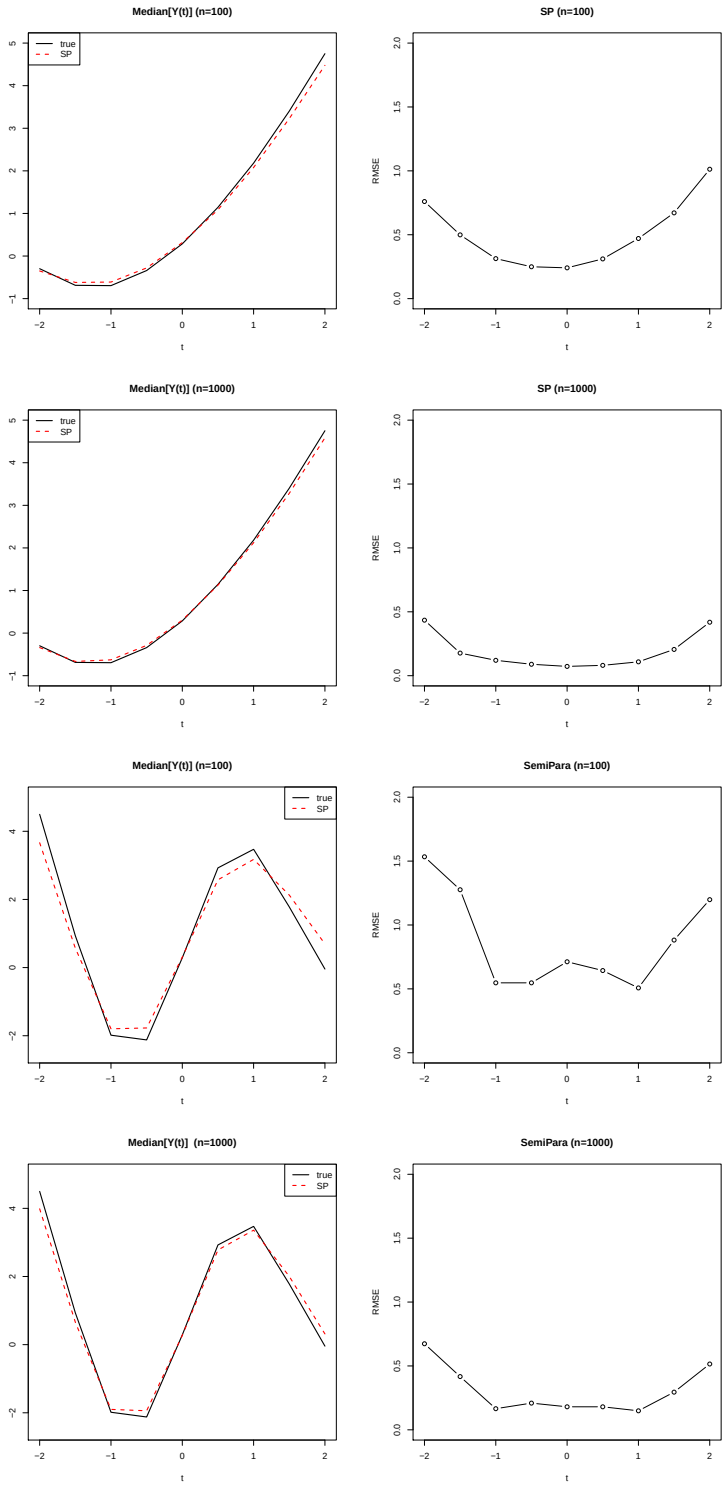

Figure 1.5 median(Y(t)). The top fours panels are for (DGP1) and the bottom fours are for (DGP2). The left panels show the average estimation over 1000 replication and the true $median[Y(t)]$. The right panels are the root-mean-square errors (RMSE).

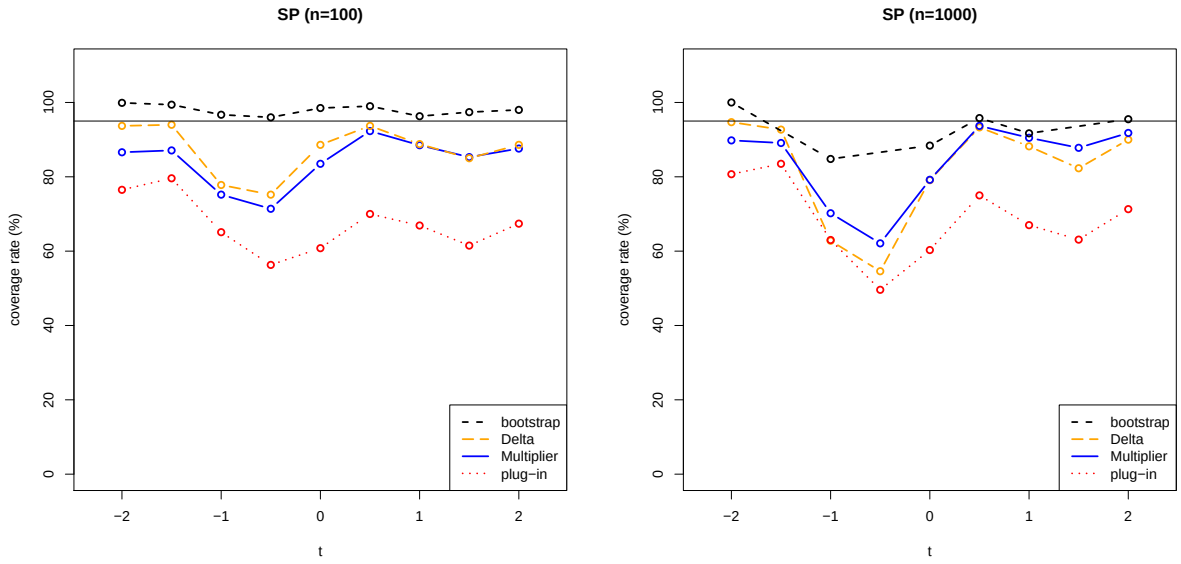

Figure 1.6 (DGP1) Coverage rates of the 95% confidence intervals for the mean $E[Y(t)]$ by the *SP* estimator

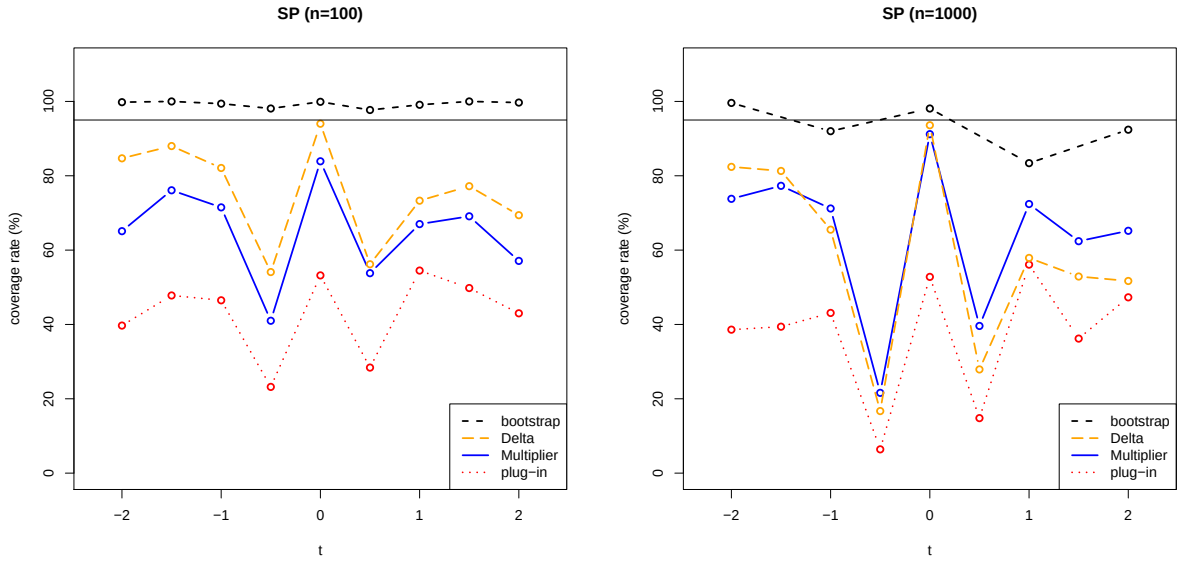

Figure 1.7 (DGP2) Coverage rates of the 95% confidence intervals for the mean $E[Y(t)]$ by the *SP* estimator

	Mean	Std. dev.	Min.	25%-quant.	Median	75%-quant.	Max.
Treatment (years)	3.414	2.144	0.678	2.048	2.960	3.996	8.829
Income (pesos)	107,300	212,150.8	0	33,330	72,180	140,000	1,500,000

Table 1.2: Descriptive statistics

The sample contains 8,851 households whose head is married male, working, not disabled, and not attending school. The households own their dwellings, and there is only one household living in each dwelling.

	Mean	Std. dev.	Min.	25%-quant.	Median	75%-quant.	Max.
Female	$n = 5,416$						
Treatment (years)	3.5490	2.0829	0.6804	2.8720	2.9850	3.2970	8.8310
Education (years)	2.094	2.9785	0	0	0	4	16
Male	$n = 7,895$						
Treatment (years)	3.3320	1.9737	0.6804	2.7190	2.9690	3.2530	8.8290
Education (years)	2.741	3.3659	0	0	1	5	17

Table 1.3: Descriptive statistics for ineglible children who did not finish high school before treatment

	Female			Male		
Treatment (quantile)	1.5 (15.8%)	3 (53%)	4.5 (81%)	1.5 (20%)	3 (56%)	4.5 (83.7%)
overall	4.97 (3.64, 6.27)	2.33 (1.79, 2.85)	2.42 (0.98, 3.84)	7.95 (6.63, 9.31)	5.17 (4.46, 5.85)	3.49 (1.85, 5.07)
1.5-yr treated	5.25 (3.90, 6.58)	2.52 (1.92, 3.13)	2.61 (0.97, 4.26)	8.68 (7.27, 10.07)	5.33 (4.54, 6.15)	4.14 (2.20, 6.11)
3-yr treated	5.03 (3.72, 6.31)	2.38 (1.85, 2.92)	2.45 (0.95, 3.90)	8.03 (6.71, 9.36)	5.22 (4.53, 5.92)	3.53 (1.91, 5.18)
4.5-yr treated	4.83 (3.47, 6.15)	2.24 (1.73, 2.74)	2.31 (0.97, 3.69)	7.44 (6.09, 8.76)	5.08 (4.43, 5.72)	3.02 (1.63, 4.39)

Table 1.4: High-school completion rates (%) for inegible children who did not finish high school before treatment

The 95% confidence intervals are in the parentheses, calculated by the multiplier method in Section 5.1. The bandwidth for the treatment is 0.79 year for female and 0.69 year for male. The parentheses below Treatment indicate sample quantiles of the observed years of exposure, for example, there are 15.8% of females whose length of exposure to the program is smaller than 1.5 years.

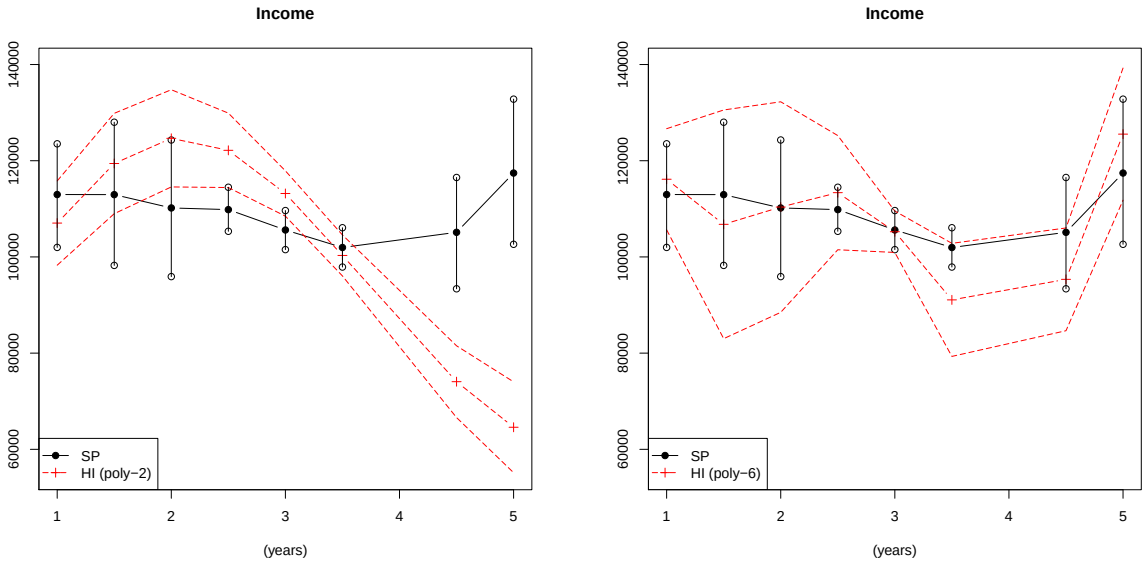

Figure 1.8 The mean potential household income with respect to years of exposure in the program, $E[Y(t)]$. In the left panel, *HI poly-2* uses second-order polynomial regression for the second-step regression on the treatment and the generalized propensity score. In the right panel, *HI poly-6* uses sixth-order polynomial regression in the second step. The pointwise confidence intervals are calculated by the multiplier method proposed in Section 5.1 for *SP* and bootstrap for *HI*.

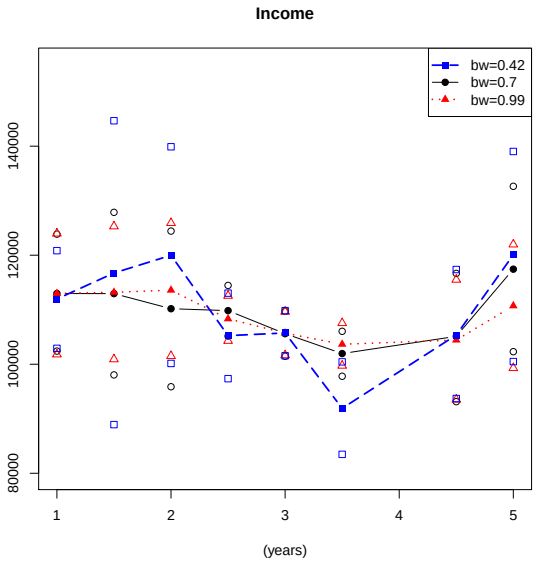

Figure 1.9 Bandwidth robustness check for the *SP* estimator. The bandwidths for the treatment variable, years in the program, are 0.42, 0.7, and 0.99 years for choosing the constants $C = 3, 5, 7$, respectively. The hollow symbols represent the confidence intervals, calculated by the multiplier method. The bandwidth 0.7 years is chosen to balance the bias and variance by an eyeball metric.

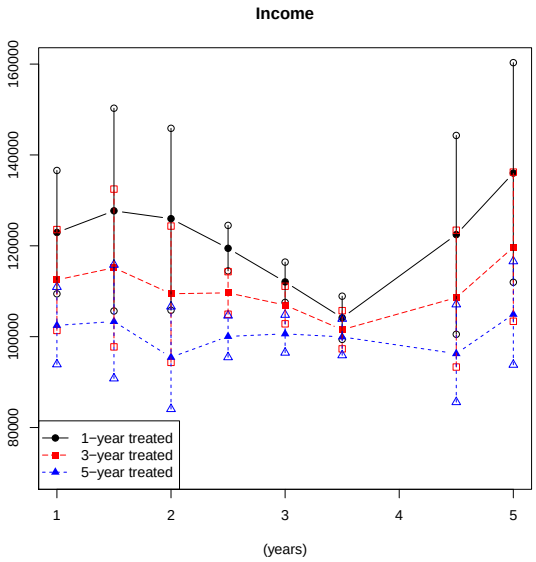

Figure 1.10 The treatment on the treated $E[Y(t)|T = \bar{t}]$: the mean potential income for those currently being treated for $\bar{t}$ years, where $\bar{t} \in \{1, 3, 5\}$.

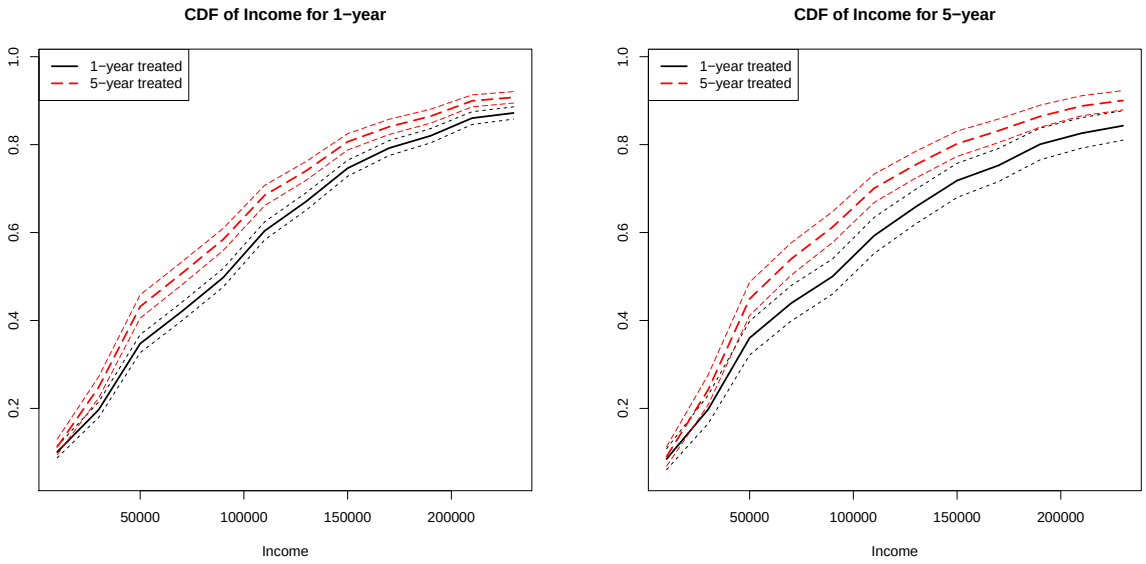

Figure 1.11 The left panel shows the distributions of the potential income of one-year exposure for the one-year treated group $F_{Y(1)|T}(y|1)$ (black solid line) and for the five-year treated group $F_{Y(1)|T}(y|5)$ (red dashed line), respectively. The right panel shows the distributions of the potential income of five-year exposure for the one-year treated group $F_{Y(5)|T}(y|1)$ (black solid line) and for the five-year treated group $F_{Y(5)|T}(y|5)$ (red dashed line), respectively.

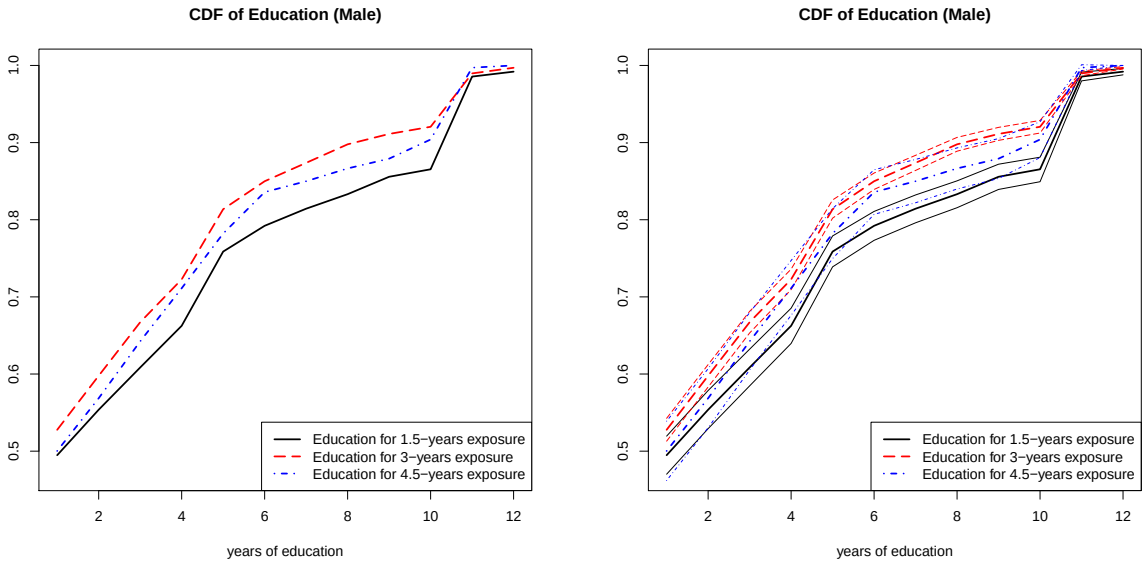

Figure 1.12 The distributions of potential education levels $F_{Y(t)}$ for $t = 1.5, 3, 4.5$ years of exposure for male ineligible (between 18 to 28 years old) children in a household. The confidence intervals calculated by the multiplier method are added in the right panel. The bandwidth for treatment is 0.66 years. After selecting for the common support and trimming the boundaries, the estimation is based on 7,990 (16.6% dropped) observations.

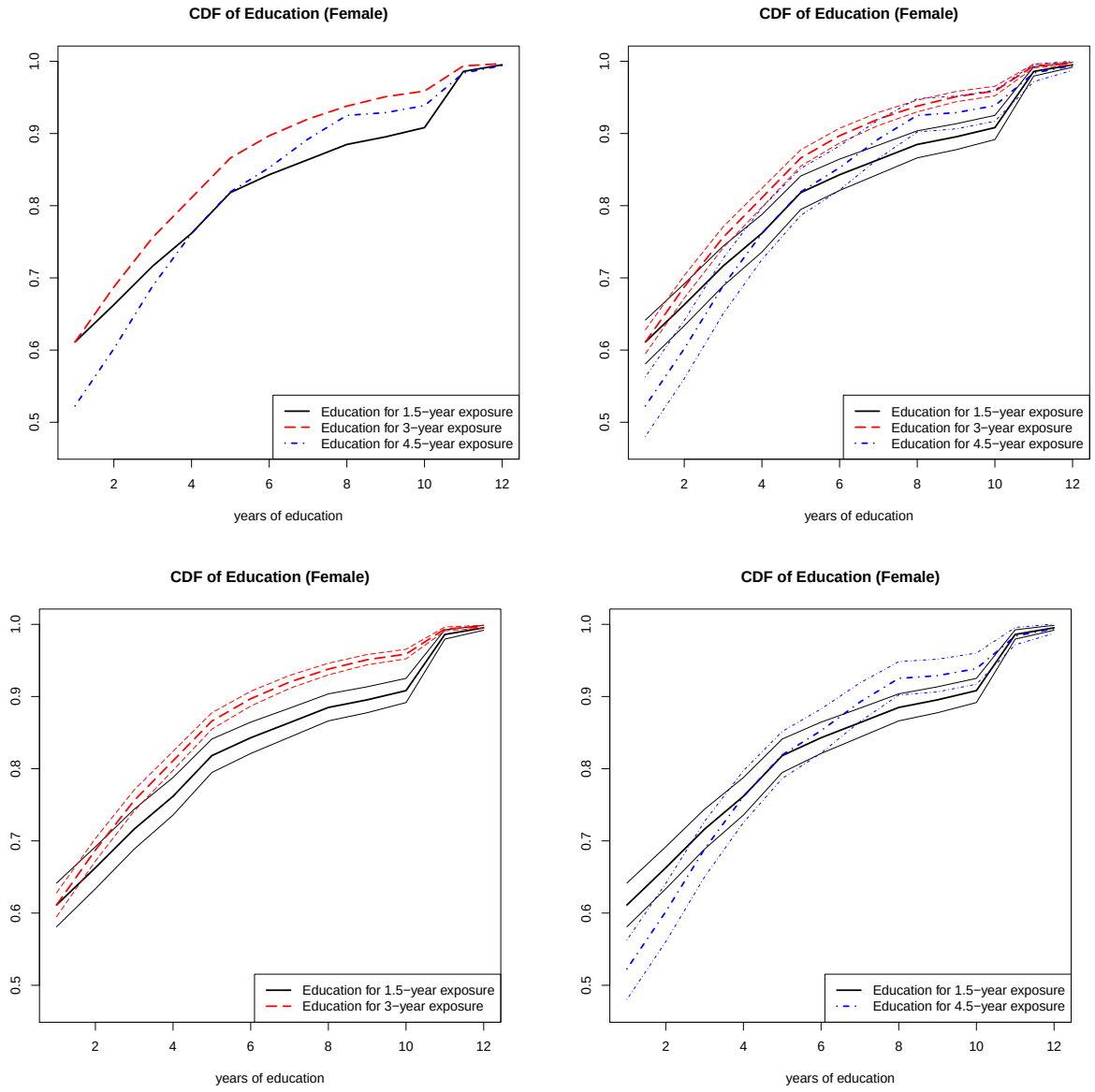

Figure 1.13 The distributions of potential education levels $F_{Y(t)}$ for $t = 1.5, 3, 4.5$ years of exposure for female ineligible (between 18 to 28 years old) children in a household. The estimations are shown in the top left panel. The confidence intervals based on the multiplier method are added in the other panels. The bottom panels display only two of the estimations for clarity. The bandwidth for treatment is 0.76 years. After selecting for the common support and trimming the boundaries, the estimation is based on 6,057 (21% dropped) observations.

Chapter 2

Nonparametric Density-Weighted Average Quantile Derivative

2.1 Introduction

The average quantile derivative (AQD) is the mean of the partial derivatives of the conditional quantile function (CQF), where the function forms of the distributions and the CQF are not specified and will be estimated nonparametrically. The proposed estimator is shown to be $\sqrt{n}$-consistent and asymptotic normal. Therefore, the AQD offers sensible and economical summary statistics for the marginal effect of the covariates on the CQF and can be viewed as the nonparametric quantile regression (QR) coefficient.

The linear QR is introduced by the seminal work of Koenker and Bassett (1978), where the CQF is specified to be linear in the covariates. However, it is known that the linear CQF induces the quantile crossing problem which implies the linear model is implicitly misspecified. For example, Angrist et al. (2006) study the approximation property of the liner QR under misspecification, and Chernozhukov et al. (2010) propose a method to rearrange the original estimated non-monotone curve into a monotone curve. Then the AQD is of interest to serve as a nonparametric summary statistic that is robust to misspecification and conveys the information of the marginal effect of the covariates on the CQF. When the economic theory implies some index structure or the researchers would like to spacify some semiparametirc models, the AQD also identifies the parameter in the semiparametric single-index and partial linear models (Chaudhuri et al. (1997), Khan (2001), Lee (2003),

Wu et al. (2010), Kong and Xia (2012)). In the growing econometrics literature of the nonparametric nonseparable structural models, the derivative of the CQF can identify useful features such as the structural derivative and continuous quantile treatment effect, for example, Chesher (2003), Chernozhukov and Hansen (2005), Hoderlein and Mammen (2007), Matzkin (2007). Therefore, the AQD carries a causal interpretation of the economic models and can potentially be applied to welfare and policy analysis.

Let the τth CQF of the dependent variable Y given the q-dimensional continuous regressor X be $Q_\tau(Y|X) := \inf\{y : F_Y(y|X) \geq \tau\}$, where $F_Y(y|X)$ is the conditional cumulative distribution function (cdf) of Y given X and $\tau \in (0,1)$. Assume $Q_\tau(Y|X)$ is continuously differentiable in X almost surely, and the density function of X, $f(X)$, is sufficiently smooth. I define the *nonparametric QR parameter* of interest to be the *density-weighted average quantile derivative* (AQD):

$$\begin{aligned}\beta(\tau) &= E[\nabla Q_\tau(Y|X) \cdot f(X)] \\ &= Q_\tau(Y|X) \cdot f^2(X)\Big|_{\partial\mathcal{X}} - 2E[Q_\tau(Y|X) \cdot \nabla f(X)] = -2E\Big[Q_\tau(Y|X) \cdot \nabla f(X)\Big] \qquad (2.1)\end{aligned}$$

where the second equality follows by integration by parts. And the third equality is by assuming $\lim_{X\to\partial\mathcal{X}} Q_\tau(Y|X)f^2(X) = 0$, where $\partial\mathcal{X}$ is the boundary of the support of X, $\mathcal{X}$. [1] I propose a two-step kernel based estimator: in the first step, the unknown functions $Q_\tau(Y|X)$ and $\nabla f(X)$ are estimated by a nonparametric kernel method; in the second step, the expectation is replaced by its sample analogue with a stochastic trimming function for the small density near the boundary. The estimation is direct without iterative optimization algorithms for a nonsmooth objective function. I first provide a Bahadur-type linear representation of the CQF estimator using the uniform convergence results of kernel estimation in Hansen (2008). Then the limit theory is derived for the final estimator.

[1] This assumption excludes a compact support of X. A similar assumption has also been made for the density-weighted average mean derivatives in Powell and Stoker (1996).

The weighted average derivative for quantile regression has been defined by Chaudhuri et al. (1997) as

$$\beta_W(\tau) \equiv E\left[\nabla Q_\tau(Y|X) \cdot W(X)\right] \tag{2.2}$$

$$= -E\Big[Q_\tau(Y|X) \cdot \Big(\nabla W(X) + \frac{\nabla f(X)}{f(X)} W(X)\Big)\Big] \tag{2.3}$$

where the known weighting function $W(X)$ is sufficiently smooth with a compact support within the interior of $\mathcal{X}$. Chaudhuri et al. (1997) propose estimators for (2.2) and (2.3) using local polynomial estimators for the unknown functions. Lee (2003) proposes an average quantile regression based on (2.2) using a local polynomial estimator for the coefficient in $\nabla Q_\tau(Y|X)$ in a partial linear model. Recently, Chernozhukov et al. (2011) develop a nonparametric QR series framework and perform inference on the entire CQF and its linear functional, covering the unweighted average quantile derivative in (2.2) defined on a compact support. This paper studies a kernel-based estimator for (2.3) with two distinct features from the previous work: (1) a stochastic trimming function is involved to estimate the AQD defined on the whole support, which can be unbounded but has zero density at the boundary, and (2) a density weight $W(X) = f(X)$ is estimated nonparametrically. The two features involve different technical issues described in the following.

The weighting function $W(X)$ works as a trimming function to avoid "the denominator problem" for nonparametrically estimating the CQF when $f(X)$ is small near the boundary. The weighting function $W(X)$ in Chaudhuri et al. (1997) removes the tail region in the support of X, so their $\beta_W(\tau)$ is a different object from the AQD which is defined on the whole support. On the other hand, a fixed trimming function will not affect consistently estimating the coefficients in the semiparametric single index and partial linear models. Lee (2003) concerns the optimal weight to estimate the partial linear coefficient efficiently and his estimator involves a fixed trimming function.

To estimate the AQD defined on the whole support and to overcome the denominator problem, I use a stochastic trimming function $\mathbf{1}_{\{\hat{f}(X_i)>\delta_n\}}$, which is an indicator of the estimated density larger than a trimming bound δ_n. This positive sequence δ_n converges to

zero at some specific rate as $n \to \infty$ which restricts the tail of the covariate distribution converging to zero slowly enough. Hardle and Stoker (1989), Lavergne and Vuong (1996), among others, use the similar stochastic trimming; more detail will be discussed in the later section.

I choose the density weight following the average mean derivative (AMD) introduced by Powell et al. (1989), which has been widely studied in econometrics literature; for example, Hardle and Stoker (1989), Powell and Stoker (1996), Cattaneo et al. (2010). For the mean case, the law of iterated expectations simplifies the expression to $-2E[Y\nabla f(X)]$, which does not suffer from the denominator problem and avoids the trimming function, comparing with the unweighted AMD in Hardle and Stoker (1989). Although the law of iterated expectation does not apply to the quantile case, choosing the density weight has several advantages: Comparing with the unweighted estimator, the density-weighted estimator requires weaker assumption on the tail of the covariate distribution from the stochastic trimming function. I will show in section 2.4 that the density of the covariates near the boundary is allowed to converge to zero at a faster rate. Comparing with the estimator of (2.3) in Chaudhuri et al. (1997), the estimated density weight allows weaker smooth assumptions on the unknown functions. In addition, the density of the covariate is a natural data-dependent choice of the weighting function. The population with higher covariate density could be of particular empirical relevance and importance.

Chaudhuri et al. (1997) do not address the choice of the weighting function $W(X)$, which could restrict its empirical application in economics. My asymptotic theory shows that replacing $W(X)$ by an estimated $f(X)$ does not give the same asymptotic covariance matrix as Chaudhuri et al. (1997) derive for $W(X) = f(X)$, i.e., when choosing the estimated $f(X)$ as the weighting function in practice, Chaudhuri et al. (1997) do not give correct statistical inference. Although I will not approach the direction of alternative weighting functions, the optimal weight in terms of efficiency is concerned in semiparametric models, for example, the average mean derivative in the single index model in Newey and Stoker (1993) and the partial linear quantile regression model in Lee (2003).

The rest of the paper is organized as follows. In Section 2.2, I illustrate some applications of the AQD on econometrics models in the literature. In Section 2.3, the estimator for the density-weighted AQD is constructed by a first-step nonparametric estimation for the unknown functions which is then plugged into the sample-analogue in the second step. In Section 2.4, I first show a uniform linear or Bahadur representation of the nonparametric kernel-based CQF estimation. Using the U-statistics theory, the estimator for the AQD is asymptotically linear and hence enjoys the parametric convergence rate, being $\sqrt{n}-$consistent and asymptotically normal. A consistent estimation for the asymptotic covariance matrix is suggested. In Section 2.5, I compare the proposed AQD estimator with the AMD estimator in Powell et al. (1989), the linear QR estimator in Koenker and Bassett (1978), and the OLS for the semiparametric partial linear and single index models. All assumptions and proofs are in Appendix.

2.2 Econometrics examples

The sth component of average derivative $\beta(\tau)$ measures the marginal response of the sth covariate on the τth conditional quantile of Y. In addition to the statistical interest of quantile regression, the AQD can be motivated by econometric models. The AQD identifies coefficients in semiparametric partial linear and single index models, which have been widely studied in mean regression, because it achieves dimension-reduction and relaxes restrictive parametric assumptions. The following examples also demonstrate how the AQD captures informative features of general nonparametric structural models under the conditional independence assumption.

Example 1 (Semiparametric partial linear model)

$Y = X_1'\beta_0 + \phi(X_2) + \epsilon$. The AQD identifies the coefficient β_0 up to scale. Lee (2003) proposes an efficient weighted average quantile regressor, which is similar to my average derivative estimator.

Example 2 (Semiparametric single index model)

Following Chaudhuri et al. (1997), I consider the nonseparable single-index model $Y = \phi(X'\beta_0, \epsilon)$, where ϵ is an unobserved stochastic term and ϕ is an unknown function strictly increasing in the second argument. [2] Then $Q_\tau(Y|X) = \phi(X'\beta_0, Q_\tau(\epsilon|X))$ by the equivariance property of quantiles. Assume quantile independence: $Q_\tau(\epsilon|X) = Q_\tau(\epsilon)$ is constant free from X, which allows heteroskedasticity from possible dependence between X and ϵ. Then the AQD is $\beta(\tau) = \beta_0 \cdot E[\phi_1(X'\beta_0, Q_\tau(\epsilon)) \cdot f(X)]$, where ϕ_1 is the partial derivative of ϕ with respect to the first argument. That is, the density-weighted AQD $\beta(\tau)$ identifies the index β_0 up to scale. Further normalizing ϵ by $Uniform[0,1]$, the structural function is identified as $Q_\tau(Y|X) = \phi(X'\beta_0, \tau)$. So the structural function ϕ can be estimated in the second step by a nonparametric quantile regression of Y on the one-dimensional index $X'\hat{\beta}$. This specification includes many models as special cases:

Separable single-index model $Y = \phi(X'\beta_0) + \epsilon$. By quantile independence and normalization $Q_\tau(\epsilon|X) = 0$, $Q_\tau(Y|X) = \phi(X'\beta_0)$. Ichimura and Lee (2010) study an M-estimator by estimating the link function using a local linear quantile regression in the first step. Wu et al. (2010) and Kong and Xia (2012) propose an iterative algorithm and an adaptive estimation procedure. When the linear CQF is correctly specified, i.e., there exists $\beta_0 \in R^q$ such that $Q_\tau(Y|X) = X'\beta_0$ almost surely, the average derivative $\beta(\tau) = \beta_0 \cdot E[f(X)]$.

Censored Tobit model In Powell (1986), $Y = Y^* \mathbf{1}_{\{Y^* \geq 0\}}$ and the unobserved latent variable $Y^* = \phi(X'\beta_0) + \epsilon$. Then $Q_\tau(Y|X) = \max\{0, Q_\tau(Y^*|X)\}$. β_0 is identified up to scale by the AQD under quantile independence, if $Q_\tau(Y^*|X)$ is positive with high probability.

Selection model $Y = X_1'\beta_1 + \epsilon_1$, X_1, and X_2 are observed only if the unobserved $Z_2^* = X_2'\beta_2 + \epsilon_2 > 0$. By assuming homoskedasticity: (ϵ_1, ϵ_2) is independent of (X_1, X_2), $Q_\tau(Y|X_1, X_2, Z_2^* > 0) = X_1'\beta_1 + Q_\tau(\epsilon_1|Z_2^* > 0)$. Then the AQD identifies the structural parameter β_1 and the selection parameter β_2 up to scale, if X_2 has no varaibles in common with X_1. If $X_1'\beta_1$ and $X_2'\beta_2$ are the same, then it is the truncated Tobit model, as discussed

[2] Khan (2001) develops a rank estimator if ϕ is monotonic in both the index and error.

in Stoker (1986).

Example 3 (Nonseparable structural model with monotonicity)
As noted in the survey paper of Matzkin (2007), when the unobservable random terms in an economic model have important interpretations such as tastes of consumers or productivity shocks in production functions, it is common that these unobservable random terms influence the dependent variables in an non-additive way. Consider the outcome variable $Y = \phi(D, X, U)$, where the structural function is strictly increasing in the third argument. Assume there is some external variable Z such that *conditional quantile independence* $Q_\tau(U|D, X, Z) = Q_\tau(U|X, Z)$ holds for the endogenous variable of interest D. Then the CQF of Y given D, X, Z can identify the structural function ϕ up to a normalization on $Q_\tau(U|X, Z)$,

$$Q_\tau(Y|D, X, Z) = \phi(D, X, Q_\tau(U|D, X, Z)) = \phi(D, X, Q_\tau(U|X, Z)). \tag{2.4}$$

Therefore the partial derivative of the CQF with respect to D, $\partial_D Q_\tau(Y|D, X, Z)$, identifies the structural derivative $\partial_D \phi(D, X, Q_\tau(U|X, Z))$, which is the causal effect of D while leaving the value of the unobserved variable U unchanged at $Q_\tau(U|X, Z)$. Further assume quantile independence and normalize (2.4) $\phi(D, X, Q_\tau(U)) = \phi(D, X, \tau)$, which is the τth quantile treatment response defined in Chernozhukov and Hansen (2005).

The nonparametric estimation of the partial derivative of the CQF can be imprecise due to the slow rate of convergence. The AQD summarizes the quantile treatment effects averaging over (Z, D, X), weighted by their joint density, and its estimator is $\sqrt{n}$-consistent. A similar argument for increasing the precision is made by Altonji and Matzkin (2005) who nonparametrically estimate weighted averages of the local average response, instead of imposing parametric assumption. Ma and Koenker (2006) also use this "weighted average" idea to estimate the weighted average quantile treatment effect for Chesher (2003)'s triangular simultaneous equation model under parametric specification. So my AQD can be a complement and different object of interest to the existing literature.

Example 4 (Nonseparable structural model without monotonicity)

For a more general nonseparable structural model, $Y = \phi(X, \epsilon)$, Hoderlein and Mammen (2007) identify the local average structural derivative:

$$E\big[\partial_{X_1}\phi(X,\epsilon)\big|X = x, Y = Q_\tau(Y|X = x)\big] = \partial_{X_1}Q_\tau(Y|X = x)$$

by conditional independence assumption that ϵ and X_1 conditionally independent given $(X_2, ..., X_q)$ which can be correlation with ϵ. In words, the derivative of the CQF identifies the average marginal effects over all individual with the same observable covariates x and responses $Y = Q_\tau(Y|X = x)$. No major assumption on the dimensionality of ϵ and the structure of ϕ is imposed. Hoderlein and Mammen (2009) show that the average derivative in mean regression

$$E\Big[\nabla E[Y|X] \cdot f(X)\Big] = \int_0^1 \beta(\tau)d\tau \tag{2.5}$$

[3] which is the average of the AQD over all the quantiles $\tau \in (0, 1)$. Therefore, the AQD reveals more local and richer information at each quantile τ than average mean derivative does.

2.3 Estimator

The data consists of n observations $Z_i = (y_i, X_i')', i = 1, ..., n$, which is an i.i.d. random sample from a distribution that is absolutely continuous with respect to a σ-finite measure ν, with Random-Nikodym density $F(y, X)$. The average derivative estimator is the sample analog of $\beta(\tau)$ in (2.1) where the unknown functions are replaced by nonparametric estimation:

$$\hat{\beta}(\tau) = -\frac{2}{n}\sum_{i=1}^{n} \hat{Q}_\tau(Y|X_i)\nabla\hat{f}(X_i)\mathbf{1}_{\{X_i \in \hat{S}\}} \tag{2.6}$$

[3]

$$\int_{\mathcal{X}} E\Big[\nabla\phi(X,\epsilon)\Big|X = x\Big]f^2(x)dx = \int_{\mathcal{X}}\int_0^1 E\Big[\nabla\phi(X,\epsilon)\Big|X = x, Y = Q_\tau(Y|X = x)\Big]d\tau f^2(x)dx$$
$$= \int_{\mathcal{X}}\int_0^1 \nabla Q_\tau(Y|X = x)d\tau f^2(x)dx$$

for $\tau \in \mathcal{T} \equiv [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$. Although any nonparametric estimation for the unknown functions might work, I use Nadaraya-Watson kernel estimator for its mathematical and practical tractability. The CQF is estimated by inverting the estimated conditional distribution function (cdf) by a smooth Nadaraya-Watson estimator:

$$\hat{F}_Y(y|X_i) = \frac{\frac{1}{|H|(n-1)} \sum_{j \neq i} K(H^{-1}(X_j - X_i)) G(\frac{y - Y_j}{h_0})}{\frac{1}{|H|(n-1)} \sum_{j \neq i} K(H^{-1}(X_j - X_i))} \tag{2.7}$$

where the νth-order kernel K has the bandwidth matrix H, the $q \times q$ identity matrix multiplied by $h = h_n$, a positive sequence of n. The indicator function $\mathbf{1}_{\{Y_j \leq y\}}$ for the dependent variable is smoothed by a cumulative kernel $G(z) = \int^u g(t)dt$ with a second-order kernel g and a bandwidth sequence h_0. Solving the inverse function for the CQF by the smooth kernel G could improve the computation time, comparing with solving optimization algorithms for the nonsmooth indicator function or the check function by the local polynomial estimator in Chaudhuri et al. (1997) and Lee (2003). [4] The leave-one-out method is standard for the preliminary plug-in nonparametric estimator and is convenient for the U-statistics theory.

Since the higher-order or bias-reducing kernel is used, the cdf estimator is not strictly increasing in y, Chernozhukov et al. (2010) propose the rearrangement method to get a monotonized version of the estimate $\tilde{F}_Y(y|X_i)$ which preserves the same asymptotics as $\hat{F}_Y(y|X_i)$. Then the CQF can be estimated by $\hat{Q}_\tau(Y|X_i) := \inf_y\{\tilde{F}_Y(y|X_i) \geq \tau\}$.

To avoid the denominator problem of $\hat{f}(X)$ in estimating $F_Y(y|X)$ for $Q_\tau(Y|X)$, I follow Hardle and Stoker (1989) and Lavergne and Vuong (1996) to estimate $\beta(\tau)$ in (2.1) by a trimmed estimated density weight, $\hat{f}(X)\mathbf{1}_{\{\hat{f}(X) \geq \delta\}}$. Define the compact set $S = S_n \equiv \{X : f(X) \geq \delta\}$ and $\hat{S} = \hat{S}_n \equiv \{X : \hat{f}(X) \geq \delta\}$, where $\delta = \delta_n$ is a trimming bound such that $\delta \to 0$ as $n \to \infty$. [5] I use the uniform convergence results of kernel estimation in Hansen (2008) where the uniformity is over values of x in expanding sets of the form

[4] Lee (2003) estimates the index coefficient or the derivative of the CQF at each point by a local polynomial estimation in the first step. In the second step, the estimated coefficient is averaged out by a weighting function and a fixed trimming function.

[5] The sample analogue in (2.6) should be divided by $n_s = \sum_{i=1}^n \mathbf{1}_{\{\mathbf{X_i} \in \hat{\mathbf{S}}\}}$ instead of n in practice. Since they are equivalent asymptotically, I use n for notational ease.

$\{x : \|x\| \leq c_n\} = S$ with $\delta_n = \inf_{\|x\| \leq c_n} f(X)$ and sequences c_n either bounded or diverging slowly to infinity.

The leave-one-out kernel estimator for the density function of X at X_i is

$$\hat{f}(X_i) = \frac{1}{|H_1|(n-1)} \sum_{j \neq i} K(H_1^{-1}(X_j - X_i)) \tag{2.8}$$

where the kernel is of order ν_1. Here, H_1 and ν_1 can be generally different from H and ν used in the CQF estimation (2.7). Therefore, the sth component of $\nabla \hat{f}(X_i)$ is

$$\partial_{X_s} \hat{f}(X_i) \equiv \frac{\partial \hat{f}(X_i)}{\partial X_{si}} = \frac{1}{|H_1|(n-1)} \sum_{j \neq i} \frac{1}{h_1} k'\Big(\frac{X_{si} - X_{sj}}{h_1}\Big) \Pi_{t \neq s} k\Big(\frac{X_{ti} - X_{tj}}{h_1}\Big)$$

where X_{si} is the sth component of X_i.

The AQD can only be calculated for the continuous covariates. When the covariates contain discrete components $X = (X^{(c)}, X^{(d)})$, the same estimation works for each point in a finite set of the realized values of $X^{(d)}$. That is, if $X_i^{(d)} = X_s^{(d)}$, $\hat{f}(X_i) = \frac{1}{|H_1|(n_s - 1)} \sum_{j \neq i} K(H_1^{-1}(X_j^{(c)} - X_i^{(c)})) \mathbf{1}_{\{X_j^{(d)} = X_s^{(d)}\}}$ where $n_s = \sum_{j=1}^n \mathbf{1}_{\{X_j^{(d)} = X_s^{(d)}\}}$. The CQF is then estimated by $\hat{F}_Y(y|X_i) = \sum_{j \neq i} K(H^{-1}(X_j - X_i)) G(\frac{y - y_j}{h_0}) \mathbf{1}_{\{X_j^{(d)} = X_s^{(d)}\}} \Big/ \sum_{j \neq i} K(H^{-1}(X_j - X_i)) \mathbf{1}_{\{X_j^{(d)} = X_s^{(d)}\}}$. For example, I can calculate the AQD for women and men separately.

2.3.1 Scaled AQD

Following Powell et al. (1989), a more interpretable rescaled coefficient might be defined as $\beta^* \equiv \beta / E[f(X)]$ so that the density weight is normalized $W^*(X) = f(X)/E[f(X)]$ so that $EW^*(X) = 1$. The scaling parameter $\alpha \equiv E[f(X)]$ can be similarly estimated by $\hat{\alpha} \equiv n^{-1} \sum_{i=1}^n \hat{f}(X_i)$, where $\hat{f}(X_i)$ is estimated by (2.8). Then the scaled AQD estimator is

$$\hat{\beta}^*(\tau) = \frac{\hat{\beta}(\tau)}{\hat{\alpha}} = \Big[-\frac{2}{n} \sum_{i=1}^n \hat{Q}_\tau(Y|X_i) \nabla \hat{f}(X_i) \mathbf{1}_{\{X_i \in \hat{S}\}} \Big] \Big/ \Big[\frac{1}{n} \sum_{i=1}^n \hat{f}(X_i) \Big]. \tag{2.9}$$

2.4 Asymptotic Properties

I first state my main results. The compact set S defined by the true density function $f(X)$ is $S \equiv \{X : f(X) \geq \delta\}$. Limits are taken as $n \to \infty$ unless otherwise noted. I

consider any $\tau \in \mathcal{T}$, so I drop τ for notational ease, i.e., the scaled AQD is $\beta^* = \beta/\alpha$. I first derive a Bahadur-type uniform linear representation for the nonparametrically estimated CQF, $\hat{Q}_\tau(Y|X)$. I make use the uniform convergence results of kernel estimation in Hansen (2008). Chaudhuri et al. (1997), Bhattacharya and Gangopadhyay (1990) and Dabrowska (1992) derive the Bahadur representation for different nonparametrically estimation of the CQF. The following Bahadur representation of the kernel estimated CQF could be of separate interest.

Proposition 2 (Bahadur representation)**.** *Suppose Assumptions A.FX, A.FY, and A.K hold. Let the smoothness parameters $p_X, p_Y \geq \nu$. Choose the bandwidths h, h_0, the trimming parameter δ, and the order of the kernel ν to satisfy $\delta^{-2}(nh_0h^q)^{-1/2} \to 0$, $(nh^q)^{1/2}(h_0^2+h^\nu) \to 0$, and the positive sequences $h, h_0, \delta \to 0$. For any $\tau \in \mathcal{T}$ and $X \in \mathcal{S}$,*

$$\hat{Q}_\tau(Y|X) - Q_\tau(Y|X) = \frac{1}{(n-1)|H|} \frac{\sum_{j=1}^n K\big(H^{-1}(X_j - X)\big) \cdot \big(\tau - G\big(\frac{Q_\tau(Y|X)-Y_j}{h_0}\big)\big)}{f(X) f_Y(Q_\tau(Y|X)|X)} + R_n(X) \tag{2.10}$$

$$= O_p\Big(\frac{1}{\delta}\Big(\frac{\log n}{nh^q}\Big)^{1/2}\Big).$$

The remaining term $R_n(X)$ satisfies $\sup_{X\in\mathcal{S}} |R_n(X_i)| = O_p\Big(\frac{\log n}{\delta^2 nh^q\sqrt{h_0}}\Big)$.

Following Hardle and Stoker (1989), the asymptotic theorem will be first derived for $\tilde{\beta} = -\frac{2}{n}\sum_{i=1}^n \hat{Q}_i \nabla \hat{f}_i \mathbf{1}_{\{f(X)\geq\delta\}}$, trimmed based on the true density. Then I will show $\sqrt{n}(\tilde{\beta} - \hat{\beta}) = o_p(1)$. The trimming method needs the assumption on the tail behavior, $E[\|Q_\tau(Y|X) \cdot \nabla f(X)\| \mathbf{1}_{\{X:f(X)<\delta\}}] = o(n^{-1/2})$. The similar tail assumption has been made in Lavergne and Vuong (1996) and Khan and Tamer (2010) for the denominator problem.

Following the idea of the proof in Powell et al. (1989) and Chaudhuri et al. (1997), $\tilde{\beta}$ can be decomposed as

$$-\frac{2}{n}\sum_{i=1}^{n}\hat{Q}_\tau(X_i)\nabla\hat{f}_i\mathbf{1}_{X_i} = \underbrace{-\frac{2}{n}\sum_{i=1}^{n}(\hat{Q}_i - Q_i)\nabla f_i\mathbf{1}_{X_i}}_{(I)}\underbrace{-\frac{2}{n}\sum_{i=1}^{n}Q_i\nabla\hat{f}_i\mathbf{1}_{X_i}}_{(II)}$$
$$\underbrace{-\frac{2}{n}\sum_{i=1}^{n}\big(\hat{Q}_i - Q_i\big)\big(\nabla\hat{f}_i - \nabla f_i\big)\mathbf{1}_{X_i}}_{(III)},$$

where $\mathbf{1}_{X_i} \equiv \mathbf{1}_{\{X_i\in S\}} = \mathbf{1}_{\{f(X_i)\geq\delta\}}$. The asymptotic properties for (I) and (II) can be derived by the U-statistics theory. The third term (III) will be made smaller order term by appropriately choosing h, h_0, h_1, ν, and ν_1.

The following is the main theorem for $\hat{\beta}$ and the scaled $\hat{\beta}^*$.

Theorem 2.1. *Suppose all Assumptions in Appendix hold and the smoothness parameters $p_Q \geq \nu$, $p_X \geq \max\{\nu, \nu_1\}$, and $p_Y \geq \nu$. Then*

1. *$\hat{\beta}$ is asymptotically linear*

$$\sqrt{n}(\hat{\beta}-\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg[\frac{2\nabla f(X_i)}{f_Y(Q_\tau(Y|X_i)|X_i)}\Big(\mathbf{1}_{\{Y_i\leq Q_\tau(Y|X_i)\}}-\tau\Big) + 2f(X_i)\nabla Q_\tau(Y|X_i) - 2E[f(X)\nabla Q_\tau(Y|X)]\Bigg] + o_p(1)$$
$$\equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_\beta(Z_i) + o_p(1) \xrightarrow{d} \mathcal{N}(0,\Sigma),$$

where $\Sigma = var\big(r_\beta(Z_i)\big) = 4\tau(1-\tau)E\Big[\frac{\nabla f(X)\nabla f(X)'}{f_Y^2(Q_\tau(Y|X)|X)}\Big] + 4var\Big(f(X)\nabla Q_\tau(Y|X)\Big)$. Denote $Q_i \equiv Q_\tau(Y|X_i)$. The bias $E[\hat{\beta}-\beta] =$

$$-2E\Big[Q_i\frac{h^\nu}{\nu!}\kappa_\nu\sum_{k=1}^{q}\partial_k^\nu\nabla f(X_i)\Big] + 2E\Big[\frac{\nabla f(X_i)}{f(X_i)f_Y(Q_i|X_i)}\Big\{\frac{h_0^2}{2}\kappa_{G2}f_Y'(Q_i|X_i)f(X_i) + h^\nu\kappa_\nu\sum_{l=1}^{\nu}\frac{1}{l!(\nu-l)!}\sum_{k=1}^{q}\partial_k^l F_Y(Q_i|X_i)\cdot\partial_k^{\nu-l}f(X_i)\Big\}\Big] + o(h^\nu+h_0^2) = O(h^\nu+h_0^2).$$

2. *The scaling parameter, $\alpha = E[f(X)]$, is estimated by $\hat{\alpha}-\alpha = \frac{1}{n}\sum_{i=1}^{n} r_\alpha(Z_i) + o_p(n^{-1/2})$, where the influence function $r_\alpha(Z_i) = 2\big(f(X_i) - E[f(X)]\big)$.*

3. For the scaled AQD, $\beta^* = \beta/\alpha$,

$$\sqrt{n}\Big(\hat{\beta}^* - \beta^*\Big) \xrightarrow{d} \mathcal{N}(0, V^*),$$

where $V^* = V/\alpha^2$ *and* $V = var\Big(r_\beta(Z_i) - r_\alpha(Z_i)\beta^*\Big)$.

Consequently, an asymptotically pivotal test statistic, a confidence interval or the corresponding hypothesis test can be constructed by a studentized version of $\hat{\beta}$ using Slusky's theorem: $\sqrt{n}\hat{V}^{*-1/2}(\hat{\beta}^* - \beta^*) \xrightarrow{d} \mathcal{N}(0, I_q)$ with a consistent covariance matrix estimator $\hat{V}^* \xrightarrow{p} V^*$ in the next section. A hypothesis test of the quantile treatment effect can be carried out by testing the null hypothesis the coefficient for the treatment variable $\beta_1^* = 0$.

The influence function $r_\beta(Z_i) = 2(r_{Ii} + r_{\mathit{II}i}) - \beta$. The first part of the influence function $r_{Ii} = r_I(Z_i) = \frac{\nabla f(X_i)}{f_Y(Q_i|X_i)}\Big(\mathbf{1}_{\{y_i \le Q_i\}} - \tau\Big)$ comes from the estimation error of the unknown CQF with the known $f(X)$. The second part $r_{\mathit{II}i} = r_{\mathit{II}}(Z_i) = f(X_i)\nabla Q_i - E\big[\nabla(Q_i f(X_i))\big]$ comes from estimating the density weight. For the case when $W(X) = f(X)$ is a known function, Chaudhuri et al. (1997) show that $\sqrt{n}(\hat{\beta}_W - \beta) \to \mathcal{N}(0, \Sigma_1)$, where $\Sigma_1 = var(2r_{Ii} + r_{\mathit{II}i} - \beta)$. Comparing with my result when the density function is known $W(X_i) = f(X_i)$, the estimation to the unknown density $f(X_i)$ contributes an extra $\frac{1}{n}\sum_{i=1}^n f(X_i)\nabla Q_i - E[f(X_i)\nabla Q_i]$ in the influence function. Therefore, when applying the AQD in practice by choosing $W(X) = \hat{f}(X)$, the estimation error is not first-order ignorable.

For the choice of the bandwidths, trimming parameter, and order of kernels, I illustrate by the following sufficient condition. Let $h \propto n^{-a}$, $h_1 \propto n^{-c}$, $h_0 \propto n^{-d}$, and $\delta \propto n^{-b}$, for some constants, $a, b, c, d > 0$. Choose $\nu > \frac{4q}{3}$, $a \in (\frac{1}{2\nu}, \frac{3}{8q})$, $\nu_1 > \frac{q+2}{2-2aq}$, $c \in (\frac{1}{2\nu_1}, \frac{1-aq}{q+2})$, $d \in (\frac{1}{4}, 1-2aq)$, and $b < \min\{\frac{1}{4}(1-2aq-d), \frac{1}{2}(1-aq-c(q+2))\}$. Choosing $\nu_1 > \frac{4}{5}(q+2)$ is sufficient. The conditions on the smoothness parameters is weaker than the estimator for the known weighted AQD in (2.3) in Chaudhuri et al. (1997), by $3 + \frac{3}{2}q > \max\{\frac{4q}{3}, \frac{4}{5}(q+2)\}$. The curse of dimensionality from nonparametric estimation goes to the order of bias-reducing kernel ν and ν_1. Hence, the distributions of Y and X (e.g., $Q_\tau(Y|X)$ and $f(X)$) need to be increasingly smooth as q increases. To achieve asymptotic linearity, a large bandwidth is needed. To make the bias vanish at rate $\sqrt{N}$, I need small bandwidths and higher-order

kernels. It is undersmooth comparing with nonparametric estimations of the curves (e.g., nonparametric density and cdf estimations). That is, the bandwidth needs to be shrunk more rapidly to zero than the typical bandwidth for curve estimation, which is common in the literature. Take an example for the dimension of X, $q = 4$. I can choose a sixth-order biweight kernel, the positive sequences of the bandwidths $h \propto n^{-0.09}$, $h_1 \propto n^{-0.1}$, $h_0 \propto n^{-0.26}$, and the trimming parameter $\delta \propto n^{-0.004}$.

2.4.1 Asymptotic Covariance Matrix

The covariance matrix Σ could be consistently estimated as the sample variance of uniformly consistent estimators of the influence function $r_\beta(Z_i) \equiv r_{\beta i}$. And the influence function can be estimated by any uniformly consistent estimators of $f(X_i), \nabla f(X_i), f_Y(Q_i|X_i), Q_i$, and ∇Q_i, under some regularity conditions. So I estimate the influence function of (I), r_{I_i}, by replacing the unknown functions with the uniformly consistent estimators. For $r_{I\!I i}$, I follow Hardle and Stoker (1989) using the projection structure in its U-statistic. Therefore, the estimator for Σ can be constructed as

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} \hat{r}_{\beta i} \hat{r}'_{\beta i} \mathbf{1}_{\{\hat{f}(X_i) \geq \delta\}} - \bar{r}\bar{r}'$$

where

$$\begin{aligned}
\bar{r} &= n^{-1} \sum_{i=1}^{n} \hat{r}_{\beta i} \mathbf{1}_{\{\hat{f}(X_i) \geq \delta\}} \\
\hat{r}_{\beta i} &= 2(\hat{r}_{Ii} + \hat{r}_{I\!I i}) - \hat{\beta} \\
\hat{r}_{Ii} &= \frac{\nabla \hat{f}(X_i)}{\hat{f}_Y(\hat{Q}_i|X_i)} \big(\mathbf{1}_{\{y_i \leq \hat{Q}_i\}} - \tau \big) \\
\hat{r}_{I\!I i} &= \frac{-1}{n-1} \sum_{j \neq i} \frac{1}{h^{q+1}} \nabla K\big(H^{-1}(X_i - X_j)\big) \big(\hat{Q}_i \hat{\mathbf{1}}_{X_i} - \hat{Q}_j \hat{\mathbf{1}}_{X_j} \big).
\end{aligned}$$

Theorem 2.2. *Under the Assumptions, $\hat{\Sigma}$ is a consistent estimator of Σ.*

The influence function of $\hat{\beta}^* \equiv \hat{\beta}/\hat{\alpha}$ can be estimated by

$$\hat{r}_i^* \equiv \Big(2\hat{r}_{Ii} + 2\hat{r}_{I\!I i} - 2\hat{\beta} - 2\Big(\hat{f}(X_i) - \hat{\alpha} \Big) \hat{\beta}^* \Big) \frac{1}{\hat{\alpha}}.$$

Then the asymptotic covariance matrix of the scaled AQD estimator, V/α^2, can be estimated by $\hat{V}^* \equiv n^{-1}\sum_{i=1}^n \hat{r}_i^*\hat{r}_i^{*\prime}\mathbf{1}_{\{\hat{f}(X_i)\geq\delta\}} - \bar{r}^*\bar{r}^{*\prime}$, where $\bar{r}^* = n^{-1}\sum_{i=1}^n \hat{r}_i^*\mathbf{1}_{\{\hat{f}(X_i)\geq\delta\}}$.

2.5 Monte Carlo Simulations

I compare the finite-sample performance of the AQD estimator with the average mean derivative (AMD) in Powell et al. (1989), the conventional Koenker and Bassett (1978) linear quantile regression (denoted by KB), and the OLS. I consider two semiparametric model: partial linear and single index models. The data generating processes are modified from the experiments in Lee (2003).

1. Partial linear model with homoscedasticity (PL-homo):

   $Y = X_1 + X_2 + 30\exp(-X_1^2)/\sqrt{2\pi} + \epsilon$

2. Partial linear model with heteroscedasticity (PL-hetero):

   $Y = X_1 + X_2 + 30\exp(-X_1^2)/\sqrt{2\pi} + 2\exp((X_1+X_2)/4)\epsilon$

3. Single index model with homoscedasticity (SI-homo):

   $Y = 20 + 10\sin(X'\beta/2) + X'\beta/2 + \epsilon$, where $X'\beta = 5X_1 + X_2$.

4. Single index model with heteroscedasticity (SI-hetero):

   $Y = 20 + 10\sin(X'\beta/2) + X'\beta/2 + 2\exp(X'\beta/12)\epsilon$, where $X'\beta = 5X_1 + X_2$.

I consider two error distributions: $\epsilon \sim \mathcal{N}(0,1)$ and $\epsilon \sim t(2)$ for the fat-tailed distribution.

I use fourth-order Epanechnikov kernel. The bandwidths are $h_x = C\sigma_x n^{-0.15}$ and $h_y = C\sigma_y n^{-0.3}$, where the powers satisfy Assumption A.B, the constant $C = 3.12$ from Silverman rule-of-thumb, and σ is the inter-quantile range robust to fat-tailed design Silverman (1986).

There are 1,000 replications in each experiment. For the normal error in partial linear model in Figure 2.1, the nonparametric estimators, AMD and AQD, outperform the linear estimators, OLS and KB. Both AMD and AQD identify the coefficient for X_2 for this partial linear model. AMD is more efficient than AQD, which could be explained by (2.5) that AMD

integrates more information. For the fat-tailed error in Figure 2.2, the quantile regressions (AQD and KB) outperform the mean regressions (AMD and OLS).

The rule-of-thumb constant 3.12 is close to minimize the MSE. The optimal bandwidth for the AQD is smaller than that of the AMD. This is because AQD involves additional nonparametric estimation of the CQF, the nonparametric estimator is more undersmooth. When the optimal bandwidth is chosen, the nonparametric estimators performs well in finite sample. The similar results can be observed for the single-index models in Figures 2.3 and 2.4. AQD outperforms the linear KB.

2.6 Discussion and Outlook

Efficiency In Newey (1990), "In models where the parameter is an explicit function of the distribution and the distribution is unrestricted, there is only one influence function for a regular asymptotically linear estimator." That is, the influence function of any asymptotically linear and regular estimator for the density-weighted AQD is unique and hence efficient. It follows that my estimator for the density-weighted AQD reaches the efficiency bound. Other nonparametric estimations for the first-step unknown functions, such as series or local polynomial, will give the same asymptotic distribution. More explicitly, Newey and Stoker (1993) calculate the efficiency bounds for the weighted average derivative for general loss functions, including conditional mean and quantiles, where the weighting function is a known function. By proceeding as in the proof for Theorem 3.1 in Newey and Stoker (1993), I can calculate the efficiency bounds for the density-weighted average quantile/mean derivatives where the density weight is to be estimated. It confirms that the estimators proposed in this paper and by Powell et al. (1989) are semiparametrically efficient, as implied by the result in Newey (1990).

The choice of the weighting functions may affect the efficiency of estimating the index parameter in semiparametric models, such as single index models in Newey and Stoker (1993) and partial linear models in Lee (2003). For the partial linear model $Y = X_1'\beta_0 + \phi(X_2) + \epsilon$, Lee (2003) derives the optimal weight so that his weighted average quantile regression

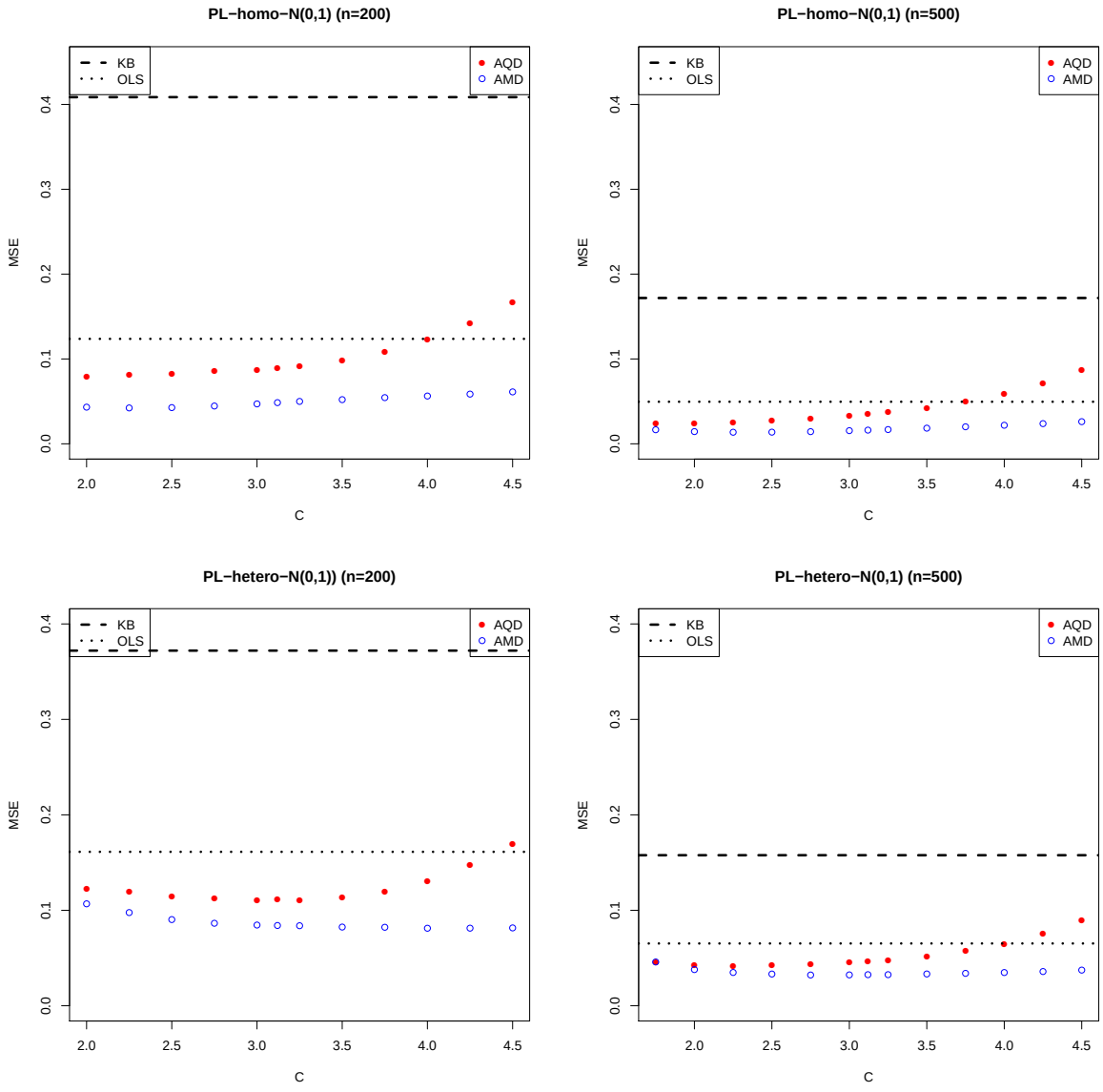

Figure 2.1 (Partial linear - $N(0,1)$) The true parameter is 1, the coefficient of X_2.

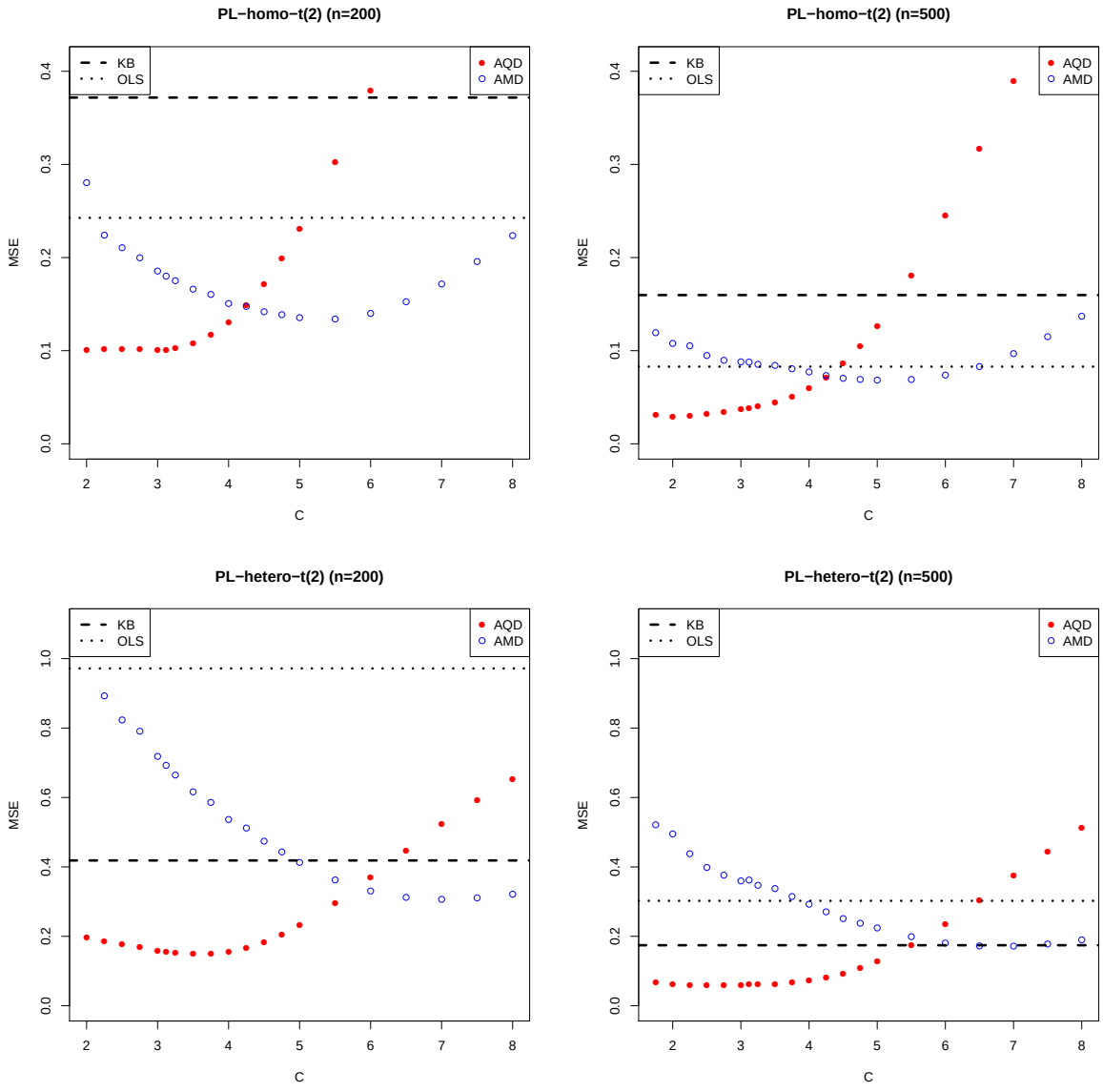

Figure 2.2 (Partial linear - $t(2)$) The true parameter is 1, the coefficient of X_2.

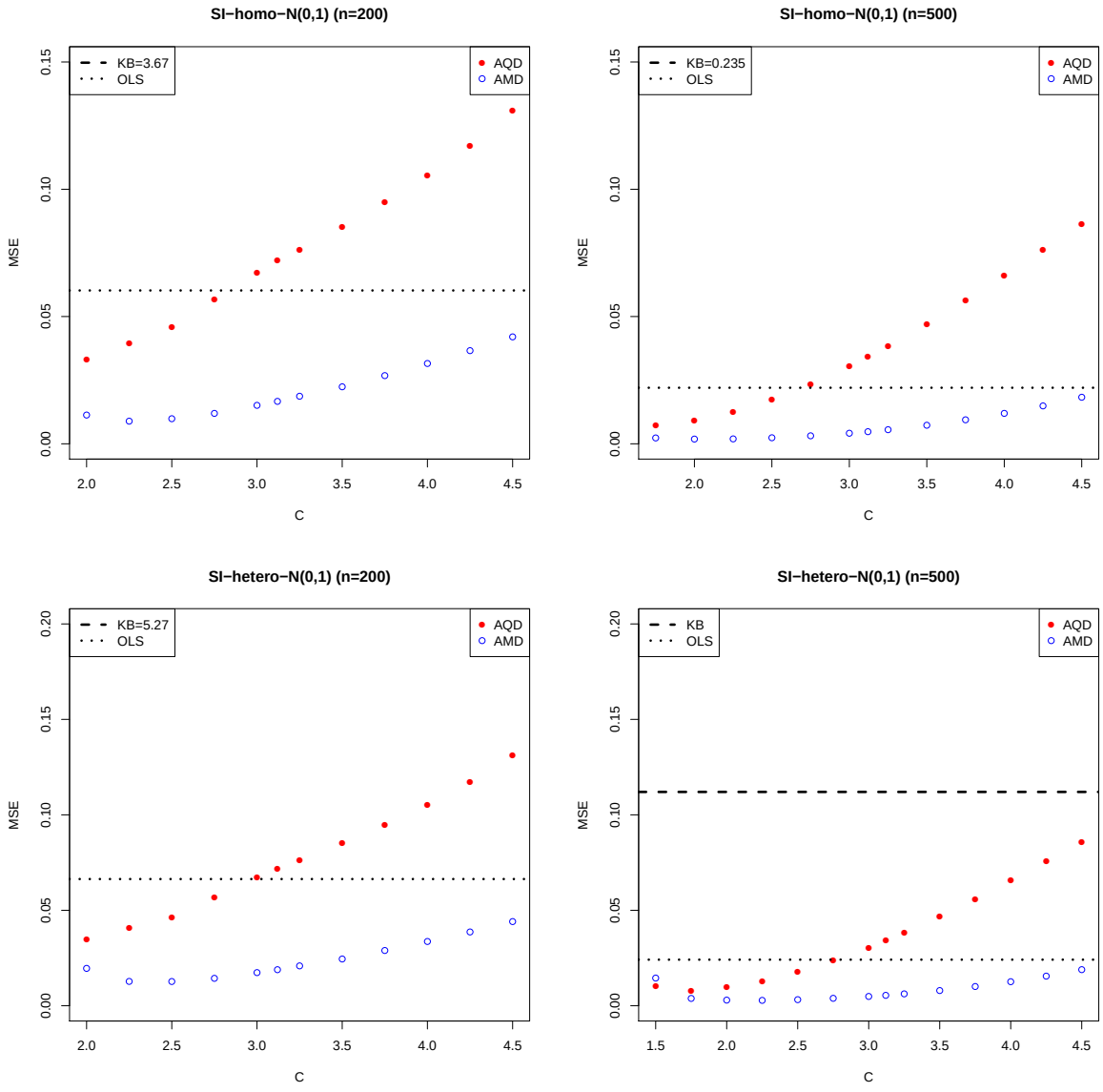

Figure 2.3 (Single Index - $N(0,1)$)

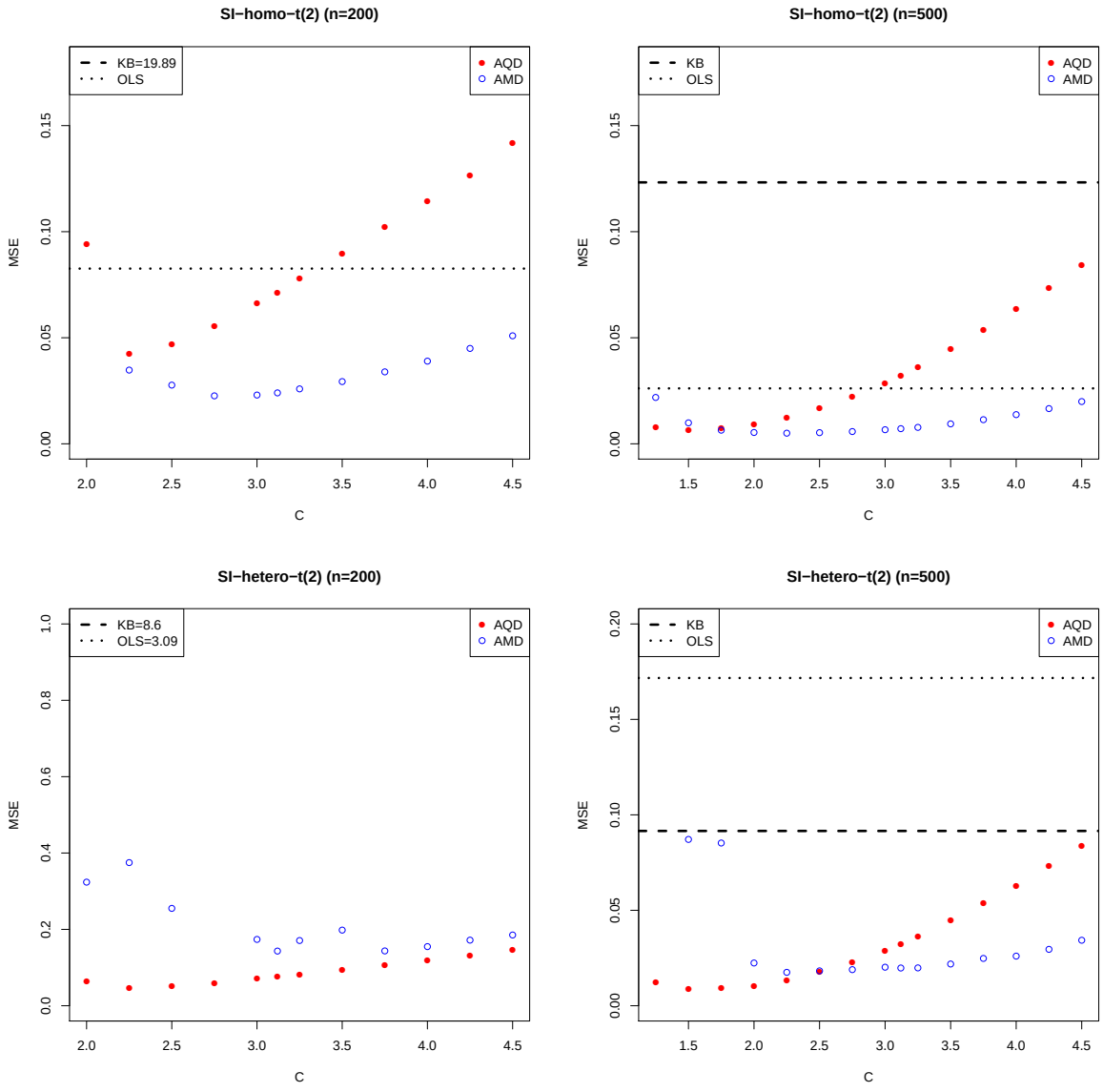

Figure 2.4 (Single Index - $t(2)$)

reaches the semiparametric efficiency bound for β_0. Newey and Stoker (1993) derive the optimal weight for the weighted average mean derivative in the single index model, when X has an ellipitically symmetric distribution. Then the declining weight implicit in the density weighted estimator would tend to have high efficiency when the X distribution has thinner tails than normal. This may motivate the density weight used for the mean case in Powell et al. (1989). Since AMD integrates more information than AQD as in (2.5), the AMD estimator might be more efficient than the AQD estimator in semiparametric models. However, Newey and Stoker (1993) note that the AQD estimator may be more efficient than the AMD estimator if the distribution of Y is fat-tailed. The Monte-Carlo experiment in Section 2.5 illustrates this point.

Bandwidth Choice The criterion of choosing bandwidths and trimming bound for small sample is to be investigated. As discussed in Lavergne and Vuong (1996), the optimal bandwidths need not have this form $C \cdot n^{-a}$ for some positive constant C, a. The optimal bandwidths can be chosen by minimizing the mean-square error of β as in Powell and Stoker (1996). However, for the AQD, the additional influence function from the estimation of the CQF in (I) complicates the problem.

Robust Inference As the dimension of the covariates increases, a higher-order kernel is needed. It is known that the finite-sample performance of the average mean derivative might deteriorate, e.g., Cattaneo et al. (2010). The classical first-order, asymptotically linear large sample theory, which ignores the remainder terms, may not capture the finite-sample behavior of $\hat{\beta}^*$. My variance estimation for the density-weighted AQD shares the same spirit of Powell et al. (1989), which is lack of robustness with respect to the bandwidth noted by CCJ (2010). Alternative inference method is to be investigated. A higher-order asympotic expansion for the AQD might be needed.

Studentized Estimator The density-weighted AQD is defined on all the support of X and I require $f(X)$ goes to zero on the boundary of the support of X. To make my

estimator converge at $\sqrt{n}$-rate by the U-statistics theory and stochastic trimming, I need the second moment (at least) $E[\|Q_\tau(Y|X) \cdot \nabla f(X)\|^2]$ to be finite. In Khan and Tamer (2010), they study the case when the second moment is not finite and the convergence rate of the estimator $\hat{\beta}$ is not regular ($\sqrt{n}$). They studentize the estimator as $\sqrt{n}\hat{\Sigma}^{-1/2}(\hat{\beta} - \beta)$, where $\hat{\Sigma}$ is an estimator for the asymptotic variance Σ if conditions were such that the asymptotic variance were finite. They show that this studentized estimator converges to a standard normal distribution, regardless of the rate of convergence of the un-studentized estimator. Their idea might be applicable to my density-weighted AQD.

Chapter 3

Interpretation and Semiparametric Efficiency in Quantile Regression under Misspecification

3.1 Introduction

This article revisits the approximation properties of the linear quantile regression under misspecification (Angrist et al. (2006); Kim and White (2003); Hahn (1997)). I study the quantile regression parameter, which is the best linear predictor of the outcome under the asymmetric check loss function without assuming that the true conditional quantile function is linear. I calculate the semiparametric efficiency bound of this parameter. The quantile regression estimator, introduced by Koenker and Bassett (1978), offers parsimonious summary statistics for the conditional quantile function and is computationally tractable. Since the development of the estimator, researchers have frequently used quantile regression, in conjunction with ordinary least squares regression, to analyze how the outcome variable responds to the explanatory variables. For example, to model wage structure in labor economics, Angrist et al. (2006) study returns to education at different points in the wage distribution and changes in inequality over time. A thorough review of recent development in quantile regression can be found in Koenker (2005).

The topic of interest is the conditional *cumulative distribution function (cdf)* of a continuous response variable Y given the regressor vector X, denoted as $F_Y(y|X)$. A convenient alternative for the conditional cdf is the τth *conditional quantile function (CQF)* of Y given

X, defined as $Q_\tau(Y|X) := \inf\{y : F_Y(y|X) \geq \tau\}$. Assuming integrability, the CQF minimizes the check loss function

$$Q_\tau(Y|X) \in \arg\min_{q(X)\in\mathcal{Q}} E\Big[\rho_\tau(Y - q(X))\Big]$$

for any quantile index $\tau \in (0,1)$, where $\mathcal{Q}$ is the set of measurable functions of X, $\rho_\tau(u) = u(\tau - \mathbf{1}_{\{u\leq 0\}})$ is known as the check function, and $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Therefore, a natural and simple summary statistic for the CQF is the *quantile regression (QR) parameter* $\beta(\tau)$, which solves the population minimization problem

$$\beta(\tau) := \arg\min_{\beta\in R^d} E\Big[\rho_\tau(Y - X'\beta)\Big] \tag{3.1}$$

assuming integrability and uniqueness of the solution and d is the dimension of X. The *QR estimator* introduced by Koenker and Bassett (1978) is the sample analogue

$$\hat{\beta}(\tau) \in \arg\min_{\beta\in R^d} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - X_i'\beta) \tag{3.2}$$

for the random sample $(Y_i, X_i', i \leq n)$ on the random variables $(Y, X')'$. By the equivalent first-order condition, this estimator $\hat{\beta}(\tau)$ is also the generalized method of moments (GMM) estimator based on the unconditional moment restriction (Powell, 1984, 1986):

$$E[(\tau - \mathbf{1}_{\{Y\leq X'\beta(\tau)\}})X] = 0. \tag{3.3}$$

In this article I study how the QR parameter defined in (3.1) or (3.3) approximates the CQF and calculates its semiparametric efficiency bound.

If the CQF is modeled to be linear in the covariates $Q_\tau(Y|X) = X'\beta(\tau)$ or $F_Y(X'\beta(\tau)|X) = \tau$, the coefficient $\beta(\tau)$ satisfies the conditional moment restriction

$$E[\tau - \mathbf{1}_{\{Y\leq X'\beta(\tau)\}}|X] = 0 \tag{3.4}$$

almost surely. In the theoretical and applied econometrics literature, this *linear QR model* is often assumed to be *correctly specified*. Nevertheless, a well-known crossing problem arises: the CQF for different quantiles may cross at some values of X, except when $\beta(\tau)$ is the same

for all τ. A logical monotone requirement is violated for $Q_\tau(Y|X)$ or its estimator to be weakly increasing in the probability index τ given X. The crossing problem for estimation could be treated by rearranging the estimator (for example, see Chernozhukov et al. (2010) and the references therein). [1] However, the crossing problem remains for the population CQF, suggesting that the linear QR model is inherently *misspecified*. That is, there is no $\beta(\tau) \in R^d$ satisfying the conditional moment (3.4) almost surely. Therefore, the parameter of interest in this article is the QR parameter $\beta(\tau)$ defined by (3.1) or (3.3) without the linear CQF assumption in (3.4). We can view $\beta(\tau)$ as the pseudo-true value of the linear QR model under misspecification.

While $\beta(.5)$ is the least absolute derivations estimation, the QR parameter $\beta(\tau)$ for other quantiles is the *best linear predictor* for a response variable under the asymmetric loss function $\rho_\tau(\cdot)$ in (3.1). Angrist et al. (2006) (henceforth ACF) note that the prediction under the asymmetric check loss function is often not the object of interest in empirical work, with the exception of the forecasting literature, for example, Komunjer (2005). For the mean regression counterpart, ordinary least squares (OLS) consistently estimates the linear conditional expectation and minimizes mean-squared error loss for fitting the conditional expectation under misspecification. The attractive features of OLS, robustness and interpretability, under misspecification, motivate the investigation of parallel properties in QR.

The equivalent first-order condition can be understood as the orthogonality condition of the covariates X and the *distribution error*, $F_Y(X'\beta(\tau)|X) - \tau$, in the projection model. I show that the QR parameter $\beta(\tau)$ minimizes the mean-squared distribution error inversely weighted by the conditional density function at the *best linear approximation* $X'\beta(\tau)$. ACF (2006) find that QR is the best linear approximation of the CQF, using a weighted mean-squared error loss function and a weight primarily determined by the conditional density $f_Y(Q_\tau(Y|X)|X)$. ACF's study as well as my own results suggest that QR approximates the CQF more accurately at points with more observations, but the corresponding conditional

[1] Chernozhukov et al. (2010) rearrange an estimator $\hat{Q}_\tau(Y|X)$ to be monotonic. The original estimator can be computationally tractable. The rearranged monotonic estimated conditional cdf is $\hat{F}_Y(y|X) = \int_0^1 \mathbf{1}_{\{\hat{Q}_\tau(Y|X) \leq y\}} d\tau$. The rearranged quantile estimation is $\hat{Q}^*_\tau(Y|X) = \inf\{y : \hat{F}_Y(y|X) \geq \tau\}$.

cdf evaluated at the approximated point $F_Y(X'\beta(\tau)|X)$ is more distant from the targeted probability τ. This trade-off is controlled by the conditional density, which is distinct from OLS approximating the conditional mean, because the distribution and quantile functions are generally nonlinear operators. This observation is novel and increases the understanding of how the QR summarizes the outcome distribution. A numerical example in Figure 3.1 in Section 4 illustrates this finding.

For the misspecified linear regression model, Chamberlain (1987) proves the semiparametric efficiency of the OLS estimator based on differentiable moment restrictions, which provides additional justification for the widespread use of OLS. However, Chamberlain's results cannot be applied to semiparametric efficiency for QR, due to the lack of moment function differentiability in (3.3) and (3.4). Although Ai and Chen (2012)'s general results for sequential moment restrictions containing unknown functions could cover the quantile regression setting, I calculate the efficiency bound accommodating regularity conditions specifically for the QR parameter $\beta(\tau)$ using the method outlined in Severini and Tripathi (2001). It follows that the misspecification-robust asymptotic variance of the QR estimator $\hat{\beta}(\tau)$ in (3.2) attains this bound, which means no regular [2] estimator for (3.3) has smaller asymptotic variance than $\hat{\beta}(\tau)$. This result might be expected for an M-estimator, but, to my knowledge, the QR application has not been demonstrated and discussed rigorously in any publication. Further, I calculate the efficiency bounds for jointly estimating QR parameters at finite number of quantiles for both misspecified (3.3) and correctly specified (3.4) models. Employing the widely used method outlined in Newey (1990), Newey and Powell (1990) find the semiparametric efficiency bound for $\beta(\tau)$ of the correctly specified linear CQF in (3.4). Note that the efficiency bounds for (3.3) do not imply the bounds for (3.4); nor does the converse hold.

In Section 2, I discuss the interpretation of misspecified QR model in terms of approximating the conditional cdf and the CQF. The theorems for the semiparametric efficiency bounds are in Section 3. In Section 4, I discuss the parallel properties of QR and OLS. The

[2] See Newey (1990) for the definition of regular estimators.

article is concluded by a review of some existing efficient estimators for correctly specified (3.4) and misspecified (3.3) linear QR models.

3.2 Interpreting QR Under Misspecification

Y is a continuous response variable and X is a $d \times 1$ regressor vector. The quantile-specific residual is defined as the distance between the response variable and the CQF, $\epsilon_\tau := Y - Q_\tau(Y|X)$ with the conditional density $f_{\epsilon_\tau}(e|X)$ at $\epsilon_\tau = e$ or $f_Y(y|X)$ at $Y = y = e + Q_\tau(Y|X)$ for any $\tau \in (0,1)$. This is a semiparametric problem in the sense that the distribution functions of ϵ_τ and X as well as the CQF are unspecified and unrestricted other than by the following assumptions which are standard in QR models. Throughout this article, I assume the regularity conditions borrowed from Theorem 3 in ACF (2006):

(R1) $(Y_i, X_i, i \leq n)$ are independent and identically distributed on the probability space $(\Omega, \mathcal{F}, P)$ for each n;

(R2) the conditional density $f_Y(y|X = x)$ exists, and is bounded and uniformly continuous in y, uniformly in x over the support of X;

(R3) $J(\tau) := E\Big[f_Y(X'\beta(\tau)|X)XX'\Big]$ is positive definite for all $\tau \in (0,1)$, where $\beta(\tau)$ is uniquely defined in (3.1);

(R4) $E\|X\|^{2+\epsilon} < \infty$ for some $\epsilon > 0$.

The identification of the pseudo-true parameter $\beta(\tau)$ is assumed in (R3). If X contains a constant component and the unique solution to (3.1) exists, then the intercept in $\beta(\tau)$ is identified. This is different from the case of the correctly specified model (3.4), where the intercept in $\beta(\tau)$ is not identified. The bounded conditional density function of the continuous response variable Y given X in (R2) is needed for the existence of the CQF for any $\tau \in (0,1)$. The uniform continuity guarantees the existence and differentiability of the distribution function, i.e., $dF_Y(y|X)/dy = f_Y(y|X)$ and $F_Y(y|X) = \int_{-\infty}^{y} f_Y(u|X)du$

with probability one. (R4) is used for the asymptotic normality of $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$. The covariates X are allowed to contain discrete components.

The parameter of interest $\beta(\tau)$ is equivalent to solving

$$E\Big[X\Big(F_Y\big(X'\beta(\tau)\big|X\big) - \tau\Big)\Big] = 0 \tag{3.5}$$

by applying the law of iterated expectations on equation (3.3). Equation (3.5) states that X is orthogonal to the *distribution error* $F_Y\big(X'\beta(\tau)\big|X\big) - \tau$. The following theorem interprets QR through a weighed mean-squared error loss function on the distribution error.

Theorem 3.1. *Assume the regularity conditions (R1)-(R4). Further assume $f_Y(X'\beta(\tau)|X)$ to be bounded away from zero such that the objective function in (3.6) is finite $\forall \beta \in R^d$, where $\beta(\tau)$ is the parameter of interest uniquely defined by (3.1). Then $\bar{\beta}(\tau) = \beta(\tau)$ solves the equation*

$$\bar{\beta}(\tau) = \arg\min_{\beta \in R^d} E\big[\big(f_Y(X'\bar{\beta}(\tau)|X)\big)^{-1}\big(F_Y(X'\beta|X) - \tau\big)^2\big]. \tag{3.6}$$

Furthermore, if $E\big[\big(f_Y(X'\beta|X) + (F_Y(X'\beta|X) - \tau)f_Y'(X'\beta|X)/f_Y(X'\beta|X)\big)XX'\big]$ is positive definite at $\beta = \beta(\tau)$, then $\bar{\beta}(\tau) = \beta(\tau)$ is unique to this problem (3.6).

Proof of Theorem 3.1 *The objective function in (3.6) is finite by the assumptions. Any fixed point $\beta = \bar{\beta}(\tau)$ would solve the first-order condition, $E\big[X\big(F_Y\big(X'\beta\big|X\big) - \tau\big)\big] = 0$. By (3.3), $\beta(\tau)$ solves (3.6). When the second-order condition holds, i.e., $E\big[\big(f_Y(X'\beta|X) + (F_Y(X'\beta|X) - \tau)f_Y'(X'\beta|X)/f_Y(X'\beta|X)\big)XX'\big]$ is positive definite at $\beta = \beta(\tau)$, $\beta(\tau)$ solves (3.6) uniquely.* □

Theorem 3.1 states that the parameter $\beta(\tau)$ is the unique fixed point to an iterated minimum distance approximation, with a weight of a function of X only. The mean-squared loss makes it clear how the linear function matches the conditional cdf to the targeted probability of interest. The loss function puts more weight on points where the conditional density $f_Y(X'\beta(\tau)|X)$ is small. As a result, the distribution error is smaller at points with smaller conditional density.

ACF (2006) interpret QR as the minimizer of the weighted mean-squared error loss function for *specification error*, defined as the deviation between the approximation point $X'\beta(\tau)$ and the true CQF $Q_\tau(Y|X)$:

$$\beta(\tau) = \arg\min_{\beta \in R^d} E\big[\bar{w}_\tau(X, \beta(\tau)) \cdot \big(X'\beta - Q_\tau(Y|X)\big)^2\big] \tag{3.7}$$

where the importance weight

$$\bar{w}_\tau(X, \beta(\tau)) = \frac{1}{2}\int_0^1 f_{\epsilon_\tau}\Big(u\big(X'\beta(\tau) - Q_\tau(Y|X)\big)\Big|X\Big)du.$$

The importance weights $\bar{w}_\tau(X, \beta(\tau))$ are the averages of the response variable over a line connecting the approximation point $X'\beta(\tau)$ and the true CQF $Q_\tau(Y|X)$. ACF (2006) note that the regressors contribute disproportionately to the QR estimate and the primary determinant of the importance weight is the conditional density. Moreover, I observe that the first-order condition from ACF's result (3.7) is $E\Big[\bar{w}_\tau(X, \beta(\tau)) \cdot X\Big(X'\beta(\tau) - Q_\tau(Y|X)\Big)\Big] = 0$, which is a weighted orthogonal condition of the specification error. A Taylor expansion provides intuition to connect the distribution error and the specification error: $[f_Y(X'\beta|X)]^{-1}\big(F_Y(X'\beta|X) - \tau\big)^2 \approx f_Y(X'\beta|X)\big(Q_\tau(Y|X) - X'\beta\big)^2$. This observation implies the specification error is smaller at points where the conditional density $f_Y(X'\beta|X)$ is larger. On the other hand, the distribution error is larger at points with larger $f_Y(X'\beta|X)$. Comparing with the OLS where the mean operator is linear, the distribution function and its inverse operator, the quantile function, are generally nonlinear. The distribution error can be interpreted as the distance after a nonlinear transformation by the conditional cdf, $F_Y(X'\beta(\tau)|X) - F_Y(Q_\tau(Y|X)|X)$. A Taylor expansion linearizes the distribution function to the specification error multiplied by the conditional density function. The conditional density plays a crucial role on weighting the distribution error and the specification error. The above discussion provides additional insights to how the QR parameter approximates the CQF and fits the conditional cdf to the targeted probability.

3.3 The Semiparametric Efficiency Bounds

3.3.1 QR under Misspecification

I calculate the semiparametric efficiency bound for the unconditional moment restriction (3.3) by Severini and Tripathi (2001)'s approach.

Theorem 3.2. *Assume the regularity conditions (R1)-(R4). The semiparametric efficiency bound for estimating the population QR parameter $\beta(\tau)$, that minimizes the expected weighted mean-squared approximation error (3.7) or equivalently (3.1), is $J(\tau)^{-1}\Gamma(\tau,\tau)J(\tau)^{-1}$, where $J(\tau)$ is defined in (R3) and*

$$\Gamma(\tau_i,\tau_j) := E\Big[(\tau_i - \mathbf{1}_{\{Y<X'\beta(\tau_i)\}})(\tau_j - \mathbf{1}_{\{Y<X'\beta(\tau_j)\}})XX'\Big]$$

for any $\tau_i,\tau_j \in \mathcal{T}$:= a closed subset of $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$.

In general, the semiparametrically efficient joint asymptotic covariance of the estimators for $(\beta'(\tau_1),\beta'(\tau_2),...,\beta'(\tau_m)'$ is $J(\tau_i)^{-1}\Gamma(\tau_i,\tau_j)J(\tau_j)^{-1}$, for any $\tau_i,\tau_j \in \mathcal{T}$, $i,j = 1,2,...,m$, for a finite integer $m \geq 1$.

Proof of Theorem A.3 *See Appendix.* □

My proof accommodates the regularity assumptions for quantile regression and modifies Section 9 of Severini and Tripathi (2001). For example, the covariate X can contain discrete components, by constructing two tangent spaces for the conditional density of Y given X and the marginal density of X, respectively. In the efficiency bound, $J(\tau) := E\Big[f_Y(X'\beta(\tau)|X)XX'\Big]$ is obtained by assuming the exchangeability of integration and differentiation for the nonsmooth check function. [3]

The method in Severini and Tripathi (2001) has been used in the monotone binary model in Magnac and Maurin (2007), Lewbel (1998) latent variable models in Jacho-Chávez (2009), for example. I work in the Hilbert space of tangent vectors of the square-root density

[3]Severini and Tripathi (2001) construct the tangent space for the continuous and bounded joint density $f(X,Y)$ in their Secton 9. And they define J on the derivative of the moment restriction.

functions and using the Riesz-Fréchet representation theorem. Another equivalent approach by Newey (1990) works in a Hilbert space of random variables and uses the projection on the linear space spanned by the scores from the one-dimensional subproblems to find the efficient influence function. The efficiency bound is then the second moment of the efficient influence function, $J(\tau)^{-1}X(\tau - \mathbf{1}_{\{Y \leq X'\beta\}})$. Newey's efficient influence function is the score function evaluated at the unique representers by the Riesz-Fréchet theorem used in Severini and Tripathi (2001); a more detailed comparison of these two approaches is discussed in Severini and Tripathi (2001).

ACF (2006) show that the QR process $\hat{\beta}(\cdot)$ is asymptotically mean-zero Gaussian with the covariance function $J(\tau_1)^{-1}\Gamma(\tau_1, \tau_2)J(\tau_2)^{-1}$ for any $\tau_1, \tau_2 \in \mathcal{T}$ which is the semiparametric efficiency bound in Theorem A.3. This asymptotic covariance under misspecification for a single quantile, $J(\tau)^{-1}\Gamma(\tau, \tau)J(\tau)^{-1}$, has been presented in Hahn (1997) and Kim and White (2003). Hahn (1997) further shows the QR estimator is well approximated by the bootstrap distribution even when the linear quantile restriction is misspecified. An alternative estimator for the misspecification-robust asymptotic covariance matrix of $\hat{\beta}(\tau)$ is the nonparametric kernel method in ACF (2006).

3.3.2 QR for Correct Linear Specification

Assuming the model is correctly specified, i.e., $Q_\tau(Y|X) = X'\beta(\tau)$ almost surely, the asymptotic covariance for the QR process $\hat{\beta}(\cdot)$ derived by ACF (2006) is simplified to $J(\tau_1)^{-1}\Gamma_0(\tau_1, \tau_2)J(\tau_2)^{-1}$, where $\Gamma_0(\tau_1, \tau_2) := [min(\tau_1, \tau_2) - \tau_1\tau_2] \cdot E[XX']$ for any $\tau_1, \tau_2 \in (0, 1)$. The asymptotic covariance $J(\tau)^{-1}\Gamma_0(\tau, \tau)J(\tau)^{-1}$ for a single quantile τ, first derived by Powell (1986), is widely used for inference in most empirical studies which implicitly assume correct specification.

The semiparametric efficiency bound for the correctly specified quantile regression (3.4) is $\tau(1-\tau)\Big\{E\big[XX'f^2_{\epsilon_\tau}(0|X)\big]\Big\}^{-1}$, where $f_Y(X'\beta(\tau)|X) = f_{\epsilon_\tau}(0|X)$ a.s. and $E\big[XX'f^2_{\epsilon_\tau}(0|X)\big]$ is assumed to be finite and nonsingular. This is first calculated by Newey and Powell (1990) by the method developed in Newey (1990). If, in addition, the conditional density function

of Y given X is independent of X, i.e., $f_{\epsilon_\tau}(\cdot|X) = f_{\epsilon_\tau}(\cdot)$, the "homoskedastic" condition in QR, and $f_{\epsilon_\tau}(0) > 0$, the semiparametric efficiency bound becomes $\frac{\tau(1-\tau)}{f^2_{\epsilon_\tau}(0)}\Big(E[XX']\Big)^{-1}$. This asymptotic covariance is attained by $\hat{\beta}(\tau)$ first shown in Koenker and Bassett (1978), who assume a homoskedastic, correctly specified linear quantile regression model. This has an interesting resemblance to the fact that the OLS estimator is semiparametrically efficient in a homoskedastic regression model, i.e., $e = Y - X'\beta$, $E[e|X] = 0$, and $E[e^2|X] = E[e^2]$.

I further show, in general, the semiparametrically efficient joint asymptotic covariance of the estimators for $(\beta'(\tau_1), ..., \beta'(\tau_m))'$ is

$$[min(\tau_i, \tau_j) - \tau_i\tau_j]\Big\{E\big[XX' f_{\epsilon_{\tau_i}}(0|X) f_{\epsilon_{\tau_j}}(0|X)\big]\Big\}^{-1} \tag{3.8}$$

for any $\tau_i, \tau_j \in \mathcal{T}$, $i, j = 1, 2, ..., m$, for any finite integer $m \geq 1$. The regularity conditions imposed, (R1), (R2), and (R4), are weaker than the assumptions in Newey and Powell (1990); for example, they assume $f(\epsilon, X)$ is absolutely continuous in ϵ which implies uniform continuity in (R2). See Appendix B for the detailed proof for (3.8).

3.4 Discussion and Conclusion

Misspecification is a generic phenomenon; especially in quantile regression (QR), the true conditional quantile function (CQF) might be nonlinear, or different functions of the covariates at different quantiles. Table 3.1 summarizes the parallel properties of QR and OLS. Under misspecification, the pseudo-true OLS coefficient can be interpreted as the best linear predictor of the conditional mean function, $E[Y|X]$, in the sense that the coefficient minimizes the mean-squared error of the linear approximation to the conditional mean. With respect to the QR counterpart, I present the inverse-density-weighted mean-squared error loss function based on the distribution error $F_Y(X'\beta|X) - \tau$. The equivalent first-order condition for QR is analogous to the unconditional moment for OLS, which is the orthogonality condition for the covariates and error, the deviation of the approximation from the true conditional mean. The approximation properties of OLS have been well studied (see, for example, White (1980)). My results imply that Koenker and Bassett (1978)'s

QR estimator $\hat{\beta}(\tau)$ is semiparametrically efficient for misspecified linear projection models and "homoskedastic" correctly specified linear quantile regression models. Alternatively, the smoothed empirical likelihood estimator using the unconditional moment restriction in Whang (2006) has the same asymptotic distribution as Koenker and Bassett (1978)'s estimator and hence attains the efficiency bound.

Under correct specification (i.e., the linear quantile regression model) Koenker and Bassett's estimator consistently estimates the true $\beta(\tau)$, although it is not semiparametrically efficient given heteroskedasticity. Researchers have proposed many efficient estimators for the correctly specified linear quantile regression parameter, for example, the one-step score estimator in Newey and Powell (1990), the smoothed conditional empirical likelihood estimator in Otsu (2008), and the sieve minimum distance (SMD) estimator in Chen and Pouzo (2009). However, for all these estimators, the pseudo-true values under misspecification are different and their interpretations have not been thoroughly studied. So the semiparametric efficiency bounds of these pseudo-true values are also different. For example, the SMD estimator converges to a pseudo-true value β_{SMD} that minimizes $E\big[\big(F_Y(X'\beta|X)-\tau\big)^2\big]$. The first-order condition is $E\Big[X\big(F_Y(X'\beta_{SMD}|X)-\tau\big)\cdot f_Y(X'\beta_{SMD}|X)\Big]=0$, which is the unconditional moment used in Newey and Powell (1990) for the semiparametrically efficient GMM estimator under correct specification. The conditional density weight is similar to the generalized least squares (GLS) in the mean regression in that it uses a weight function of the conditional variance to construct an efficient estimator.

It is interesting to note that the pseudo-true value of the SMD estimator minimizes $E\big[\big(F_Y(X'\beta|X)-\tau\big)^2\big] \approx E\big[f_Y^2(Q_\tau(Y|X)|X)\big(X'\beta - Q_\tau(Y|X)\big)^2\big]$. The distribution error is weighted evenly over the support of X for β_{SMD}, in contrast to the QR parameter, which is weighted more at points with smaller conditional density in Theorem 3.1. Therefore, the SMD estimator might have more desirable and reasonable approximation properties than QR. Nevertheless, the SMD estimator is computationally more demanding than the Koenker and Bassett (1978) (KB) estimator. A numerical example in Figure 3.1 illustrates how KB and SMD estimators approximate the CQF and the conditional cdf. The red solid line is

for the QR parameter β_{KB}, defined in (3.1) and estimated by Konenker and Bassett (KB) (1978). The blue dashed line is the approximation by the SMD estimator β_{SMD} minimizing $E[(F_Y(X'\beta|X) - \tau)^2]$. The left panel shows the linear approximations $X'\beta_{KB}$, $X'\beta_{SMD}$, and the true CQF. The right panel shows the corresponding conditional cdfs $F_Y(X'\beta_{KB}|X)$ and $F_Y(X'\beta_{SMD}|X)$. For smaller x where the conditional density is larger, the specification error of SMD is smaller than that of KB in the left panel. For the distribution error in the right panel, SMD weights more evenly over the support of X, while KB has smaller distribution error at larger x with smaller density. This numerical example is constructed by $e|X = x \sim Uniform[0, x]$, $X \sim Uniform[1, 2]$, and $Y = \cos(2X) + e$. So $f_Y(y|X) = 1/X$, $F_Y(y|X) = (y - \cos(2X))/X$, and $Q_\tau(Y|X) = \tau X + \cos(2X)$. Set $\tau = 0.5$ for the median. The approximations are $X'\beta_{KB} = -0.324 + 0.161X$ and $X'\beta_{SMD} = -0.204 + 0.078X$.

This discussion leads to open-ended questions: What is an appropriate linear approximation or a meaningful summary statistic for the nonlinear CQF? How should economists measure the marginal effect of the covariates on the CQF? An approach that circumvents this problem is measuring the average marginal response of the covariates on the CQF directly. The *average quantile derivative*, defined as $E[W(X)\nabla Q_\tau(Y|X)]$ where $W(X)$ is a weight function, offers such a succinct summary statistic (Chaudhuri et al. (1997)). Lee (2011) shows that the nonparametric estimator for the density-weighted average quantile derivative enjoys $\sqrt{n}$-consistency, asymptotic normality, and semiparametric efficiency.

	OLS	QR
	Linear Projection Model	
objective minimized	$E[(Y - X'\beta)^2]$	$E[\rho_\tau(Y - X'\beta)]$
(interpretation)	$E[(E[Y\|X] - X'\beta)^2]$	$E[\bar{w}_\tau \cdot (Q_\tau(Y\|X) - X'\beta)^2]$
		$E[f_Y(X'\beta(\tau)\|X)^{-1} \cdot (F_Y(X'\beta\|X) - \tau)^2]$
unconditional moment	$E[X(Y - X'\beta)] = 0$	$E[X(\mathbf{1}_{\{Y \le X'\beta(\tau)\}} - \tau)] = 0$
(interpretation)	$E[X(E[Y\|X] - X'\beta)] = 0$	$E[X(F_Y(X'\beta(\tau)\|X) - \tau)] = 0$
		$E\big[\bar{w}_\tau \cdot X\big(X'\beta(\tau) - Q_\tau(Y\|X)\big)\big] = 0$
efficient estimators	$\arg\min_{\beta \in R^d} \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\beta)^2$	$\arg\min_{\beta \in R^d} \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - X_i'\beta)$
	$= (\sum_{i=1}^n X_i X_i')^{-1}(\sum_{i=1}^n X_i Y_i)$	(Koenker & Bassett, Whang)
asymptotic covariance	$Q^{-1}\Omega Q^{-1}$ *	$J^{-1}\Gamma J^{-1}$
efficiency bounds	Chamberlain (1987)	Theorem A.3
	Linear Regression Model	
conditional moment	$E[Y\|X] = X'\beta$	$Q_\tau(Y\|X) = X'\beta(\tau)$ or $F_Y(X'\beta(\tau)\|X) = \tau$
efficiency bounds	Chamberlain (1987) †	Newey & Powell (1990)
	Homoskedastic Linear Regression Model	
condition	$var[Y\|X] = \sigma^2$	$f_{\epsilon_\tau}(0\|X) = f_{\epsilon_\tau}(0)$
efficient estimators	OLS	Koenker & Bassett

Table 3.1 Summary Properties of OLS and QR

* $Q = E[XX']$ and $\Omega = E[XX'e^2]$ where $e = Y - X'\beta$.

† Feasible GLS estimator is semiparametrically efficient, for example.

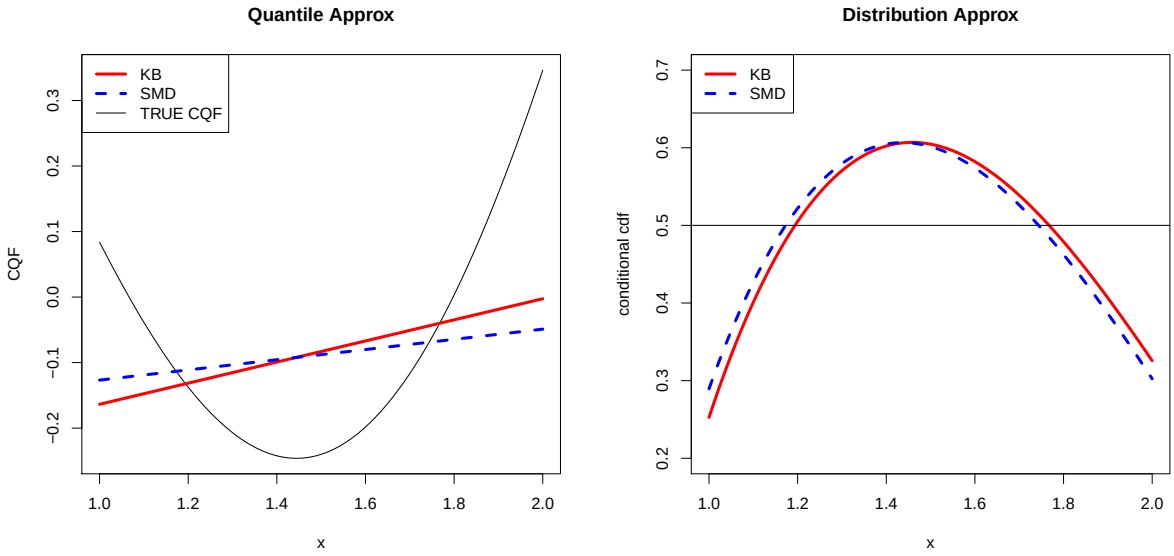

Figure 3.1 Approximations by the linear quantile regression and the sieve minimum distance estimators

DISCARD THIS PAGE

Appendix A: Supplmentary Appendix to Chapter 1

The Appendix is organized as follows. Section A.1 collects the lemmas of stochastic equicontinuity, whose proofs are collected in Section A.6. Section A.2 is the proof of Theorem 1.1, the weak convergence of the partial mean process. Section A.3 is the proof of Theorem 1.2 for generated regressors. Section A.4 is the proof of Theorem 1.3 for estimating the weight. Section A.5 collects the proofs for the inference for the policy effect in Section 1.6.

Let $(Z_1, Z_1, ..., Z_n)$ be an i.i.d. sequence of random variables taking values in $(\mathcal{Z}, \mathcal{B})$ with distribution P. For some measurable function $\phi : \mathcal{Z} \to \mathbb{R}$, define $G_n\phi = \sqrt{n}(P_n - P)\phi$ for the empirical process at ϕ. Define $\bar{o}_p(a_n)$ as $o_p(a_n)$ uniformly in $y \in \mathcal{Y}$. Let $f_{T|\Lambda}(t|\lambda)$ be the density with respect to a σ-finite measure $\mu_\Lambda(\cdot)$ of T conditional on $\Lambda = \lambda$ and evaluated at $t \in \mathcal{T}$. Let $\|\cdot\|_2$ be the $L_2(P)$ norm, i.e., $\|f\|_{2,P}^2 = \int f^2 dP$. When P is clear from the context, the subscript P is omitted. Let C denote a generic constant.

A.1 Stochastic Equicontinuity

The section collects results for stochastic equicontinuity. The first two lemmas use the bracketing CLT in Theorem 2.7.1 in van der Vaart and Wellner (1996). Lemma A.2 is for estimating the weight. Proposition 3 is for generated regressors in the second-step regression, which serves as an intermediate step for Theorem 1.2. Lemmas A.3 and A.4 are modified from Lemma 1 in Mammen et al. (2012a) using chaining arguments. The conditions for the following lemmas are specified for a general function space $\mathcal{M}$. The following Remark A.1 shows that $\mathcal{C}_M^\alpha(\mathcal{S})$ will satisfy the conditions, so I assume the functions of interest belong to this space.

Remark A.1. *Assume $\mathcal{M}$ to be $\mathcal{C}_M^\alpha(\mathcal{S})$ and $\alpha > d/2$ is sufficient for the bracketing number assumption in the following lemmas for stochastic equicontinuity. By Theorem 2.7.1 of van der Vaart and Wellner (1996), there exists a constant C depending only on $M, \alpha, diam(\mathcal{S}), d$ such that $\log N(\epsilon, \mathcal{C}_M^\alpha, \|\cdot\|_\infty) \leq C\epsilon^{-d/\alpha} < C\epsilon^{-2}$ for a bounded convex $\mathcal{S}$.*

Ichimura and Lee (2010) discuss some sufficient conditions for the condition $P(\hat{V} \in \mathcal{M}) \to 1$ *in their footnote 11: Suppose a function of interest* $V \in \mathcal{M}$*. The* q*th derivative of its estimator* $\hat{V}$ *converges in probability uniformly to the* q*th derivative of* V *for any* q *such that* $q. \leq \underline{\alpha}$*. Escanciano et al. (2012) also provide similar primitive conditions in their Appendix C.*

Lemma A.1 (Stochastic Equicontinuity I)**.** *Consider any fixed* $t \in \mathcal{T}$*. Define* $\mathcal{F}$ *to be a class of uniformly bounded functions* $f : \mathcal{Y} \times \mathcal{T} \times \mathbf{\Lambda} \mapsto R$ *such that there exists an universal constant* C_L *satisfying a Hölder continuity condition:*

$$\|f(y_1, t, \cdot) - f(y_2, t, \cdot)\|_\infty \leq C_L |y_1 - y_2|^{1/2}, \tag{A.1}$$

for any $f \in \mathcal{F}$*. For each fixed* $\bar{y} \in \mathcal{Y}$*, the subclass* $\{f(\bar{y}, t, \cdot) \in \mathcal{F}\}$ *is* $\mathcal{M}$*, where the class* $\mathcal{M}$ *is a class of functions such that* $\log N(\epsilon, \mathcal{M}, \|\cdot\|_\infty) \leq C\epsilon^{-\nu}$ *for some* $\nu < 2$*.*

Suppose for any $y \in \mathcal{Y}$*,* $F_{Y|T\Lambda}(y|t, \cdot) \in \mathcal{F}$ *and* $\hat{F}_{Y|T\Lambda}(y|t, \cdot) \in \mathcal{F}$ *with probability approaching one. Suppose the weight function* $W(S_w)$ *is uniformly bounded. Then*

$$\sup_{y \in \mathcal{Y}} \Bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\hat{F}_{Y|T\Lambda}(y|t, \Lambda_i) - F_{Y|T\Lambda}(y|t, \Lambda_i) \Big) W(S_{wi}) \\ - \sqrt{n} E\Big[\Big(\hat{F}_{Y|T\Lambda}(y|t, \Lambda) - F_{Y|T\Lambda}(y|t, \Lambda) \Big) W(S_w) \Big] \Bigg| = o_p(1).$$

Remark A.2. *The stochastic equicontinuity argument is similarly used in the proof of Theorem 2.1 in Escanciano et al. (2012), using the bracketing CLT. The difference here is that the process is not indexed by the regressor. I modify Lemma B.2 in* **?** *to specify the complexity of the function space where* $F_{Y|T\Lambda}(y|t, \cdot)$ *belongs. Lemma 1 in Rothe (2010) shares the same spirit, but the "with probability approaching one" statement is needed for* $\hat{F}_{Y|T\Lambda}(y|t, \cdot) \in \mathcal{C}_M^\alpha(\mathbf{\Lambda})$ *for any* $y \in \mathcal{Y}$*.*

The followings are comments on the conditions:

1. *The condition that* $P\big(\forall y \in \mathcal{Y}, \hat{F}_{Y|T\Lambda}(y|t,\cdot) \in \mathcal{M} = \mathcal{C}_M^{\alpha}(\mathbf{\Lambda})\big) \to 1$ *can be checked by Remark A.1. Assumption 1.5 assumes* $V = F_{Y|T\Lambda}(y|t,\Lambda) \in \mathcal{C}_M^{\alpha}(\mathbf{\Lambda})$. *By Proposition 1,*

$$\left\| D^q \hat{F}_{Y|T\Lambda}(y|t,\cdot) - D^q F_{Y|T\Lambda}(y|t,\cdot) \right\|_{\infty} = O_p\Big(\sqrt{\frac{\log n}{n h^{d_2+2q}}} + h^r\Big)$$

for $q. \leq \underline{\alpha}$. *The uniform convergence is made in Assumptions 1.5 and 1.6.*

2. *In addition to* $P(\forall y \in \mathcal{Y}, \hat{F}_{Y|T\Lambda}(y|t,\cdot) \in \mathcal{M}) \to 1$, *it remains to show* $P(\forall y \in \mathcal{Y}, \hat{F}_{Y|T\Lambda}(y|t,\cdot) \in \mathcal{F}) \to 1$. *That is, the Hölder continuity needs to be satisfied with probability approaching one. I check the following sufficient high-level assumption modifying Assumption 3.4 in* **?**: *For any* $\epsilon > 0$ *and* $\delta > 0$, *there exists* n_0 *such that for all* $n \geq n_0$, *for any* $y_1, y_2 \in \mathcal{Y}$,

$$Pr\Bigg\{ \Big\| \hat{F}_{Y|T\Lambda}(y_1|t,\cdot) - \hat{F}_{Y|T\Lambda}(y_2|t,\cdot) - \big(F_{Y|T\Lambda}(y_1|t,\cdot) - F_{Y|T\Lambda}(y_2|t,\cdot)\big) \Big\|_{\infty} \leq \delta |y_1 - y_2|^{1/2} \Bigg\} \geq 1 - \epsilon. \tag{A.2}$$

(A.2) and the Hölder continuity of $F_{Y|T\Lambda}$ *imply* $\|\hat{F}_{Y|T\Lambda}(y_1|t,\cdot) - \hat{F}_{Y|T\Lambda}(y_2|t,\cdot)\|_{\infty} \leq C_L |y_1 - y_2|^{1/2}$, *with probability approaching one. (A.2) is satisfied by Chebyshev's inequality and the mean-square-errors of my kernel estimator for the regressor* $E[\mathbf{1}_{\{y_2 < Y \leq y_1\}} | T = t, \Lambda]$, *assuming* $y_1 > y_2$.

Because of the nonsmooth estimator $\hat{F}_{Y|T\Lambda}(y|t,\Lambda)$, *the function space* $\mathcal{F}$ *is allowed to be less smooth in* y *by assuming a Hölder continuity (A.1). Alternatively, as discussed in* **?**, *a smoothed cdf estimator is needed for a stronger Lipschitz continuity assumption.*

Lemma A.2 (Stochastic Equicontinuity II)**.** *The class* $\mathcal{M}$ *is a class of uniformly bounded functions such that* $\log N(\epsilon, \mathcal{M}, \|\cdot\|_{\infty}) \leq C\epsilon^{-\nu}$ *for some* $\nu < 2$. *Suppose* $W \in \mathcal{M}$, $\|\hat{W} - W\|_{\infty} = o_p(1)$, *and* $P(\hat{W} \in \mathcal{M}) \to 1$.

The function $A(y,S)$ *is uniformly bounded and satisfies a Hölder continuity:* $\|A(y_1,\cdot) - A(y_2,\cdot)\|_{\infty} \leq C_A |y_1 - y_2|^{\gamma}$, *for some constant* C_A, *positive* γ, *and any* $y_1, y_2 \in \mathcal{Y}$.

Then

$$\sup_{y \in \mathcal{Y}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A(y, S_i)\Big(\hat{W}(S_{wi}) - W(S_{wi})\Big) - \sqrt{n} E\Big[A(y,S)\Big(\hat{W}(S_w) - W(S_w)\Big)\Big] \right| = o_p(1).$$

Proposition 3 (Stochastic Equicontinuity)**.** *Suppose Assumptions 1.3, 1.4, 1.6, 1.7, and 1.8 hold.*

$$\sup_{y\in\mathcal{Y},t,\in\mathcal{T},v\in\mathcal{V}}\Bigg|\hat{F}_{Y|T\hat{V}}(y|t,v)-\hat{F}_{Y|TV}(y|t,v)-\frac{1}{f_{TV}(t,v)}E\Big[f_{T|S_v}(t|S_v)\Big(F_{Y|TS_v}(y|t,S_v)$$
$$-F_{Y|TV}(y|t,v)\Big)\Big(K_h(\hat{V}(t,S_v)-v)-K_h(V(t,S_v)-v)\Big)\Big]\Bigg|=O_p(R_n)$$

where the remaining term $R_n=O_p\big(n^{-\kappa_1}+n^{-\kappa_2}+n^{-r_2(\eta)_{min}}\big)$. $\kappa_2<\min\{1-\eta_+,2(\delta-\eta)_{min},(\delta-\eta)_{min}+\frac{1}{2}(1-\eta_+)\}$ *and* $0<(\delta-\eta)_{min}<\kappa_1<\frac{1}{2}(1-\eta_+)+(\delta-\eta)_{min}-\frac{1}{2}(\delta\beta+\xi)_{max}$, *where* $(ab)_{min}=\min_{1\le j\le d_2}(a_jb_j)$, $(ab)_{max}=\max_{1\le j\le d_2}(a_jb_j)$ *for any vectors* a,b. *Denote* $\eta_+\equiv\sum_{j=1}^{d_2}\eta_j<1$.

Lemma A.3 (Lemma 1, Mammen et al. (2012a))**.** *Suppose the conditions of Proposition 3 hold. Then*

$$\sup_{t\in\mathcal{T},v\in\mathcal{V},y\in\mathcal{Y},V_1,V_2\in\bar{\mathcal{M}}_n}\Bigg|\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{Y_j\le y\}}K_h(T_j-t)\Big(K_h(V_1(T_j,S_{vj})-v)-K_h(V_2(T_j,S_{vj})-v)\Big)$$
$$-E\Big[\mathbf{1}_{\{Y\le y\}}K_h(T-t)\Big(K_h(V_1(T,S_v)-v)-K_h(V_2(T,S_v)-v)\Big)\Big]\Bigg|=O_p(n^{-\kappa_1}).$$

and

$$\sup_{t\in\mathcal{T},v\in\mathcal{V},V_1,V_2\in\bar{\mathcal{M}}_n}\Bigg|\frac{1}{n}\sum_{j=1}^{n}K_h(T_j-t)\Big(K_h(V_1(T_j,S_{vj})-v)-K_h(V_2(T_j,S_{vj})-v)\Big)$$
$$-E\Big[K_h(T-t)\Big(K_h(V_1(T,S_v)-v)-K_h(V_2(T,S_v)-v)\Big)\Big]\Bigg|=O_p(n^{-\kappa_1}).$$

Lemma A.4. *Suppose the conditions of Proposition 3 hold. Then for any* $t\in\mathcal{T}$,

$$\sup_{y\in\mathcal{Y},V_1,V_2\in\bar{\mathcal{M}}_n}\sup_{S\in\mathcal{S}}\Bigg|\frac{1}{n}\sum_{i=1}^{n}A(y,t,W_i,V_i;S_v)\big(K_h(V_1(t,S_v)-V_i)-K_h(V_2(t,S_v)-V_i)\big)$$
$$-E_{WV}\Big[A(y,t,W,V;S_v)\big(K_h(V_1(t,S_v)-V)-K_h(V_2(t,S_v)-V)\big)\Big]\Bigg|=O_p(n^{-\kappa_1})$$

where

$$A(y,t,W_i,V_i;S_v) \equiv \frac{W_i \ f_{T|S_v}(t|S_v)}{f_{TV}(t,V_i)}\Big(F_{Y|TS_v}(y|t,S_v) - F_{Y|TV}(y|t,V_i)\Big).$$

Lemma A.5. *Suppose the conditions of Proposition 3 hold. Then for any $t \in \mathcal{T}$,*

$$\sup_{y\in\mathcal{Y}}\Bigg|\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{V}(T_i,S_{vi}) - V(T_i,S_{vi})\Big)'\nabla_v F_{Y|TV}(y|t,V(T_i,S_{vi}))\cdot W_i$$
$$- E\Big[\Big(\hat{V}(T,S_v) - V(T,S_v)\Big)'\nabla_v F_{Y|TV}(y|t,V(T,S_v))\cdot E[W|T,S_v]\Big]\Bigg| = O_p(n^{-\kappa_{10}}),$$

where $\kappa_{10} < \frac{1}{2} + (\delta)_{min} - \frac{1}{2}(\delta\beta + \xi)_{max}$.

A.2 Proof of Theorem 1.1

The proof follows the decomposition and linearization in Theorem 1 in Rothe (2010). The main difference is that the influence function is not standard Donsker and contains the kernel and bandwidth. And the stochastic equicontinuity argument is modified.

$$\sqrt{nh^{d_t}}(\hat{\theta}_t(y|\Lambda,W) - \theta_t(y|\Lambda,W)) = \sqrt{nh^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y|T\Lambda}(y|t,\Lambda_i)W_i - EF_{Y|T\Lambda}(y|t,\Lambda)W\Big)$$
$$= \sqrt{h^{d_t}}\ G_n\big[\hat{F}_{Y|T\Lambda}(y|t,\Lambda_i)W_i - F_{Y|T\Lambda}(y|t,\Lambda_i)W_i\big] \tag{A.3}$$
$$+ \sqrt{h^{d_t}}\ G_n\big[F_{Y|T\Lambda}(y|t,\Lambda_i)W_i\big] \tag{A.4}$$
$$+ \sqrt{nh^{d_t}}\ E\big[\hat{F}_{Y|T\Lambda}(y|t,\Lambda)W - F_{Y|T\Lambda}(y|t,\Lambda)W\big]. \tag{A.5}$$

The first term (A.3) is $\bar{o}_p(\sqrt{h^{d_t}})$ by the stochastic equicontinuity result in Lemma A.1. The second term (A.4) is $\bar{O}_p(\sqrt{h^{d_t}}) = \bar{o}_p(1)$, by the Donsker property of $\mathcal{F} \equiv \{F_{Y|T\Lambda}(y|t,\Lambda(S))$ $W(S_w) : (S,S_w) \mapsto R, y \in \mathcal{Y}\}$, for any $t \in \mathcal{T}$. The asymptotic distribution is dominated by the third term (A.5). I derive the functional central limit theorem for (A.5) in the following section.

A.2.1 (A.5)

Denote $E[W|\Lambda = \lambda] = W_\Lambda(\lambda)$.

$$E\big[\hat{F}_{Y|T\Lambda}(y|t,\Lambda)W - F_{Y|T\Lambda}(y|t,\Lambda)W\big]$$

$$= \int \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)\ K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{\hat{f}_{T\Lambda}(t,\lambda)}dF_\Lambda(\lambda) \tag{A.6}$$

$$+ \int \frac{1}{n}\sum_{i=1}^{n}\Big(F_{Y|T\Lambda}(y|t,\Lambda_i) - F_{Y|T\Lambda}(y|t,\lambda)\Big)\ K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{\hat{f}_{T\Lambda}(t,\lambda)}dF_\Lambda(\lambda). \tag{A.7}$$

For the first term (A.6),

$$\int \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)\ K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{\hat{f}_{T\Lambda}(t,\lambda)}dF_\Lambda(\lambda)$$

$$= \int \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)\ K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f_{T\Lambda}(t,\lambda)}dF_\Lambda(\lambda) \tag{A.8}$$

$$- \int \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)\Big(\hat{f}_{t\lambda} - f_{t\lambda}\Big)K_h(t-T_i)K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f_{t\lambda}^2}dF_\Lambda(\lambda) \tag{A.9}$$

$$+ O_p(\|\hat{f}_{T\Lambda} - f_{T\Lambda}\|_\infty^2)$$

uniformly in y, t, where $f_{t\lambda} \equiv f_{T\Lambda}(t,\lambda)$. Because $|\mathbf{1}_{\{Y_i\le y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)| \le 1$ and integration takes over a compact set, the last term is made $o_p((nh^{d_t})^{-1/2})$ by Assumption 1.6 (iii). I will show (A.8) contributes the main influence function $\psi_{tin}(y)$. (A.9) will be of smaller order by the U-process theory. (A.7)$= o_p((nh^{d_t})^{-1/2})$ by similar arguments.

Define $u(\lambda) = \frac{W_\Lambda(\lambda)f_\Lambda(\lambda)}{f_{T\Lambda}(t,\lambda)}$. Using the standard technique in the kernel literature, i.e., change of variables, Taylor expansion, and the dominated convergence theorem,

$$\int K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f_{T\Lambda}(t,\lambda)}dF_\Lambda(\lambda) = \int K(v)\ u(\Lambda_i + vh)dv = u(\Lambda_i) + O_p(h^r).$$

Then (A.8) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)K_h(T_i-t)\Big(\frac{W_\Lambda(\Lambda_i)}{f_{T|\Lambda}(t|\Lambda_i)}+O_p(h^r)\Big)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)K_h(T_i-t)\frac{W_\Lambda(\Lambda_i)}{f_{T|\Lambda}(t|\Lambda_i)}+o_p\big((nh^{d_t})^{-1/2}\big), \qquad (A.10)$$

by Assumption 1.6 (ii).

Lemma A.6 (Functional Central Limit Theorem)**.** *The process* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(\cdot)$ *weakly converges to a Gaussian process* $\mathbb{G}_t(\cdot)$ *as defined in Theorem 1.1, where*

$$\Big\{\psi_{tin}(y)\equiv\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)\cdot\frac{1}{\sqrt{h^{d_t}}}K\Big(\frac{T_i-t}{h}\Big)\frac{W_\Lambda(\Lambda_i)}{f_{T|\Lambda}(t|\Lambda_i)}:y\in\mathcal{Y}\Big\}.$$

The proof is in Section A.2.1.2. The bias is dominated by the bias of the influence function. For $d_t=1$,

$$\frac{1}{\sqrt{h}}E\psi_{tin}(y)=E\bigg[\Big(\mathbf{1}_{\{Y\le y\}}-F_{Y|T\Lambda}(y|t,\Lambda)\Big)\cdot\frac{1}{h}K\Big(\frac{T-t}{h}\Big)\frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\bigg]$$
$$=E\bigg[\Big(F_{Y|T\Lambda}(y|T,\Lambda)-F_{Y|T\Lambda}(y|t,\Lambda)\Big)\cdot\frac{1}{h}K\Big(\frac{T-t}{h}\Big)\frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\bigg]$$
$$=E\bigg[\Big(\partial_t^r\Big(F_{Y|T\Lambda}(y|t,\Lambda)\cdot f_{T|\Lambda}(t|\Lambda)\Big)-F_{Y|T\Lambda}(y|t,\Lambda)\cdot\partial_t^r f_{T|\Lambda}(t|\Lambda)\Big)\frac{h^r}{r!}\int u^rK(u)du\frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\bigg]$$
$$=Ch^rE\Big[\partial_t^rF_{Y|T\Lambda}(y|t,\Lambda)\cdot W_\Lambda(\Lambda)\Big]=Ch^r\frac{\partial^r}{\partial t^r}E\Big[F_{Y|T\Lambda}(y|t,\Lambda)\cdot W_\Lambda(\Lambda)\Big], \qquad (A.11)$$

where the rth-order derivative with respect to t is $\partial_t^r\equiv\frac{\partial^r}{\partial t^r}$.

I now show (A.9) is $o_p((nh^{d_t})^{-1/2})$,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)K_h(t-T_i)$$
$$\times\int K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f_{T\Lambda}^2(t,\lambda)}\Big(\frac{1}{n}\sum_{j=1}^{n}K_h(\Lambda_j-\lambda)K_h(T_j-t)-f_{T\Lambda}(t,\lambda)\Big)dF_\Lambda(\lambda)$$
$$\equiv A_{21}-A_{22}.$$

Define here $u(\lambda) = \frac{f_\Lambda(\lambda)W_\Lambda(\lambda)}{f^2_{T\Lambda}(t,\lambda)}$.

$$\int K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f^2_{T\Lambda}(t,\lambda)}K_h(\Lambda_j-\lambda)dF_\Lambda(\lambda) = \int K(v)\ K_h(\Lambda_j-\Lambda_i-hv)\ u(\Lambda_i+hv)dv$$
$$= u(\Lambda_i)\bar{K}_h(\Lambda_i-\Lambda_j) + \int u'(\bar{\Lambda}_i)vK(v)K\Big(\frac{\Lambda_j-\Lambda_i}{h}-v\Big)dv,$$

where $\bar{\Lambda}_i$ is between Λ_i and Λ_i+hv. Since the kernel is of bounded support, the second term is $o_p(h)$. The convolution kernel is defined as

$$\bar{K}_h(\Lambda_i-\Lambda_j) \equiv \frac{1}{h^{d_\lambda}}\bar{K}\Big(\frac{\Lambda_i-\Lambda_j}{h}\Big) = \frac{1}{h^{d_\lambda}}\int K(v)K\Big(v-\frac{\Lambda_i-\Lambda_j}{h}\Big)dv \tag{A.12}$$
$$= \frac{1}{h^{2d_\lambda}}\int K\Big(\frac{\Lambda_i-\lambda}{h}\Big)K\Big(\frac{\Lambda_j-\lambda}{h}\Big)d\lambda. \tag{A.13}$$

For A_{22},

$$\int K_h(\lambda-\Lambda_i)\frac{W_\Lambda(\lambda)}{f_{T\Lambda}(t,\lambda)}f_\Lambda(\lambda)d\lambda = f_{T\Lambda}(t,\Lambda)_i)u(\Lambda_i) + O_p(h^r).$$

Therefore, (A.9) becomes

$$\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{y|t\Lambda_i}\Big)u(\Lambda_i)K_h(T_i-t)\cdot\Big(\bar{K}_h(\Lambda_i-\Lambda_j)K_h(T_j-t)-f_{T\Lambda}(t,\Lambda)_i)\Big)$$
$$+\bar{o}_p((nh^{d_t})^{-1/2}).$$

For the projection of the U-process, define

$$\bar{f}_{T\Lambda}(t,\Lambda_i) \equiv E\Big(\bar{K}_h(\Lambda_i-\Lambda_j)\cdot K_h(T_j-t)\Big|Z_i\Big)$$
$$= \int\int\frac{1}{h^{d_t}}K\Big(\frac{T-t}{h}\Big)\cdot\frac{1}{h^{2d_\lambda}}\int K\Big(\frac{\Lambda_i-v}{h}\Big)K\Big(\frac{\Lambda-v}{h}\Big)dv\cdot f_{T\Lambda}(T,\Lambda)d\Lambda dT$$
$$= \int\frac{1}{h^{d_\lambda}}K\Big(\frac{\Lambda_i-v}{h}\Big)\int\int K(r)K(s)\ f_{T\Lambda}(t+rh,v+sh)drds\ dv$$
$$= \int K(w)f_{T\Lambda}(t,\Lambda_i+wh)dw + O_p(h^r) = f_{T\Lambda}(t,\Lambda)_i) + O_p(h^r).$$

Define $H(Z_i,Z_j;y,h) \equiv$

$$\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)\ u(\Lambda_i)K_h(T_i-t)\cdot\Big(\bar{K}_h(\Lambda_i-\Lambda_j)\cdot K_h(T_j-t)-\bar{f}_{T\Lambda}(t,\Lambda_i)\Big).$$

Then (A.9) becomes

$$\frac{1}{n^2}\sum_i\sum_{j\neq i} H(Z_i,Z_j;y,h)+\frac{1}{n^2}\sum_i H(Z_i,Z_i;y,h)+\bar{o}_p((nh^{d_t})^{-1/2}). \tag{A.14}$$

The second term in (A.14) is $o_p(1/\sqrt{nh^{d_t}})$ uniformly in y, because its second part is $\frac{1}{n}$(A.10) and its first part is smaller than $\frac{1}{n^2}\sum_i u(\Lambda_i)K_h^2(T_i-t)\frac{1}{h^{d_t}}\int K^2(v)dv = o_p((nh^{d_t})^{-1/2})$. The first term in (A.14) is a degenerate second order U-process. By Corollary 4 (ii) in Sherman (1994),

$$\sup_{y\in\mathcal{Y}}\Big|\frac{1}{n^2}\sum_i\sum_{j\neq i}\sqrt{h^{d_\lambda+2d_t}}H(Z_i,Z_j;y,h)\Big| = O_p\Big(\frac{1}{n}\Big).$$

Therefore,

$$\sup_{y\in\mathcal{Y}}\Big|\frac{1}{n^2}\sum_i\sum_{j\neq i}H(Z_i,Z_j;y,h)\Big| = O_p\Big(\frac{1}{n\sqrt{h^{d_\lambda+2d_t}}}\Big) = o_p\Big(\frac{1}{\sqrt{nh^{d_t}}}\Big),$$

which is implied by Assumption 1.6.

By similar reasoning, (A.7) can be shown to be $o_p((nh^{d_t})^{-1/2})$ uniformly in y.

A.2.1.1 Applying Corollary 4 in Sherman (1994)

The class of P-degenerate functions

$$\mathcal{H}\equiv\{\sqrt{h^{d_\lambda+2d_t}}H(Z_i,Z_j;y,h): y\in\mathcal{Y}\}$$

has an envelope $F(\Lambda_i,\Lambda_j) = (h^{d_\lambda+2d_t})^{-1/2}\ u(\Lambda_i)K(T_i-t)\cdot\bar{K}(\Lambda_i-\Lambda_j)\cdot K(T_j-t)$. Let $\mathcal{H}$ be a real-valued functions on $\mathcal{S}^2=\mathcal{S}\otimes\mathcal{S}$. And $P^2=P\otimes P$ denotes the product measure. We then show $\mathcal{H}$ is Euclidean for this envelope F satisfying $EF^2=P^2F^2<\infty$.

First, $\{\mathbf{1}_{\{Y_i\leq y\}}: y\in\mathcal{Y}\}$ and $\{F_{Y|T\Lambda}(y|t,\Lambda_i): y\in\mathcal{Y}\}$ are manageable by the fact that they are monotone increasing in y (p.221 in Kosorok (2008)). And $F(\Lambda_i,\Lambda_j)$ is a $\mathbb{R}$-valued function on the underlying probability space. By applying Lemma E1 in Andrews and Shi (2011), $\mathcal{H}$ is Euclidean.

We then check $EF^2<\infty$. First, we calculate

$$\frac{1}{h^{d_t}}\int K^2\Big(\frac{T-t}{h}\Big)\ f_{T|\Lambda}(T|\Lambda)dT = \int K^2(u)du\cdot f_{T|\Lambda}(t|\Lambda)+O(h).$$

W.L.O.G., we could work on the case $d_\lambda = 1$ for expositional simplicity. The extension to $d_\lambda > 1$ is straightforward.

$$\begin{aligned}
&\int \frac{1}{h} f_{T|\Lambda}(t|\Lambda) f(\Lambda) \int K(v) K\big(v - \frac{\Lambda_i - \Lambda}{h}\big) dv \cdot \int K(u) K\big(u - \frac{\Lambda_i - \Lambda}{h}\big) du \; d\Lambda \\
&\quad = \int\int\int f_{T|\Lambda}(t|(s-v)h + \Lambda_i) f((s-v)h + \Lambda_i) \cdot K(s) K(u+s-v) ds \; K(v) dv \; K(u) du \\
&\quad = f_{T|\Lambda}(t|\Lambda_i) f(\Lambda_i) \int\int\int K(s) K(u+s-v) K(v) K(u) ds du dv + O(h)
\end{aligned}$$

by change of variables $s \equiv v - (\Lambda_i - \Lambda)/h$. Therefore, $EF^2 =$

$$\Big(\int K^2(u) du\Big)^2 \int f^2_{T|\Lambda}(t|\Lambda) f^2(\Lambda) u^2(\Lambda) d\Lambda \int\int\int K(s) K(u+s-v) K(v) K(u) ds du dv < \infty.$$

A.2.1.2 Proof of Lemma A.6 (FCLT)

Note this proof is for the influence function and robust to the nonparametric estimator. For all $\omega \in \Omega$, the triangular array $f_{ni}(\omega, y, t) \equiv \frac{1}{\sqrt{n}} \psi_{tin}(y) = \big(\mathbf{1}_{\{Y_i(\omega) \le y\}} - F_{Y|T\Lambda}(y|t, \Lambda_i(\omega))\big) \cdot \frac{1}{\sqrt{nh^{d_t}}} K\Big(\frac{T_i(\omega) - t}{h}\Big) \frac{W_\Lambda(\Lambda(\omega))}{f_{T|\Lambda}(t|\Lambda_i(\omega))}$ are independent within rows. Define the $n \times 1$ vector $f_n(\omega, y, t) \equiv \big(f_{n1}(\omega, y, t), ..., f_{nn}(\omega, y, t)\big)'$ and the random set $\mathcal{F}_{n\omega,t} \equiv \big\{f_n(\omega, y, t) : y \in \mathcal{Y}\big\}$. I skip the subscript t for notational ease without loss of clarity.

(i) The triangular array processes $\{f_{ni}(\omega, y)\}$ are manageable with respect to the envelopes $F_{ni}(\omega) \equiv \frac{1}{\sqrt{nh^{d_t}}} K\Big(\frac{T_i(\omega) - t}{h}\Big) \frac{W_\Lambda(\Lambda(\omega)}{f_{T|\Lambda}(t|\Lambda_i(\omega))}$.

First, $\{\mathbf{1}_{\{Y_i \le y\}} : y \in \mathcal{Y}, i = 1, ..., n\}$ and $\{F_{Y|T\Lambda}(y|t, \Lambda_i) : y \in \mathcal{Y}, i = 1, ..., n\}$ are manageable by the fact that they are monotone increasing in y (p.221 in Kosorok (2008). And $F_n(\omega) \equiv (F_{n1}, ..., F_{nn})^T$ is a $\mathbb{R}^n$-valued function on the underlying probability space. Then (i) is proved by applying Lemma E1 in Andrews and Shi (2011).

Before I proceed to check the next conditions, it will be convenient to calculate the following expectations. Define $V(y, T, \Lambda) \equiv \big(F_{Y|T\Lambda}(y|T, \Lambda) - F_{Y|T\Lambda}(y|t, \Lambda)\big) \cdot f_{T|\Lambda}(T|\Lambda)$. By assumption, $\frac{\partial^r}{\partial T^r} V(y, T, \Lambda)$ is bounded uniformly over y, T, Λ, and $f_{T|\Lambda}$ is uniformly bounded

away from zero. Then

$$
\begin{aligned}
Ef_{ni}(y) &= \frac{1}{\sqrt{n}} E\psi_{tin}(y) \\
&= \sqrt{\frac{h^{d_t}}{n}} E\left[\int K(u)\Big(F_{Y|T\Lambda}(y|t+uh,\Lambda) - F_{Y|T\Lambda}(y|t,\Lambda)\Big)\cdot f_{T|\Lambda}(t+uh|\Lambda)du\ \frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\right] \\
&= \sqrt{\frac{h^{d_t}}{n}}\frac{h^r}{r!}\cdot\int K(u)u^r du\cdot E\left[\frac{\partial^r}{\partial T^r}V(y,T,\Lambda)\Big|_{T=t}\cdot\frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\right] = O\Big(h^r\sqrt{\frac{h^{d_t}}{n}}\Big)
\end{aligned}
$$

uniformly in y.

For any $t, s \in \mathcal{T}$,

$$
\begin{aligned}
&E\big[f_{ni}(y_1,t)f_{ni}(y_2,s)\big] = \frac{1}{n}E\big[\psi_{tin}(y_1)\psi_{sin}(y_2)\big] \\
&= \frac{1}{n}E\left[\Big(\mathbf{1}_{\{Y\le y_1\}} - F_{Y|T\Lambda}(y_1|t,\Lambda)\Big)\Big(\mathbf{1}_{\{Y\le y_2\}} - F_{Y|T\Lambda}(y_2|s,\Lambda)\Big)\frac{1}{h^{d_t}}K\Big(\frac{T-t}{h}\Big)K\Big(\frac{T-s}{h}\Big)\frac{W_\Lambda^2(\Lambda)}{f_{t|\Lambda}f_{s|\Lambda}}\right] \\
&= \frac{1}{n}E\Big[\Big(F_{y_1|T\Lambda} - F_{y_1|t\Lambda}F_{y_2|T\Lambda} - F_{y_1|T\Lambda}F_{y_2|s\Lambda} + F_{y_1|t\Lambda}F_{y_2|s\Lambda}\Big)\frac{1}{h^{d_t}}K\Big(\frac{T-t}{h}\Big)K\Big(\frac{T-s}{h}\Big)\frac{W_\Lambda^2(\Lambda)}{f_{t|\Lambda}f_{s|\Lambda}}\Big] \\
&= \frac{1}{n}E\Big[\int\Big(F_{Y|T\Lambda}(y_1|t+uh,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t+uh,\Lambda) \\
&- F_{Y|T\Lambda}(y_1|t+uh,\Lambda)F_{Y|T\Lambda}(y_2|s,\Lambda) \\
&+ F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|s,\Lambda)\Big)K(u)K\big(u+\frac{t-s}{h}\big)f_{t+uh,\Lambda}du\ \frac{W_\Lambda^2(\Lambda)}{f_{t|\Lambda}f_{s|\Lambda}}\Big] \\
&= \frac{1}{n}E\Big[\Big(F_{Y|T\Lambda}(y_1|t,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t,\Lambda)\Big)\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(s|\Lambda)} + O(h)\Big]\bar{K}\big(\frac{s-t}{h}\big)
\end{aligned}
\tag{A.15}
$$

uniformly in $y_1 \le y_2 \in \mathcal{Y}$. The convolution kernel $\bar{K}$ is defined in (A.13). When $s = t$, $\bar{K}(0) = \frac{1}{h}\int K^2\big(\frac{s-x}{h}\big)dx = \int K^2(u)du$. When $s \ne t$, $\bar{K}\big(\frac{s-t}{h}\big) = o(h)$ because K is of bounded support.

(ii) Define $\mathcal{Z}_n(y) = \sum_{i=1}^n \Big(f_{ni}(y) - Ef_{ni}(y)\Big)$. Let $y_1 \le y_2 \in \mathcal{Y}$. The covariance kernel of the limiting Gaussian process is

$$
\begin{aligned}
\lim_{h\to 0} P\mathcal{Z}_n(y_1)\mathcal{Z}_n(y_2) &= \lim_{h\to 0} E\left[\Big(\psi_{tin}(y_1) - E\psi_{tin}(y_1)\Big)\Big(\psi_{tin}(y_2) - E\psi_{tin}(y_2)\Big)\right] \\
&= \lim_{h\to 0} E\Big[\psi_{tin}(y_1)\psi_{tin}(y_2)\Big] - E\Big[\psi_{tin}(y_1)\Big]E\Big[\psi_{tin}(y_2)\Big] \\
&= \lim_{h\to 0} E\Big[\psi_{tin}(y_1)\psi_{tin}(y_2)\Big] \equiv H(y_1,y_2) \\
&= E\left[\Big(F_{Y|T\Lambda}(y_1|t,\Lambda) - F_{Y|T\Lambda}(y_1|t,\Lambda)F_{Y|T\Lambda}(y_2|t,\Lambda)\Big)\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\right]\int K^2(v)dv,
\end{aligned}
\tag{A.16}
$$

using (A.15).

(iii) Using (A.15),

$$\sum_{i=1}^{n} PF_{ni}^2 = E\Big[\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(t|\Lambda)} + O(h)\Big]\int K^2(v)dv.$$

(iv) For any $\epsilon > 0$, $\sum_{i=1}^{n} PF_{ni}^2\ \mathbf{1}\{F_{ni} > \epsilon\} \to 0$ holds. Because K is bounded, $f_{T|\Lambda}$ is bounded away from zero, and $\sqrt{nh^{d_t}} \to \infty$, $\mathbf{1}\{\frac{1}{\sqrt{nh^{d_t}}}K\Big(\frac{T_i(\omega)-t}{h}\Big)\frac{W_\Lambda(\Lambda_i(\omega))}{f_{T|\Lambda}(t|\Lambda_i(\omega))} > \epsilon\} = 0$ for n large enough.

(v) Denote $F_{Y|T,\Lambda}(y|t,\Lambda) \equiv F_{y|t,\Lambda}$.

$$\begin{aligned}
&P\Big|f_{ni}(y_1) - f_{ni}(y_2)\Big|^2 = E\Big|\frac{1}{\sqrt{nh^{d_t}}}K\Big(\frac{T-t}{h}\Big)\frac{W_\Lambda(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\cdot\Big(\mathbf{1}_{\{y_1<Y\le y_2\}} + F_{y_1|t,\Lambda} - F_{y_2|t,\Lambda}\Big)\Big|^2 \\
&= E\left[\frac{1}{nh^{d_t}}K^2\Big(\frac{T-t}{h}\Big)\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}^2(t|\Lambda)}\cdot\Big(\mathbf{1}_{\{y_1<Y\le y_2\}} + F_{y_1|t,\Lambda} - F_{y_2|t,\Lambda}\Big)^2\right] \\
&= E\Bigg[\frac{1}{nh^{d_t}}K^2\Big(\frac{T-t}{h}\Big)\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}^2(t|\Lambda)}\cdot\Big(\mathbf{1}_{\{y_1<Y\le y_2\}}\big(1 + 2F_{y_1|t,\Lambda} - 2F_{y_2|t,\Lambda}\big) - 2F_{y_1|t,\Lambda}F_{y_2|t,\Lambda} \\
&+ F_{y_1|t,\Lambda}^2 + F_{y_2|t,\Lambda}^2\Big)\Bigg] \\
&= \frac{1}{n}\int K^2(u)du\cdot E\left[\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\cdot\Big(F_{y_2|t,\Lambda} - F_{y_1|t,\Lambda} - \big(F_{y_2|t,\Lambda} - F_{y_1|t,\Lambda}\big)^2\Big)\right] + O\Big(\frac{h^{d_t}}{n}\Big)
\end{aligned}$$

uniform in y_1, y_2. The uniformity comes from the assumption that the first order partial derivative $\partial_t F_{y|t\Lambda}$ is uniformly bounded over their arguments, so that the expectations exist.

Therefore, $\rho_n(y_1, y_2) \to \rho(y_1, y_2)$

$$\equiv \left\{\int K^2(u)du\cdot E\left[\frac{W_\Lambda^2(\Lambda)}{f_{T|\Lambda}(t|\Lambda)}\cdot\Big(F_{y_2|t,\Lambda} - F_{y_1|t,\Lambda} - \big(F_{y_2|t,\Lambda} - F_{y_1|t,\Lambda}\big)^2\Big)\right]\right\}^{1/2},$$

uniformly in y_1, y_2.

A.3 Generated Regressors

Let $N(\epsilon, \mathcal{V}, \|\cdot\|)$ denote the covering number with respect to the semimetric $\|\cdot\|$ and $N_{[\cdot]}(\epsilon, \mathcal{V}, \|\cdot\|)$ be the bracketing number. I drop the subscript in S_v and denote it by S for notational ease.

A.3.1 Proof of Theorem 1.2

A.3.1.1 Functional directional derivative - estimated evaluated points

The first part uses functional directional derivative to decompose the variation from estimating the evaluated points, following the proof of Corollary 6 in Mammen et al. (2012a) for the nonparametric generated control variables for simultaneous equation models.

The true functions $\bar{f} = (\bar{f}_1, \bar{f}_2) = \Big(E\big[\mathbf{1}_{\{Y \le y\}}\big|T = t, V\big], V(S)\Big)$. Denote $f_2 = f_2(S)$, $f_1 = f_1(t, V)$, and $\theta_0 = E\big[E\big(\mathbf{1}_{\{Y \le y\}}\big|T = t, V(S)\big)W(S)\big]$. For any two functions of X, $f_1(X)$ and $f_2(X)$, denote $[f_1 + f_2](X) \equiv f_1(X) + f_2(X)$. Denote $f_1^{(v)} \equiv \nabla_v f_1(t, v)$.

Define the functional $S_n(f) \equiv \frac{1}{n}\sum_{i=1}^n f_1\big(t, f_2(S_i)\big)W_i - \theta_0$. For the estimator $\hat{f} = (\hat{f}_1, \hat{f}_2)$, I study the asymptotics of $S_n(\hat{f})$. The directional derivative

$$
\begin{aligned}
&\dot{S}_n(\bar{f})[f - \bar{f}] = \lim_{s \to 0} \frac{1}{s}\Big(S_n\big(\bar{f} + s(f - \bar{f})\big) - S_n(\bar{f})\Big) \\
&= \lim_{s \to 0} \frac{1}{s}\frac{1}{n}\sum_{i=1}^n \Big\{\Big[\bar{f}_1 + s(f_1 - \bar{f}_1)\Big]\Big(t, \big[\bar{f}_2 + s(f_2 - \bar{f}_2)\big](S_i)\Big)W_i - \bar{f}_1\big(t, \bar{f}_2(S_i)\big)W_i \\
&\qquad\qquad - \bar{f}_1\Big(t, \big[\bar{f}_2 + s(f_2 - \bar{f}_2)\big](S_i)\Big)W_i + \bar{f}_1\Big(t, \big[\bar{f}_2 + s(f_2 - \bar{f}_2)\big](S_i)\Big)W_i\Big\} \\
&= \frac{1}{n}\sum_{i=1}^n \big[f_1 - \bar{f}_1\big]\big(t, \bar{f}_2(S_i)\big)W_i + \frac{1}{n}\sum_{i=1}^n \bar{f}_1^{(v)}\big(t, \bar{f}_2(S_i)\big) \cdot \big[f_2 - \bar{f}_2\big](S_i)W_i \\
&= \frac{1}{n}\sum_{i=1}^n \big[\hat{F}_{Y|T\hat{V}}(y|t, V(S_i)) - F_{Y|TV}(y|t, V(S_i))\big]W_i \\
&+ \frac{1}{n}\sum_{i=1}^n \nabla_v F_{Y|TV}(y|t, V(S_i))'\big[\hat{V}(S_i) - V(S_i)\big]W_i \\
&\equiv T_{1,n}(f) + T_{2,n}(f).
\end{aligned}
$$

$T_{1,n}$ is for the nonparametric regression and generated regressors. $T_{2,n}$ is from the estimated evaluated points in the regression. By Lemma A.5,

$$
\sqrt{nh^{d_t}}T_{2,n} = \sqrt{nh^{d_t}}E\Big[\nabla_v F_{Y|TV}(y|t, V(T,S))'\big(\hat{V}(T,S) - V(T,S)\big)E[W|T,S]\Big] + o_p(1). \tag{A.17}
$$

Therefore, for any $f_1 = f_{1,A} + f_{2,B}$, the smaller order terms are

$$S_n(f) - S_n(\bar{f}) - \dot{S}_n(\bar{f})[f - \bar{f}] = \frac{1}{n}\sum_{i=1}^{n}\left[f_{1,A}^{(v)} - \bar{f}_1^{(v)}\right]\left(t, \bar{f}_2(S_i)\right)\cdot\left(f_2(S_i) - \bar{f}_2(S_i)\right)W_i$$
$$+ O_p(\|f_{1,B}\|_\infty) + O_p\Big(||f_2 - \bar{f}_2||_\infty^2\Big) \equiv so1.$$

By Proposition 1, $\left\|\hat{F}_{Y|TV}^{(v)} - F_{Y|TV}^{(v)}\right\|_\infty = O_p\big((\log n/[nh^{d_2+2}])^{-1/2}\big)$. Proposition 3 implies

$$f_1 - \bar{f}_1 = f_{1,A} - \bar{f}_1 + f_{1,B} = \hat{F}_{Y|T\hat{V}} - \hat{F}_{Y|TV} + \hat{F}_{Y|TV} - F_{Y|TV}$$
$$= \frac{1}{f_{TV}(t,v)}E\Big[f_{t|S}\big(F_{y|t,S} - F_{y|t,V}\big)\big(K_h(\hat{V} - v) - K_h(V - v)\big)\Big] + \hat{F}_{Y|TV} - F_{Y|TV} + O_p(R_n). \tag{A.18}$$

Let $f_{1,B} = O_p(R_n)$ and the first terms are $f_{1,A} - \bar{f}_1$. Therefore,

$$\left\|f_{1,A}^{(v)} - \bar{f}_1^{(v)}\right\|_\infty = O(h^{-2})\left\|\hat{V} - V\right\|_\infty + O_p\big((\log n/[nh^{d_2+2}])^{-1/2}\big).$$

$$\big|so1\big| \le O_p\Big(\left\|f_{1,A}^{(v)} - \bar{f}_1^{(v)}\right\|_\infty \cdot \left\|\hat{V} - V\right\|_\infty\Big) + O_p(R_n) + O_p\Big(\left\|\hat{V} - V\right\|_\infty^2\Big)$$
$$= O_p\Big(\frac{1}{h^2}\|\hat{V} - V\|_\infty^2 + \frac{\log n}{\sqrt{nh^{d_2+2}}}\|\hat{V} - V\|_\infty + R_n\Big) = O_p\Big(R_n\Big) = o_p((nh^{d_t})^{-1/2})$$

by the bandwidth assumption.

A.3.1.2 Nonparametric regression with generated regressors

Proof of Proposition 3

The following proof can allow for additional known regressors X, which is ignored for notational ease. Linearize the regression estimator for any generated regressor $V_1 \in \mathcal{V}$,

$$\hat{F}_{Y|TV_{1i}} \equiv \hat{F}_{Y|TV_1}(y|t, v_i) = \frac{\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{Y_j \le y\}}K_h(T_j - t)K_h(V_1(T_j, S_j) - v_i)}{\frac{1}{n}\sum_{j=1}^{n}K_h(T_j - t)K_h(V_1(T_j, S_j) - v_i)} \equiv \frac{\hat{g}_{1i}}{\hat{f}_{1i}}$$
$$= \frac{\hat{g}_{1i}}{f_i} + \frac{F_{Y|TV}(y|t, v_i)}{f_i}\Big(f_i - \hat{f}_{1i}\Big) + \frac{1}{f_i}\Big(f_i - \hat{f}_{1i}\Big)\Big(\hat{F}_{Y|TV_{1i}} - F_{Y|TV_i}\Big)$$

where $f_i \equiv f_{TV}(t, v_i)$ and $F_{Y|TV_i} \equiv F_{Y|TV}(y|t, v_i)$ with the true regressor $V(T, S)$. The first two terms will dominate the first-order asymptotics of $\hat{F}_{Y|TV_{1i}} - \hat{F}_{Y|TV_{2i}}$ and the third term is controlled to be a smaller-order term $so2$.

By Lemma A.3, uniformly in $y \in \mathcal{Y}$,

$$\begin{aligned}
\frac{\hat{g}_{1i} - \hat{g}_{2i}}{f_i} &= \frac{1}{f_i} E\Big[F_{Y|TS}(y|T,S) K_h(T-t)\Big(K_h(V_1(T,S)-v_i) - K_h(V_2(T,S)-v_i)\Big)\Big] + O_p(n^{-\kappa_1}) \\
&= \frac{1}{f_i} E\Big[F_{Y|TS}(y|t,S) f_{T|S}(t|S)\big(K_h(V_1(t,S)-v_i) - K_h(V_2(t,S)-v_i)\big)\Big] + O_p(n^{-\kappa_1}) + O_p(h_2^{r_2}) \\
&= O_p\Big(\frac{\|V_1 - V_2\|_\infty}{h}\Big) + O_p(n^{-\kappa_1}) + O_p(h_2^{r_2}). \\
&\frac{F_{Y|TV}(y|t,v_i)}{f_i}\Big(\hat{f}_{2i} - \hat{f}_{1i}\Big) \\
&= \frac{F_{Y|TV}(y|t,v_i)}{f_i} E\Big[K_h(T-t)\Big(K_h(V_2(T,S)-v_i) - K_h(V_1(T,S)-v_i)\Big)\Big] + O_p(n^{-\kappa_1}) + O_p(h_2^{r_2}) \\
&= \frac{F_{Y|TV}(y|t,v_i)}{f_i} E\Big[f_{T|S}(t|S)\Big(K_h(V_2(t,S)-v_i) - K_h(V_1(t,S)-v_i)\Big)\Big] + O_p(n^{-\kappa_1}) + O_p(h_2^{r_2}) \\
&= O_p\Big(\frac{\|V_1 - V_2\|_\infty}{h}\Big) + O_p(n^{-\kappa_1}) + O_p(h_2^{r_2}). \qquad \text{(A.19)}
\end{aligned}$$

Note that $E\Big[\mathbf{1}_{\{Y \le y\}} K_h(T-t) K_h(V_1(T,S) - v_i)\Big] = F_{Y|TV_1}(y|t,v_i) f_{TV_1}(t,v_i) + O(h^{r_2})$. If the Lipschitz condition was assumed on the regression function with respect to the generated regressor as Assumption 4 in Mammen et al. (2012a), then the above terms are dominated by $O_p(\|V_1 - V_2\|_\infty)$ (assuming $\kappa_1 > \delta$), instead of $O_p\Big(\|V_1 - V_2\|_\infty / h\Big)$ (assuming $\kappa_1 > \delta - \eta$).

Let $V_2 = V$. The smaller term is

$$\begin{aligned}
\big|so2\big| &\le \Big|\frac{1}{f_i}\Big(f_i - \hat{f}_{1i}\Big)\Big(\hat{F}_{Y|TV_{1i}} - F_{Y|TV_i}\Big)\Big| + \Big|\frac{1}{f_i}\Big(f_i - \hat{f}_{2i}\Big)\Big(\hat{F}_{Y|TV_{2i}} - F_{Y|TV_i}\Big)\Big| \\
&= O_p\Big(\Big|\frac{1}{f_i}\Big(f_i - \hat{f}_i + \hat{f}_i - \hat{f}_{1i}\Big)\Big(\hat{F}_{Y|TV_{1i}} - \hat{F}_{Y|TV_i} + \hat{F}_{Y|TV_i} - F_{Y|TV_i}\Big)\Big|\Big) \\
&= O_p\Big(\Big(\frac{\log n}{\sqrt{n h^{d_2}}} + \frac{\|V_1 - V\|_\infty}{h}\Big)^2\Big) = O_p(n^{-\kappa_2}).
\end{aligned}$$

by (A.19), Proposition 1, and using a bias-reducing kernel. □

For the partial sum, by Proposition 3 and Lemma A.4, uniformly in $y \in \mathcal{Y}$,

$$\begin{aligned}
&\sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n} W_i\Big(\hat{F}_{Y|T\hat{V}}(y|t,v_i) - \hat{F}_{Y|TV}(y|t,v_i)\Big) \\
&= \sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n} \frac{W_i}{f_{TV}(t,v_i)} E_S\Big[f_{T|S}(t|S)\Big(F_{Y|TS}(y|t,S) - F_{Y|TV}(y|t,v_i)\Big) \\
&\qquad\qquad \big(K_h(V_1(t,S)-v_i) - K_h(V_2(t,S)-v_i)\big)\Big] + o_p(1) \\
&= \sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n} E_S\Big[A(y,t,W_i,V_i;S)\big(K_h(V_1(t,S)-V_i) - K_h(V_2(t,S)-V_i)\big)\Big] + o_p(1) \\
&= \sqrt{nh^{d_t}} E_{WV}\Big[E_S\Big[A(y,t,W,V;S)\big(K_h(V_1(t,S)-V) - K_h(V_2(t,S)-V)\big)\Big]\Big] + o_p(1)
\end{aligned} \tag{A.20}$$

where E_{WV} denotes the expectation over the joint density of (W,V), and define

$$A(y,t,W_i,V_i;S) \equiv \frac{W_i f_{T|S}(t|S)}{f_{TV}(t,V_i)}\Big(F_{Y|TS}(y|t,S) - F_{Y|TV}(y|t,V_i)\Big).$$

The stochastic equicontinuity (A.20) is

$$\begin{aligned}
\sup_{y\in\mathcal{Y},V_1,V_2\in\bar{\mathcal{M}}_n} &\sqrt{nh^{d_t}}\Bigg|\frac{1}{n}\sum_{i=1}^{n} E_S\Big[A(y,t,W_i,V_i;S)\big(K_h(V_1(t,S)-V_i) - K_h(V_2(t,S)-V_i)\big)\Big] \\
&- E_{WV}\Big[E_S\Big[A(y,t,W,V;S)\big(K_h(V_1(t,S)-V) - K_h(V_2(t,S)-V)\big)\Big]\Big]\Bigg| = o_p(1)
\end{aligned}$$

which is implied by Lemma A.4 through

$$\begin{aligned}
\sup_{y\in\mathcal{Y},V_1,V_2\in\bar{\mathcal{M}}_n} &E_S\Bigg[\sqrt{nh^{d_t}}\Bigg|\frac{1}{n}\sum_{i=1}^{n} A(y,t,W_i,V_i;S)\big(K_h(V_1(t,S)-V_i) - K_h(V_2(t,S)-V_i)\big) \\
&- E_{WV}\Big[A(y,t,W,V;S)\big(K_h(V_1(t,S)-V) - K_h(V_2(t,S)-V)\big)\Big]\Bigg|\Bigg] = o_p(1).
\end{aligned}$$

Define $W_V(v) = E[W(S)|V = v]$. Then in (A.20),

$$
\begin{aligned}
&E_{WV}\Big[E_S\Big[A(y,t,W,V;S)\big(K_h(V_1(t,S)-V)-K_h(V_2(t,S)-V)\big)\Big]\Big] \\
&= E_{WV}\Big[\frac{W}{f_{TV}(t,V)}E_S\Big[f_{T|S}(t|S)\Big(F_{Y|TS}(y|t,S)-F_{Y|TV}(y|t,V)\Big) \\
&\quad \big(K_h(V_1(t,S)-V)-K_h(V_2(t,S)-V)\big)\Big]\Big] \\
&= E_S\Big[E_V\Big[\frac{W_V(V)}{f_{TV}(t,V)}\Big(F_{Y|TS}(y|t,S)-F_{Y|TV}(y|t,V)\Big)f_{T|S}(t|S) \\
&\quad \big(K_h(V_1(t,S)-V)-K_h(V_2(t,S)-V)\big)\Big]\Big] \\
&= E_S\Big[\int \frac{W_V(v)}{f_{TV}(t,v)}\Big(F_{Y|TS}(y|t,S)-F_{Y|TV}(y|t,v)\Big)\big(K_h(V_1(t,S)-v) \\
&\quad - K_h(V_2(t,S)-v)\big)f_V(v)dv\ f_{T|S}(t|S)\Big] \\
&= E_S\Big[\Big(\frac{W_V(V_1(t,S))}{f_{T|V}(t|V_1(t,S))}\Big(F_{Y|TS}(y|t,S)-F_{Y|TV}(y|t,V_1(S))\Big) \\
&\quad - \frac{W_V(V_2(S))}{f_{T|V}(t|V_2(t,S))}\Big(F_{Y|TS}(y|t,S)-F_{Y|TV}(y|t,V_2(t,S))\Big)\Big)\ f_{T|S}(t|S)\Big] + O_p(h_2^{r_2}) \\
&= E_S\Big[\Big(F_{Y|TS}(y|t,S)\Big(-\frac{W_V(V_2(t,S))}{f_{T|V}(t|V_2(t,S))}\nabla_v f_{T|V}(t|V_2(t,S))+\nabla_v W_V(V_2(t,S))\Big) \\
&\quad - W_V(V_2(t,S))\nabla_v F_{Y|TV}(y|t,V_2(t,S)) + W_V(V_2(S))F_{Y|TV}(y|t,V_2(t,S))\frac{\nabla_v f_{T|V}(t|V_2(t,S))}{f_{T|V}(t|V_2(t,S))} \\
&\quad - F_{Y|TV}(y|t,V_2(t,S))\nabla_v W_V(V_2(t,S))\Big)'\Big(V_1(t,S)-V_2(t,S)\Big)\frac{f_{T|S}(t|S)}{f_{T|V}(t|V_2(t,S))}\Big] \\
&\quad + O\big(\|V_1-V_2\|_\infty^2\big) + O_p(h_2^{r_2}).
\end{aligned}
$$

Together with $T_{2,n}$ in (A.17), the result is derived.

A.3.2 Proof of Corollary 1.1 and 1.2

Take $V(S) = F_{T|Z}(T|Z)$ for example. By the same linearization in (A.6), the dominating term is

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E\Big[A(y,T,Z)\Big(\mathbf{1}_{\{T_i\leq T\}} - F_{T|Z}(T|Z)\Big)K_b(Z_i - Z)\frac{1}{f_Z(Z)}\Big] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} E\Big[A(y,T,Z_i)\Big(\mathbf{1}_{\{T_i\leq T\}} - F_{T|Z}(T|Z_i)\Big)\frac{f_{Z|T}(Z_i|T)}{f_Z(Z_i)}\Big] + O_p(h_1^{r_1}).
\end{aligned}$$

If the generated regressor V is the GPS $f_{T|X}(t|X)$, then $S = X$. Then the regression $F_{Y|TV}$ has two regressors. For all $x \in \mathcal{X}$, $f_{T|V}(t|V(x)) = f_{T|X}(t|x)$ by Hirano and Imbens (2004). The third term of the influence function (**??**) is zero. Take derivative with respect to the argument x_l, $l \in \{1, 2, ..., d_x\}$, $\partial_{x_l} f_{T|X}(t|x) = \partial_v f_{T|V}(t|V(x)) \cdot \partial_{x_l} V(x)$ by chain rule. Since $\partial_{x_l} V(x) = \partial_{x_l} f_{T|X}(t|x)$, $\partial_v f_{T|V}(t|V(x)) = 1$. So for the second term (**??**), $A(y,X) = \nabla_v F_{Y|TV}(y|t,V)\Big(W - E[W|V]\Big) + \Big(F_{Y|TV}(y|t,V(X)) - F_{Y|TX}(y|t,X)\Big)\cdot\Big(E[W|V]/f_{T|X}(t|X) - \nabla_v E[W|V]\Big)$. Follow the same steps in the proof of Theorem 1.1 to derive the influence function for

$$\sqrt{nh_1}E\Big[A(y,X)\cdot\Big(\hat{f}_{T|X}(t|X) - f_{T|X}(t|X)\Big)\Big],$$

Linearize

$$\begin{aligned}
\frac{\hat{f}_{TX}(t,X)}{\hat{f}_X(X)} - f_{T|X}(t|X) &= \frac{1}{n}\sum_{i=1}^{n}\Big[K_{h_1}(T_i - t) - f_{T|X}(t|X)\Big]K_{h_1}(X_i - X)/\hat{f}_X(X) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[K_{h_1}(T_i - t) - f_{T|X}(t|X_i)\Big]K_{h_1}(X_i - X)/\hat{f}_X(X) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\Big[f_{T|X}(t|X_i) - f_{T|X}(t|X)\Big]K_{h_1}(X_i - X)/\hat{f}_X(X).
\end{aligned}$$

Focus on the first term and the second will be $s.o.$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big[K_{h_1}(T_i-t)-f_{T|X}(t|X_i)\Big]K_{h_1}(X_i-X)/\hat{f}_X(X)\\
=&\frac{1}{n}\sum_{i=1}^{n}\Big[K_{h_1}(T_i-t)-f_{T|X}(t|X_i)\Big]K_{h_1}(X_i-X)/f_X(X)\\
+&\frac{1}{n}\sum_{i=1}^{n}\Big[K_{h_1}(T_i-t)-f_{T|X}(t|X_i)\Big]K_{h_1}(X_i-X)\Big(\frac{1}{\hat{f}_X(X)}-\frac{1}{f_X(X)}\Big)
\end{aligned}$$

where the first term will dominate the influence function and the second term is of smaller order by the U-statistic theory. $\sqrt{nh_1}\|\hat{f}_X-f_X\|^2\to 0$.

$$\begin{aligned}
&\sqrt{nh_1}E\Big[A(y,X)\cdot\frac{1}{n}\sum_{i=1}^{n}\Big(K_{h_1}(T_i-t)-f_{T|X}(t|X_i)\Big)K_{h_1}(X_i-X)/f_X(X)\Big]\\
=&\sqrt{nh_1}\frac{1}{n}\sum_{i=1}^{n}\Big(K_{h_1}(T_i-t)-f_{T|X}(t|X_i)\Big)A(y,X_i)+\sqrt{nh_1}O({h_1}^{r_1}).
\end{aligned}$$

Assume $n{h_1}^{2r_1+1}\to 0$.

Therefore, if $h_2=O(h_1)$,

$$\begin{aligned}
&\sqrt{nh_2}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_{Y|T\hat{V}}(y|t,\hat{V}(X_i))W(X_i)-E[F_{Y|TV}(y|t,V(X))W(X)]\Big)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}^{GPS}(y)\\
&+\sqrt{\frac{h_2}{h_1}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(F_{Y|TV}(y|t,V(X_i))-F_{Y|TX}(y|t,X_i)\Big)\cdot\frac{W(X_i)}{f_{T|V}(t|V(X_i))}\frac{1}{\sqrt{h_1}}K\Big(\frac{T_i-t}{h_1}\Big)+o_p(1).
\end{aligned}$$

If $h_1=h_2$, then its influence function is reduced to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\mathbf{1}_{\{Y_i\le y\}}-F_{Y|TX}(y|t,X_i)\Big)\cdot\frac{W(X_i)}{f_{T|X}(t|X_i)}\frac{1}{\sqrt{h_2}}K\Big(\frac{T_i-t}{h_2}\Big)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}(y).$$

A.3.2.1 Proof of Lemma 1.2

First note that $E\Big[var\big(B(Y)\big|T=t,X\big)\cdot A(X)\Big]=E\Big[var\big(B(Y)\big|T=t,X\big)\cdot A(X)\cdot\frac{f_T(t)}{f_{T|X}(t|X)}\Big|T=t\Big]$. And $\frac{f_T(t)}{f_{T|X}(t|X)}$ is a function of $V(X)=f_{T|X}(t|X)$. So we could abuse the notation of $A(X)$ and prove $E\Big[var\big(B(Y)\big|T=t,V(X)\big)\cdot A(X)\Big|T=t\Big]\ge E\Big[var\big(B(Y)\big|T=t,X\big)\cdot A(X)\Big|T=t\Big]$.

By the law of iterated expectations,

$$
\begin{aligned}
&E\Big[var\big(B(Y)\big|T=t,V(X)\big)\cdot A(X)\Big|T=t\Big] \\
&=E\Big[E\big[\big(B(Y)-E[B(Y)|T=t,V(X)]\big)^2\big|T=t,X\big]\cdot A(X)\Big|T=t\Big].
\end{aligned}
$$

I could skip conditioning on $T=t$ for notational ease and observe that

$$
\begin{aligned}
&E\big[\big(B(Y)-E[B(Y)|V(X)]\big)^2\big|X\big]-E\big[\big(B(Y)-E[B(Y)|X]\big)^2\big|X\big] \\
&=E[B^2(Y)|X]-2E[B(Y)|X]\cdot E[B(Y)|V(X)]+\Big(E[B(Y)|V(X)]\Big)^2 \\
&-E[B^2(Y)|X]+\Big(E[B(Y)|X]\Big)^2 \\
&=\Big(E[B(Y)|V(X)]-E[B(Y)|X]\Big)^2\geq 0.
\end{aligned}
$$

The result is implied.

A.4 Estimation of the weights

Assume the weight function does not depend on the generated regressor V.

Lemma A.7. *Assume (i)* $\big\|\hat{W}-W\big\|_\infty=O_p(n^{-\zeta})$*, where* $\zeta>(d_2-d_t)\eta/2$*; (ii)* $W\in\mathcal{C}_M^\alpha(\mathcal{S})$*; (iii)* $\big\|D^q\hat{W}-D^qW\big\|_\infty=o_p(1)$*,* $q\leq\underline{\alpha}$*.*

1. *(Observable Regressors) Suppose all the regressors* X *are observable. Assume the conditions in Theorem 1.1. Then*

$$
\begin{aligned}
\sup_{y\in\mathcal{Y}}\Bigg|&\sqrt{nh^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^n\hat{F}_{Y|TX}(y|t,X_i)\hat{W}(S_i)-E\Big[F_{Y|TX}(y|t,X)W(S)\Big]\Big) \\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_{tin}^X(y|X,W)-\sqrt{nh^{d_t}}E\Big[F_{Y|TX}(y|t,X)\big(\hat{W}(S)-W(S)\big)\Big]\Bigg|=o_p(1).
\end{aligned}
$$

2. *(Generated Regressors) Suppose some regressors* V *are unobservable and estimated as generated regressors. Assume the conditions in Theorem 1.2, where the influence function for the case of the known weight is denoted as* $\psi_{tin}^{GR}(y)$*. Let* $\zeta>(1-d_t\eta)/2-$

$(\delta - \eta)$. *Then*

$$\sup_{y\in\mathcal{Y}}\Bigg|\sqrt{nh^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y|TX\hat{V}}(y|t,X_i,\hat{V}_i)\hat{W}(S_i) - E\Big[F_{Y|TXV}(y|t,X,V)W(S)\Big]\Big) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{tin}^{GR}(y) - \sqrt{nh^{d_t}}E\Big[F_{Y|TXV}(y|t,X,V)\big(\hat{W}(S)-W(S)\big)\Big]\Bigg| = o_p(1).$$

Condition (i) controls the remaining terms to be of smaller order. Condition (iii) ensures $\hat{W} \in \mathcal{C}^{\alpha}_{\mathcal{M}}$ with probability approaching one.

A.4.1 Proof of Lemma A.7

Denote $F_i \equiv F_{Y|TX}(y|t,X_i)$ and $W_i \equiv W(X_i)$. $\sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_i\hat{W}_i - E(F_iW_i)\Big) = \sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_i\hat{W}_i - \hat{F}_iW_i\Big) + \sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_i\, W_i - E(F_iW_i)\Big)$, where the second term is known from Theorem 1.1. The first term $\sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_i\hat{W}_i - \hat{F}_iW_i\Big) =$

$$\underbrace{\sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}F_i\Big(\hat{W}_i - W_i\Big)}_{(I)} + \underbrace{\sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_i - F_i\Big)\Big(\hat{W}_i - W_i\Big)}_{(II)}.$$

The first part uses the stochastic equicontinuity result in Lemma A.2

$$(I) = \sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}F_i\Big(\hat{W}_i - W_i\Big) = \sqrt{nh^{d_t}}E\Big[F_i\Big(\hat{W}_i - W_i\Big)\Big] + \bar{o}_p(1). \tag{A.21}$$

The second part is made of smaller order,

$$(II) = \sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_i - F_i\Big)\Big(\hat{W}_i - W_i\Big) = \sqrt{nh^{d_t}}E\Big(\hat{F}_i - F_i\Big)\Big(\hat{W}_i - W_i\Big) + \bar{o}_p(1). \tag{A.22}$$

By Cauchy-Schwarz inequality, $\Big(\frac{1}{n}\sum_{i=1}^{n}\big[(\hat{F}_i - F_i)(\hat{W}_i - W_i)\big]\Big)^2 \le \frac{1}{n}\sum_{i=1}^{n}\big(\hat{F}_i - F_i\big)^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\big(\hat{W}_i - W_i\big)^2 = O_p\Big((\log n)^2/(nh^{d_2})\Big)O_p(n^{-2\zeta}) = o_p((nh^{d_t})^{-1})$.

When the generated regressors are estimated $\hat{F}_i = \hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i)$, use the results and proofs in Proposition 3 and Theorem 1.2. By , $\hat{F}_i - F_i = \hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i) - F_{Y|TV}(y|t,V_i) = \hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i) - \hat{F}_{Y|TV}(y|t,\hat{V}_i) + \hat{F}_{Y|TV}(y|t,\hat{V}_i) - F_{Y|TV}(y|t,\hat{V}_i) + F_{Y|TV}(y|t,\hat{V}_i) - F_{Y|TV}(y|t,V_i)$. So $\Big|\frac{1}{n}\sum_{i=1}^{n}(\hat{F}_i - F_i)(\hat{W}_i - W_i)\Big| \le \Big|\frac{1}{n}\sum_{i=1}^{n}\big(\hat{F}_{Y|T\hat{V}}(y|t,\hat{V}_i) - \hat{F}_{Y|TV}(y|t,\hat{V}_i)\big)(\hat{W}_i - W_i)\Big| +$

$\Big|\frac{1}{n}\sum_{i=1}^n \big(\hat{F}_{Y|TV}(y|t,\hat{V}_i)-F_{Y|TV}(y|t,\hat{V}_i)\big)(\hat{W}_i-W_i)\Big|+\Big|\frac{1}{n}\sum_{i=1}^n \big(F_{Y|TV}(y|t,\hat{V}_i)-F_{Y|TV}(y|t,V_i)\big)(\hat{W}_i-W_i)\Big|$. The second term is the same as the case when the regressors are observable. The third term is of smaller order than the first term. For the first term, by (A.18), $\|\hat{F}_{Y|T\hat{V}}-F_{Y|TV}\| = O_p\Big(\frac{\|\hat{V}-V\|}{h}+\sqrt{\frac{\log n}{nh^{d_2}}}+R_n\Big)$. Assumption (vi) makes $\frac{\|\hat{V}-V\|^2}{h^2}\cdot n^{-2\zeta} = o((nh^{d_t})^{-1})$.

A.4.2 Proof of Theorem 1.3

I exchange t and $\bar{t}$ in the proof. I also skip the subscript for S_w for notational simplicity. By Lemma A.7, the influence from estimating the weight estimation is dominated by

$$\sqrt{nb^{d_t}}E\big[F_{y|\bar{t}\Lambda}\big(\hat{W}(S)-W(S)\big)\big] = \sqrt{nb^{d_t}}E\Big[F_{y|\bar{t}\Lambda}(\hat{f}_{t|S}-f_{t|S})\Big]\Big[\frac{1}{f_t}+\Big(\frac{1}{\hat{f}_t}-\frac{1}{f_t}\Big)\Big] \tag{A.23}$$

$$+\sqrt{nb^{d_t}}E\Big[F_{y|\bar{t}\Lambda}f_{t|S}\Big]\Big(\frac{1}{\hat{f}_t}-\frac{1}{f_t}\Big). \tag{A.24}$$

Consider (A.23) first,

$$\hat{f}_{t|S}-f_{t|S} = \frac{\hat{f}_{tS}}{\hat{f}_S}-\frac{f_{tS}}{f_S} = \frac{\hat{f}_{tS}-f_{tS}}{f_S}+\frac{f_{t|S}}{f_S}(f_S-\hat{f}_S)+\frac{1}{f_S}(f_S-\hat{f}_S)(\hat{f}_{t|S}-f_{t|S}).$$

The contribution of the third term is made to be of smaller order: $\sqrt{nb^{d_t}}\,O_p\Big(\|f_S-\hat{f}_S\|_\infty\|\hat{f}_{t|S}-f_{t|S}\|_\infty\Big) = O_p\Big(\sqrt{nb^{d_t}}\frac{\log n}{\sqrt{nb^{d_w}}}\frac{\log n}{\sqrt{nb^{d_w+d_t}}}\Big) = o_p(1)$ by assuming $nb^{2d_w}/(\log n)^2\to\infty$. The second term contributes to a full-mean,

$$-\sqrt{nb^{d_t}}E\big[F_{Y|\bar{t}\Lambda}\frac{f_{t|S}}{f_S}\big(\hat{f}_S-f_S\big)\big] = -\sqrt{nb^{d_t}}\int E[F_{y|\bar{t},\Lambda}|S]\Big(\frac{1}{n}\sum_{i=1}^n K_b(S_i-S)-f_S(S)\Big)f_{t|S}dS$$

$$= -\sqrt{nb^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^n E[F_{y|\bar{t},\Lambda}|S_i]f_{t|S_i}+\bar{O}_p(b^r)-E\Big[E[F_{y|\bar{t},\Lambda}|S]f_{t|S}\Big]\Big) = \bar{O}_p(\sqrt{b}) = \bar{o}_p(1),$$

by the Donsker property of $F_{Y|T\Lambda}(y|\bar{t},\Lambda)$.

The first term is

$$\sqrt{nb^{d_t}}E\big[E[F_{y|\bar{t}\Lambda}|S]\big(\frac{1}{f_S}(\hat{f}_{tS}-f_{tS})\big)\big] = \int E[F_{y|\bar{t},\Lambda}|S]\cdot\sqrt{nb^{d_t}}\Big(\frac{1}{n}\sum_{i=1}^n K_b(T_i-t)K_b(S_i-S)-f_{tS}\Big)dS$$

$$= \sqrt{nb^{d_t}}\Big(\frac{1}{n}\sum_i K_b(T_i-t)\big(E[F_{y|\bar{t},\Lambda}|S_i]+\bar{O}_p(b^r)\big)-\theta_{\bar{t}}f_t\Big)$$

$$= \frac{1}{\sqrt{n}}\sum_i\big(\psi^t_{\bar{t}1i}(y)-E\psi^t_{\bar{t}1i}(y)\big)f_t+\bar{O}_p(\sqrt{nb^{d_t+2r}})$$

where $\psi^t_{\bar{t}1i}(y) = \sqrt{b^{d_t}}K_b\big(T_i - t\big)E\big[F_{Y|T\Lambda}(y|\bar{t},\Lambda)\big|S = S_i\big]/f_T(t)$ and $E\psi^t_{\bar{t}1i}(y) = \sqrt{b^{d_t}}\theta_{\bar{t}} + \bar{O}(b^{r+d_t/2})$. Therefore, (A.23) is

$$\sqrt{nb^{d_t}}E\Big[F_{y|\bar{t}\Lambda}(\hat{f}_{t|S} - f_{t|S})\Big]\Big[\frac{1}{f_t} + \Big(\frac{1}{\hat{f}_t} - \frac{1}{f_t}\Big)\Big] = \frac{1}{\sqrt{n}}\sum_i \psi^t_{\bar{t}1i}(y) + \bar{O}_p(\sqrt{nb^{d_t+2r}}) + \bar{O}_p(\|\hat{f}_t - f_t\|_\infty).$$

(A.24) is

$$\begin{aligned}&-\sqrt{nb^{d_t}}E\big[F_{y|\bar{t}\Lambda}\frac{f_{t|S}}{f_t^2}\big]\big(\hat{f}_t - f_t\big) + \sqrt{nb^{d_t}}\ \bar{O}_p\big(\|\hat{f}_t - f_t\|_\infty^2\big)\\ &= -\sqrt{nb^{d_t}}\big(\hat{f}_t - f_t\big)\cdot\theta_{\bar{t}}/f_T(t) + \bar{o}_p(1) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(\psi^t_{\bar{t}2i} - E\psi^t_{\bar{t}2i}\Big) + O_p(\sqrt{nb^{d_t+2r}}),\end{aligned}$$

where $\psi^t_{\bar{t}2i} = \sqrt{b^{d_t}}K_b\big(T_i - t\big)\theta_{\bar{t}}/f_T(t)$ and $E\psi^t_{\bar{t}2i} = \sqrt{b^{d_t}}\theta_{\bar{t}} + \bar{O}(b^{r+d_t/2}) = \bar{O}(b^r)$.

Therefore,

$$\sqrt{nb^{d_t}}\frac{1}{n}\sum_{i=1}^n F_i\Big(\hat{W}_i - W_i\Big) = \frac{1}{\sqrt{n}}\sum_i\big(\psi^t_{\bar{t}1i}(y) - \psi^t_{\bar{t}2i}(y)\big) + \bar{o}_p(1).$$

A.4.2.1 FCLT

The weak convergence is proved by checking the conditions for FCLT in Pollard (1990) in the following subsection. I first show the weak convergence of

$$\sqrt{nb^{d_t}}\frac{1}{n}\sum_{i=1}^n F_{Y|T\Lambda}(\cdot|\bar{t},\Lambda_i)\Big(\hat{W}(\Lambda_i) - W(\Lambda_i)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\psi^t_{\bar{t}i}(\cdot) + \bar{o}_p(1) \Rightarrow \mathbb{G}^w(\cdot).$$

(i) $\{E[F_{Y|T\Lambda}(y|\bar{t},\Lambda)|S_w]\}$ and $\{F_{Y(\bar{t})|T}(y|t)\}$ are monotone increasing in y, so $\{\psi^t_{\bar{t}}/n\}$ are manageable with envelopes $F_{ni} = \frac{\sqrt{b^{d_t}}}{\sqrt{n}}K_b\big(T_i - t\big)\frac{1}{f_t}$.

$$\begin{aligned}E[\psi^t_{\bar{t}i}] &= \sqrt{b^{d_t}}E\Big[\int K_b\big(T - t\big)f_{T|S_w}(T|S_w)dT\frac{1}{f_t}\Big(E[F_{Y|T\Lambda}(y|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y|t)\Big)\Big]\\ &= \sqrt{b^{d_t}}E\Big[f_{T|S_w}(t|S_w)\frac{1}{f_t}\Big(E[F_{Y|T\Lambda}(y|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y|t)\Big)\Big] + O_p(b^{r+d_t/2}).\end{aligned}$$

$$
\begin{aligned}
&E[\psi^t_{\bar{t}i}(y_1)\psi^t_{\bar{t}i}(y_2)] \\
&= E\Big[\frac{b^{d_t}K_b^2\big(T-t\big)}{f_t^2}\Big(E[F_{Y|T\Lambda}(y_1|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y_1|t)\Big)\Big(E[F_{Y|T\Lambda}(y_2|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y_2|t)\Big)\Big] \\
&= \int K^2(v)dvE\Big[\frac{f_{t|S_w}}{f_t^2}\Big(E[F_{Y|T\Lambda}(y_1|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y_1|t)\Big)\Big(E[F_{Y|T\Lambda}(y_2|\bar{t},\Lambda)|S_w] \\
&- F_{Y(\bar{t})|T}(y_2|t)\Big)\Big] + O_p(b)
\end{aligned}
$$

uniformly in y_1 and y_2.

(ii) Define $\mathcal{Z}_n(y) = \sum_{i=1}^n \Big(f_{ni}(y) - Ef_{ni}(y)\Big)$. Let $y_1 \le y_2 \in \mathcal{Y}$. The covariance kernel of the limiting Gaussian process is

$$
\begin{aligned}
&\lim_{b\to 0} P\mathcal{Z}_n(y_1)\mathcal{Z}_n(y_2) = \lim_{b\to 0} E\Big[\psi^t_{\bar{t}i}(y_1)\psi^t_{\bar{t}i}(y_2)\Big] - E\Big[\psi^t_{\bar{t}i}(y_1)\Big]E\Big[\psi^t_{\bar{t}i}(y_2)\Big] \\
&= \lim_{b\to 0} E\Big[\psi^t_{\bar{t}i}(y_1)\psi^t_{\bar{t}i}(y_2)\Big] \equiv H^t(y_1,y_2) \\
&= \int K^2(v)dvE\Big[\frac{f_{t|S_w}}{f_t^2}\Big(E[F_{Y|T\Lambda}(y_1|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y_1|t)\Big)\Big(E[F_{Y|T\Lambda}(y_2|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y_2|t)\Big)\Big].
\end{aligned}
$$

(iii)

$$
\sum_{i=1}^n PF_{ni}^2 = \frac{1}{f_t}\int K^2(v)dv.
$$

(iv) For any $\epsilon > 0$, $\sum_{i=1}^n PF_{ni}^2\ \mathbf{1}\big\{F_{ni} > \epsilon\big\} \to 0$ holds. Because K is bounded, f_T is bounded away from zero, and $\sqrt{nb^{d_t}} \to \infty$, $\mathbf{1}\big\{\frac{\sqrt{b^{d_t}}}{\sqrt{n}}K_b\big(T-t\big)\frac{1}{f_t} > \epsilon\big\} = 0$ for n large enough.

(v) Denote $F_{Y|T,\Lambda}(y|t,\Lambda) \equiv F_{y|t,\Lambda}$. $P\Big|f_{ni}(y_1) - f_{ni}(y_2)\Big|^2 =$

$$
\begin{aligned}
&\frac{1}{n}\int K^2(v)dv\frac{1}{f_t^2}E\Big[f_{t|\Lambda}\Big(E[F_{Y|T\Lambda}(y_1|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y_1|t) \\
&- E[F_{Y|T\Lambda}(y_2|\bar{t},\Lambda)|S_w]] + F_{Y(\bar{t})|T}(y_2|t)\Big)^2\Big] + O_p(b/n)
\end{aligned}
$$

uniform in y_1, y_2. Therefore, uniformly in y_1, y_2, $\rho_n(y_1,y_2) \to \rho(y_1,y_2) \equiv$

$$
\begin{aligned}
&\Big\{\int K^2(v)dv\frac{1}{f_t^2}E\Big[f_{t|\Lambda}\Big(E[F_{Y|T\Lambda}(y_1|\bar{t},\Lambda)|S_w] - F_{Y(\bar{t})|T}(y_1|t) \\
&- E[F_{Y|T\Lambda}(y_2|\bar{t},\Lambda)|S_w] + F_{Y(\bar{t})|T}(y_2|t)\Big)^2\Big]\Big\}^{1/2}.
\end{aligned}
$$

Next, I consider the weak convergence of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\psi_{\bar{t}i}(\cdot)+\psi^t_{\bar{t}i}(\cdot)\big)$. $E[\psi_{\bar{t}i}(y_1)\psi^t_{\bar{t}i}(y_2)]=$

$$E\Big[\big(\mathbf{1}_{\{Y\le y_1\}}-F_{y_1|\bar{t}\Lambda}\big)b^{d_t}K_b\Big(T-\bar{t}\Big)K_b\Big(T-t\Big)\frac{f_{t|\Lambda}}{f_{\bar{t}|\Lambda}f_t^2}\Big(E[F_{y_2|\bar{t}\Lambda}|S_w]-F_{Y(\bar{t})|T}(y_2|t)\Big)\Big]$$

where

$$\int\big(F_{y_1|T\Lambda}(y|\bar{t}+vb)-F_{y_1|\bar{t}\Lambda}\big)K(v)K\Big(\frac{\bar{t}-t}{b}+v\Big)\cdot f_{T|\Lambda}(\bar{t}+vb|\Lambda)dv=\bar{O}_p(b).$$

So $E[\psi_{\bar{t}i}(y_1)\psi^t_{\bar{t}i}(y_2)]=\bar{o}(1)$. Then for $\psi_{\bar{t}|t}\equiv\psi_{\bar{t}}+\psi^t_{\bar{t}}$, $E[\psi_{\bar{t}|t}(y_1)\psi_{\bar{t}|t}(y_2)]=E[\psi_{\bar{t}}(y_1)\psi_{\bar{t}}(y_2)]+E[\psi^t_{\bar{t}}(y_1)\psi^t_{\bar{t}}(y_2)]+\bar{o}(1)$.

(ii)

$$\begin{aligned}\lim_{h\to0}P\mathcal{Z}_n(y_1)\mathcal{Z}_n(y_2)&=\lim_{h\to0}E\Big[\psi_{\bar{t}|t}(y_1)\psi_{\bar{t}|t}(y_2)\Big]-E\Big[\psi_{\bar{t}|t}(y_1)\Big]E\Big[\psi_{\bar{t}|t}(y_2)\Big]\\&=\lim_{h\to0}E[\psi_{\bar{t}}(y_1)\psi_{\bar{t}}(y_2)]+E[\psi^t_{\bar{t}}(y_1)\psi^t_{\bar{t}}(y_2)]=H(y_1,y_2)+H^t(y_1,y_2).\end{aligned}$$

(iii) $E[(F_{ni}+F^t_{ni})^2]=E[F^2_{ni}]+E[{F^t_{ni}}^2]+2E[F_{ni}F^t_{ni}]$.

$$\begin{aligned}E[F_{ni}F^t_{ni}]&=\frac{1}{n}E\Big[b^{d_t}K_b\Big(T-\bar{t}\Big)K_b\Big(T-t\Big)\frac{f_{t|\Lambda}}{f_{\bar{t}|\Lambda}f_t^2}\Big]\\&=\frac{1}{n}E\Big[\Big(\bar{K}\Big(\frac{t-\bar{t}}{b}\Big)f_{\bar{t}|\Lambda}+O(b)\Big)\frac{f_{t|\Lambda}}{f_{\bar{t}|\Lambda}f_t^2}\Big]=\frac{1}{n}\bar{K}\Big(\frac{t-\bar{t}}{b}\Big)\frac{1}{f_t}+O(b/n),\end{aligned}$$

where the convolution kernel is defined in (A.13). So $\sum_{i=1}^{n}PF^2_{ni}$ converges.

(iv) $\sum_{i=1}^{n}P(F_{ni}+F^t_{ni})^2\mathbf{1}\big\{F_{ni}+F^t_{ni})>\epsilon\big\}\to0$ by the same argument as (A.25).

(v)

$$\begin{aligned}&P\Big[f_{ni}(y_1)+f^t_{ni}(y_1)-f_{ni}(y_2)-f^t_{ni}(y_2)\Big]^2\\&=P\Big[f_{ni}(y_1)-f_{ni}(y_2)\Big]^2+P\Big[f^t_{ni}(y_1)-f^t_{ni}(y_2)\Big]^2+2P\Big[f_{ni}(y_1)-f_{ni}(y_2)\Big]\Big[f^t_{ni}(y_1)-f^t_{ni}(y_2)\Big],\end{aligned}$$

where the last term is $\bar{o}(1)$. The uniform convergence of the first two terms are shown in the previous proofs.

A.5 Inference for the Treatment Effects

A.5.1 Proof for Theorem 1.4

By the functional delta method (e.g., Theorem 3.9.4 in van der Vaart and Wellner (1996)) and the linearity of the Hadamard derivative, the weak convergence to a Gaussian process is implied.

A.5.2 Proof of Corollary 1.3

Using the results in (A.15), for the diagonal term $t \neq s$, $E\big[\psi_{tin}(y_1)\psi_{si}(y_2)\big] = 0$ as $h \to 0$.

A.5.3 Mean

Note that $\int_{\mathcal{Y}} y d\mathbf{1}_{\{Y \leq y\}} = Y$ using integration by parts. Therefore, $\Gamma'\big(\mathbf{1}_{\{Y \leq y\}} - F_{Y|TX}(y|t, X)\big)$ $= Y - E(Y|t, X)$. $V_u = \lim_{h \to 0} E[\psi_i^2]$.

A.5.4 Quantile processes

The Hadamard derivative is shown in Example 3.9.24 in van der Vaart and Wellner (1996).

A.5.5 Multiplier Method: Proof of Theorem 1.5

Decompose

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \hat{\psi}_{tin}(y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \psi_{tin}(y) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \big(\hat{\psi}_{tin}(y) - \psi_{tin}(y)\big)$$

First, I use the functional CLT, Theorem 10.6 in Pollard (1990) to show $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \psi_{tin}(\cdot)$ $\Rightarrow \mathbb{G}_t(\cdot)$. Then I show the rest terms are $\bar{o}_p(1)$.

Following the notation defined in the proof of Lemma A.6, define $f_{ni}^u(y) = U_i f_{ni}(y) = U_i \frac{1}{\sqrt{n}} \psi_{tin}(y)$ which has the envelope $F_{ni}^u = U_i F_{ni} = U_i \sqrt{\frac{h^{d_t}}{n}} K_h\big(T_i - t\big) W_\Lambda(\Lambda_i) / f_{t|\Lambda_i}$. Then (i) holds. $E f_{ni}^u(y) = 0$ and $\mathcal{Z}_n^u(y) = \sum_{i=1}^{n} f_{ni}^u(y)$.

(ii) Let $y_1 \le y_2$.

$$P(\mathcal{Z}_n^u(y_1)\mathcal{Z}_n^u(y_2)) = E\big[\sum_{i=1}^n f_{ni}^u(y_1) f_{ni}^u(y_2)\big] = \frac{1}{n}\sum_{i=1}^n \psi_{tin}(y_1)\psi_{tin}(y_2)\cdot EU_i^2$$

$$= \frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(T_i - t\big)\big(\mathbf{1}_{\{Y_i \le y_1\}} - F_{y_1|t\Lambda_i}\big)\big(\mathbf{1}_{\{Y_i \le y_2\}} - F_{y_2|t\Lambda_i}\big) W_\Lambda^2(\Lambda_i)/f_{t|\Lambda_i}^2 \xrightarrow{p} H(y_1, y_2)$$

defined in (A.16), by the weak law of large number.

(iii)

$$\sum_{i=1}^n PF_{ni}^{u\,2} = \sum_{i=1}^n \frac{h^{d_t}}{n} K_h^2\big(T_i - t\big) W_\Lambda(\Lambda_i)^2/f_{t|\Lambda_i}^2 \to \int K^2(u)du \cdot E[W_\Lambda(\Lambda)^2 f_{t|\Lambda}^{-1}].$$

(iv) $B \equiv \inf_{\{\Lambda_i, T_i\}} f_{t|\Lambda_i}/\Big(W_\Lambda(\Lambda_i)^2 h^{d_t} K_h\big(T_i - t\big)\Big)$ exists, because $f_{T|\Lambda}$ is bounded away from zero, the weight and the kernel are uniformly bounded. For any $\epsilon > 0$, $\sum_{i=1}^n PF_{ni}^{u\,2}\,\mathbf{1}\big\{F_{ni} > \epsilon\big\} =$

$$\frac{h^{d_t}}{n}\sum_{i=1}^n K_h^2\Big(T_i - t\Big)\frac{W_\Lambda(\Lambda_i)^2}{f_{t|\Lambda_i}^2}\cdot E\Big[U_i^2\;\mathbf{1}\Big\{U_i > \frac{\sqrt{n}\epsilon f_{t|\Lambda_i}}{\sqrt{h^{d_t}} K_h\big(T_i - t\big) W_\Lambda(\Lambda_i)}\Big\}\Big]$$

$$\le \frac{h^{d_t}}{n}\sum_{i=1}^n K_h^2\Big(T_i - t\Big)\frac{W_\Lambda(\Lambda_i)^2}{f_{t|\Lambda_i}^2}\cdot E\Big[U_i^2\;\mathbf{1}\Big\{U_i > \sqrt{nh^{d_t}}\epsilon B\Big\}\Big]$$

$$\to \int K^2(u)du \cdot E\big[W_\Lambda(\Lambda_i)^2 f_{t|\Lambda}^{-1}\big]\cdot 0 = 0 \tag{A.25}$$

because $\sqrt{nh^{d_t}} \to \infty$.

(v) Denote $F_{Y|T\Lambda}(y|t,\Lambda_i) = F_y$ and $\mathbf{1}_{\{Y_i \le y\}} = \mathbf{1}_y$. Then for any $y_1 \le y_2$, $\rho_n^u(y_1,y_2)^2$

$$= \sum_{i=1}^n E\Big(f_{ni}^u(y_1) - f_{ni}^u(y_2)\Big)^2 = \frac{1}{n}\sum_{i=1}^n \Big(\psi_{tin}^2(y_1) + \psi_{tin}^2(y_2) - 2\psi_{tin}(y_1)\psi_{tin}(y_2)\Big)$$

$$= \frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(T_i - t\big) W_\Lambda(\Lambda_i)^2 \Big[\mathbf{1}_{y_1} - 2\mathbf{1}_{y_1}F_{y_1} + F_{y_1}^2 + \mathbf{1}_{y_2} - 2\mathbf{1}_{y_2}F_{y_2} + F_{y_2}^2 \tag{A.26}$$

$$- 2\mathbf{1}_{y_1} + 2\mathbf{1}_{y_1}F_{y_2} + 2\mathbf{1}_{y_2}F_{y_1} - 2F_{y_1}F_{y_2}\Big] \Big/ f_{t|\Lambda_i}^2 \tag{A.27}$$

$$= \frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(T_i - t\big) W_\Lambda(\Lambda_i)^2 \Big[\mathbf{1}_{y_1}\big(-1 - 2F_{y_1} + 2F_{y_2}\big) + \mathbf{1}_{y_2}\big(1 - 2F_{y_2} + 2F_{y_1}\big)$$

$$+ \big(F_{y_1} - F_{y_2}\big)^2\Big] \Big/ f_{t|\Lambda_i}^2$$

$$\to \int K^2(u)du \cdot E\Big[\frac{W_\Lambda(\Lambda_i)^2}{f_{t|\Lambda}}\Big(F_{y_1}(-1 - 2F_{y_1} + 2F_{y_2}) + F_{y_2}(1 - 2F_{y_2} + 2F_{y_1}) + (F_{y_1} - F_{y_2})^2\Big)\Big]$$

$$= \int K^2(u)du \cdot E\Big[\frac{W_\Lambda(\Lambda)^2}{f_{t|\Lambda}}(F_{y_2} - F_{y_1})(1 - F_{y_2} + F_{y_1})\Big] \equiv \rho^u(y_1,y_2)^2.$$

It remains to show that for all deterministic sequences $\{y_{1n}\}$ and $\{y_{2n}\}$ such that $\rho^u(y_{1n},y_{2n}) \to 0$, $\rho_n^u(y_{1n},y_{2n}) \to 0$. Using the same argument in Lemma (A.6) FCLT, $\sqrt{nh^{d_t}}\Big\{\frac{1}{n}\sum_{i=1}^n \psi_{tin}^2(y_1) - \int K^2(u)du \cdot E\Big[\frac{W_\Lambda(\Lambda)^2}{f_{t|\Lambda}}\big(F_{y_1} - F_{y_1}^2\big)\Big]\Big\}$ converges to a Gaussian process of y_1. So the first part $\frac{1}{n}\sum_{i=1}^n \big[\psi_{tin}^2(y_1) + \psi_{tin}^2(y_2)\big]$ (A.26) in $\rho_n^u(y_1,y_2)$ converges uniformly in y_1, y_2.

For the second part $-\frac{2}{n}\sum_{i=1}^n \psi_{tin}(y_1)\psi_{tin}(y_2)$ (A.27) indexed by both y_1 and y_2, I focus on one of the terms, defining $A_n(y_1,y_2) \equiv \frac{1}{n}\sum_{i=1}^n h^{d_t}K_h^2\big(T_i - t\big)W_\Lambda(\Lambda)^2(\Lambda_i)\mathbf{1}_{y_1}F_{y_2}/f_{t|\Lambda_i}^2$ and $A(y_1,y_2) \equiv \int K^2(u)du \cdot E\Big[\frac{W_\Lambda(\Lambda)^2}{f_{t|\Lambda}}F_{y_1}F_{y_2}\Big]$. It suffices to show that for all deterministic sequences $\{y_{1n}\}$ and $\{y_{2n}\}$ such that $A(y_{1n},y_{2n}) \to 0$, $A_n(y_{1n},y_{2n}) \to 0$.

By assumption, $W_\Lambda(\Lambda)^2(\Lambda)/f_{t|\Lambda} < \delta < \infty$. $A(y_{1n},y_{2n}) \to 0$ means that for any $\epsilon > 0$, there exists an integer N_0 such that for $n > N_0$,

$$A(y_{1n},y_{2n}) \le \int K^2(u)du \cdot \delta \cdot E\big[F_{Y|T\Lambda}(y_{1n}|t,\Lambda)F_{Y|T\Lambda}(y_{2n}|t,\Lambda)\big]$$

$$\le C \cdot E\big[F_{Y|T\Lambda}(\min\{y_{1n},y_{2n}\}|t,\Lambda)\big] < \epsilon, \tag{A.28}$$

defining $C = \int K^2(u)du \cdot \delta$ for notational ease. Actually, $\min\{y_{1n},y_{2n}\}$ can be either y_{1n} or y_{2n}. It's not required both the deterministic sequence to converge.

Since $F_{Y|T\Lambda}$ is increasing in y, there exists y_0 such that $C \cdot E[F_{Y|T\Lambda}(y_0|t,\Lambda)] = \epsilon$. Then for any $y < y_0$, $C \cdot E[F_{Y|T\Lambda}(y|t,\Lambda)] \leq \epsilon$. Then (A.28) implies either $y_{1n} < y_0$ or $y_{1n} < y_0$ or both for $n > N_0$.

First, note that $\frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(T_i - t\big)\mathbf{1}_{\{Y_i \leq y_0\}} \Big/ f_{t|\Lambda_i} \to \int K^2(u)du E\big[F_{y_0|t\Lambda}\big]$, i.e., for any $\epsilon_1 > 0$, there exists an integer N_1 such that

$$\Big|\frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(\frac{T_i - t}{h}\big)\mathbf{1}_{\{Y_i \leq y_0\}} \Big/ f_{t|\Lambda_i} - \int K^2(u)du E\big[F_{y_0|t\Lambda}\big]\Big| < \epsilon_1,$$

for $n > N_1$. For the case $y_{1n} < y_0$, for $n > \max\{N_1, N_0\}$,

$$\begin{aligned} A_n(y_{1n}, y_{2n}) &\leq \delta \frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(T_i - t\big)\mathbf{1}_{y_{1n}} F_{y_{2n}} \Big/ f_{t|\Lambda_i} \leq \delta \frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(T_i - t\big)\mathbf{1}_{\{Y_i \leq y_0\}} \Big/ f_{t|\Lambda_i} \\ &\leq \delta \int K^2(u)du E\big[F_{y_0|t\Lambda}\big] + \delta\epsilon_1 = \epsilon + \delta\epsilon_1. \end{aligned}$$

For the other case $y_{2n} < y_0$, use the similar argument by $\frac{1}{n}\sum_{i=1}^n h^{d_t} K_h^2\big(T_i - t\big) F_{y_0} \Big/ f_{t|\Lambda_i} \to \int K^2(u)du E\big[F_{y_0|t\Lambda}\big]$. Then it's shown $A_n(y_{1n}, y_{2n}) \to 0$. The same argument applies to other terms in (A.27).

Therefore, by the FCLT in Pollard (1990), $\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i \psi_{tin}(\cdot) \Rightarrow \mathbb{G}_t(\cdot)$.

Next, I need to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\Big(\hat{\psi}_{tin}(y) - \psi_{tin}(y)\Big) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i \sqrt{h^{d_t}} K_h\big(T_i - t\big)\Big(\hat{\varphi}_{tin}(y) - \varphi_{tin}(y)\Big) = \bar{o}_p(1)$$

where $\varphi_{tin}(y) = \Big(\mathbf{1}_{\{Y_i \leq y\}} - F_{Y|T\Lambda}(y|t,\Lambda_i)\Big)E[W(S_w)|\Lambda = \Lambda_i]/f_{T|\Lambda}(t|\Lambda_i)$ and a consistent estimator $\hat{\varphi}_{tin}(y) = \Big(\mathbf{1}_{\{Y_i \leq y\}} - \hat{F}_{Y|T\Lambda}(y|t,\Lambda_i)\Big)\hat{E}[W(S_w)|\Lambda = \Lambda_i]/\hat{f}_{T|\Lambda}(t|\Lambda_i)$. I show this empirical process converges to Gaussian processes with zero covariance kernel, by checking the conditions of FCLT in Pollard (1990).

(i) Given the sample, $\hat{F}_{y|t\Lambda_i}$ is monotone increasing in y by construction, so $\Big\{f_{ni}(y) \equiv U_i\sqrt{\frac{h^{d_t}}{n}} K_h\big(T_i - t\big) \cdot \Big(\hat{\varphi}_{tin}(y) - \varphi_{tin}(y)\Big)\Big\}$ are manageable.

$Ef_{ni}(y) = 0$. $Ef_{ni}^2(y) = \frac{h^{d_t}}{n} K_h^2\big(T_i - t\big) \cdot \big(\hat{\varphi}_{tin}(y) - \varphi_{tin}(y)\big)^2$. $\mathcal{Z}_n(y) = \sum_{i=1}^n f_{ni}(y)$. Assuming $f_{T|\Lambda}(t|\Lambda)$ and $W(S_w)$ are uniformly bounded away from zero and above, respectively, define the envelope $F_{ni} = U_i\sqrt{\frac{h^{d_t}}{n}} K_h\big(T_i - t\big)C$.

(ii)

$$\begin{aligned}\big|P\mathcal{Z}_n(y_1)\mathcal{Z}_n(y_2)\big| &= \big|P\sum_{i=1}^n f_{ni}(y_1)f_{ni}(y_2)\big| \\ &\le \sum_{i=1}^n \frac{h^{d_t}}{n}K_h^2\big(T_i-t\big)\Big|\hat{\varphi}_{tin}(y_1)-\varphi_{tin}(y_1)\Big|\Big|\hat{\varphi}_{tin}(y_2)-\varphi_{tin}(y_2)\Big| \\ &\le \sum_{i=1}^n \frac{h^{d_t}}{n}K_h^2\big(T_i-t\big)\cdot\Big\|\hat{\varphi}_{tin}-\varphi_{tin}\Big\|_\infty^2 = O_p(1)\cdot o_p(1) = o_p(1).\end{aligned}$$

(iii)

$$\sum_{i=1}^n PF_{ni}^2 = \sum_{i=1}^n \frac{h^{d_t}}{n}K_h^2\big(T_i-t\big)C^2 \to \int K^2(u)du\cdot f_T(t)C^2.$$

(iv) $\sum_{i=1}^n PF_{ni}^2\mathbf{1}\big\{U_i\sqrt{\frac{h^{d_t}}{n}}K_h\big(T_i-t\big)C>\epsilon\big\}\to 0$ by the same argument as (A.25).

(v)

$$\begin{aligned}0 &\le \sum_{i=1}^n E\Big[f_{ni}(y_1)-f_{ni}(y_2)\Big]^2 \\ &= \frac{1}{n}\sum_{i=1}^n h^{d_t}K_h^2\big(T_i-t\big)\Big(\hat{\varphi}_{tin}(y_1)-\varphi_{tin}(y_1)-\big(\hat{\varphi}_{tin}(y_2)-\varphi_{tin}(y_2)\big)\Big)^2 \\ &\le \frac{1}{n}\sum_{i=1}^n h^{d_t}K_h^2\big(T_i-t\big)\Big\|\hat{\varphi}_{tin}(y_1)-\varphi_{tin}(y_1)-\big(\hat{\varphi}_{tin}(y_2)-\varphi_{tin}(y_2)\big)\Big\|_\infty^2 \to 0.\end{aligned}$$

A.6 Proofs of Lemmas for Stochastic Equicontinuity

In this section, I drop the subscript of S_v and denote it by S for notational simplicity.

A.6.1 Proof of Lemma A.1

Define $Z_{ni}(\upsilon)\equiv\frac{1}{\sqrt{n}}f(y,t,\Lambda_i)W(X_i)$, indexed by $\upsilon\equiv(y,f)\in\Upsilon\equiv\mathcal{Y}\times\mathcal{F}$. The bracketing CLT will imply $\sum_{i=1}^n\Big(Z_{ni}(\upsilon)-EZ_{ni}(\upsilon)\Big)$ is asymptotic stochastic equicontinuous in υ with respect to the pseudo-metric $\rho(\upsilon_1,\upsilon_2)=\max\{|y_1-y_2|,\|f_1-f_2\|_\infty\}$. The conditions for the bracketing CLT from Theorem 2.11.9 in van der Vaart and Wellner (1996) are checked in the following:

(i) Since the functions are assumed to be uniformly bounded above and below, $\mathbf{1}_{\{\|Z_{ni}\|_\Upsilon > \eta\}} = 0$ for n large enough. So for any $\eta > 0$, $\sum_{i=1}^n E\Big[\|Z_{ni}\|_\Upsilon \mathbf{1}_{\{\|Z_{ni}\|_\Upsilon > \eta\}}\Big] = o_p(1)$.

(ii) It is straightforward to modify Lemma B.2 in **?** to replace their Lipschitz continuity with Hölder continuity,

$$N(\epsilon_1^{1/2} C_L + \epsilon_2, \mathcal{F}, \|\cdot\|_\infty) \leq N\Big(\epsilon_1, \mathcal{Y}, |\cdot|\Big) \times \sup_{y \in \mathcal{Y}} N\Big(\epsilon_2, \mathcal{M}, \|\cdot\|_\infty\Big).$$

Since $\mathcal{Y}$ is a compact set, the result remains.

$$N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq N\Big(\big(\epsilon/(2C_L)\big)^2, \mathcal{Y}, |\cdot|\Big) \times \sup_{y \in \mathcal{Y}} N\Big(\epsilon/2, \mathcal{M}, \|\cdot\|_\infty\Big).$$

$$N_{[\,]}(\epsilon, \Upsilon, L_2) \leq N\Big(\frac{\epsilon}{2C}, \mathcal{Y}, |\cdot|\Big) \times N\Big(\frac{\epsilon}{2C}, \mathcal{F}, \|\cdot\|_\infty\Big).$$

Therefore, $\int_0^{\delta_n} \sqrt{\log N_{[\,]}(\epsilon, \Gamma, L_2)} d\epsilon \to 0$, $\forall \delta_n \to 0$.

(iii)

$$\sum_{i=1}^n E\big(Z_{ni}(\upsilon_1) - Z_{ni}(\upsilon_2)\big)^2 = E\Big(f_1(y_1, t, \Lambda) W(X) - f_2(y_2, t, \Lambda) W(X)\Big)^2$$
$$= E\Big[\Big(f_1(y_1, t, \Lambda) - f_2(y_1, t, \Lambda) + f_2(y_1, t, \Lambda) - f_2(y_2, t, \Lambda)\Big)^2 W^2(X)\Big] = o(1)$$

for any $\rho(\upsilon_1, \upsilon_2) = o(1)$, by the Hölder continuity assumption.

A.6.2 Proof of Lemma A.2

Define $Z_{ni}(\upsilon) \equiv \frac{1}{\sqrt{n}} A(y, S_i) W(S_{wi})$, indexed by $\upsilon \equiv (y, W) \in \Upsilon \equiv \mathcal{Y} \times \mathcal{M}$. The bracketing CLT will imply $\sum_{i=1}^n \Big(Z_{ni}(\upsilon) - E Z_{ni}(\upsilon)\Big)$ is asymptotic stochastic equicontinuous in υ with respect to the pseudo-metric $\rho(\upsilon_1, \upsilon_2) = \max\{|y_1 - y_2|, \|W_1 - W_2\|_\infty\}$. The conditions for the bracketing CLT from Theorem 2.11.9 in van der Vaart and Wellner (1996) are checked in the following:

(i) Since the functions are assumed to be uniformly bounded above and below, $\mathbf{1}_{\{\|Z_{ni}\|_\Upsilon > \eta\}} = 0$ for n large enough. So for any $\eta > 0$, $\sum_{i=1}^n E\Big[\|Z_{ni}\|_\Upsilon \mathbf{1}_{\{\|Z_{ni}\|_\Upsilon > \eta\}}\Big] = o_p(1)$.

(ii)

$$N_{[\cdot]}(\epsilon, \Upsilon, L_2) \leq N\Big(\frac{\epsilon}{2C}, \mathcal{Y}, |\cdot|\Big) \times N\Big(\frac{\epsilon}{2C}, \mathcal{M}, \|\cdot\|_\infty\Big).$$

(iii)

$$\sum_{i=1}^{n} E\big(Z_{ni}(\upsilon_1) - Z_{ni}(\upsilon_2)\big)^2 = E\Big(A(y_1,S)W_1(S_w) - A(y_2,S)W_2(S_w)\Big)^2$$
$$= E\Big(A(y_1,S)\big(W_1(S_w) - W_2(S_w)\big) + \big(A(y_1,S) - A(y_2,S)\big)W_2(S_w)\Big)^2 = o(1)$$

for any $\rho(\upsilon_1,\upsilon_2) = o(1)$, by the Hölder continuity.

A.6.3 Proof of Lemma A.3

The proof modifies the proof of Lemma 1 in Mammen et al. (2012a) and is presented for completeness. The difference is I replace the residual in MRS with $\mathbf{1}_{\{Y_i \le y\}}$ and the expansion is uniform in y.

When $\kappa_1 \le (\delta-\eta)_{min}$, the results hold from a direct bound. Consider the case $\kappa_1 > (\delta-\eta)_{min}$. Define $\Delta_i(y,V_1,V_2) \equiv \mathbf{1}_{\{Y_i\le y\}}K_h(T_i-t)\Big(K_h(V_1(T_i,S_i)-v) - K_h(V_2(T_i,S_i)-v)\Big) - E\Big[\mathbf{1}_{\{Y\le y\}}K_h(T-t)\Big(K_h(V_1(T,S)-v) - K_h(V_2(T,S)-v)\Big)\Big]$. First note that (i) $|\frac{1}{n}\sum_{i=1}^n \Delta_i(y,V_1,V_2)| \le C\max_j \|V_{1j}-V_{2j}\|/h_j$. (ii) $E\Delta_i(y,V_1,V_2)^2 \le Cn^{\eta_+}\big(\max_j \|V_{1j}-V_{2j}\|/h_j\big)^2$. (iii) $|\Delta_i(y,V_1,V_2)| \le Cn^{\eta_+}\max_j \|V_{1j}-V_{2j}\|/h_j$.

For $s \ge 0$, let $\bar{\mathcal{M}}^*_{s,n,j}$ be a set of functions chosen such that for each $V_j \in \bar{\mathcal{M}}_{n,j}$, there exists $V_j^* \in \bar{\mathcal{M}}^*_{s,n,j}$ such that $\|V_j - V_j^*\|_\infty \le 2^{-s}n^{-\delta_j}$. Define $\bar{\mathcal{M}}^*_{s,n} = \bar{\mathcal{M}}^*_{s,n,1} \times ... \times \bar{\mathcal{M}}^*_{s,n,d}$. For $V_1, V_2 \in \bar{\mathcal{M}}_n$, choose $V_1^s, V_2^s \in \bar{\mathcal{M}}^*_{s,n}$ such that $\|V^s_{1,j} - V_{1,j}\|_\infty \le 2^{-s}n^{-\delta_j}$ and $\|V^s_{2,j} - V_{2,j}\|_\infty \le C2^{-s}n^{-\delta_j}$ for all j and $s \ge 0$. The functions in $\bar{\mathcal{M}}^*_{s,n,j}$ are the midpoints of a $(2^{-s}n^{-\delta_j})$-covering of $\bar{\mathcal{M}}_{n,j}$. So the cardinality $\#\bar{\mathcal{M}}^*_{s,n,j}$ is at most $C\cdot\exp\Big(\big(2^{-s}n^{-\delta_j}\big)^{-\beta_j}n^{\xi_j}\Big)$.

Consider the chain $\Delta_i(y,V_1,V_2) = \Delta_i(y,V_1^0,V_2^0) - \sum_{s=1}^{G_n}\Delta_i(y,V_1^{s-1},V_1^s)$
$+\sum_{s=1}^{G_n}\Delta_i(y,V_2^{s-1},V_2^s) - \Delta_i(y,V_1^{G_n},V_1) + \Delta_i(y,V_2^{G_n},V_2)$, where G_n is chosen to be the smallest integer that satisfies $G_n > (1+c_G)(\kappa_1 - (\delta-\eta)_{min})\log n/\log 2$ for a constant $c_G > 0$. So that for $l = 1,2$, by (i), uniformly in $y \in \mathcal{Y}$,

$$T_1 \equiv |\frac{1}{n}\sum_{i=1}^n \Delta_i(y,V_l^{G_n},V_l)| \le C2^{-G_n}n^{-(\delta-\eta)_{min}} \le Cn^{-\kappa_1}.$$

For any $a > c_G$, define the constant $c_a = (\sum_{s=1}^{\infty} 2^{-as})^{-1}$.

$$
\begin{aligned}
&Pr\Big(\sup_{V_1 \in \bar{\mathcal{M}}_n, y \in \mathcal{Y}} \Big| \frac{1}{n} \sum_{i=1}^{n} \sum_{s=1}^{G_n} \Delta_i(y, V_1^{s-1}, V_1^s) \Big| > n^{-\kappa_1} \Big) \\
&\leq Pr\Big(\sum_{s=1}^{G_n} \sup_{V_1 \in \bar{\mathcal{M}}_n, y \in \mathcal{Y}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) \Big| > \sum_{s=1}^{G_n} c_a 2^{-as} n^{-\kappa_1} \Big) \\
&\leq \sum_{s=1}^{G_n} Pr\Big(\sup_{V_1 \in \bar{\mathcal{M}}_n, y \in \mathcal{Y}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) \Big| > c_a 2^{-as} n^{-\kappa_1} \Big) \\
&= \sum_{s=1}^{G_n} Pr\Big(\max_{V_1^{s-1} \in \bar{\mathcal{M}}^*_{s-1,n}, V_1^s \in \bar{\mathcal{M}}^*_{s,n}} \sup_{y \in \mathcal{Y}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) \Big| > c_a 2^{-as} n^{-\kappa_1} \Big) \\
&\leq \sum_{s=1}^{G_n} \sum_{\bar{\mathcal{M}}^*_{s,n}} \sum_{\bar{\mathcal{M}}^*_{s-1,n}} Pr\Big(\sup_{y \in \mathcal{Y}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) \Big| > c_a 2^{-as} n^{-\kappa_1} \Big) \qquad \text{(A.29)} \\
&\leq \sum_{s=1}^{G_n} \#\bar{\mathcal{M}}^*_{s-1,n} \#\bar{\mathcal{M}}^*_{s,n} Pr\Big(\frac{1}{n} \sum_{i=1}^{n} \Delta_i(y^s_{sup}, V_1^{*,s}, V_1^{**,s}) > c_a 2^{-as} n^{-\kappa_1} \Big) \qquad \text{(A.30)} \\
&\quad + \sum_{s=1}^{G_n} \#\bar{\mathcal{M}}^*_{s-1,n} \#\bar{\mathcal{M}}^*_{s,n} Pr\Big(\frac{1}{n} \sum_{i=1}^{n} \Delta_i(y^s_{inf}, \tilde{V}_1^{*,s}, \tilde{V}_1^{**,s}) < -c_a 2^{-as} n^{-\kappa_1} \Big) \equiv T_2 + T_3, \qquad \text{(A.31)}
\end{aligned}
$$

In (A.30), denoting $c_a 2^{-as} n^{-\kappa_1} \equiv C$,

$$
\begin{aligned}
&Pr\Big(\sup_{y \in \mathcal{Y}} \Big| \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) \Big| > C \Big) \\
&\quad \leq Pr\Big(\sup_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) > C \Big) + Pr\Big(\inf_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) < -C \Big) \\
&\quad = Pr\Big(\frac{1}{n} \sum_{i=1}^{n} \Delta_i(y^s_{sup}, V_1^{s-1}, V_1^s) > C \Big) + Pr\Big(\frac{1}{n} \sum_{i=1}^{n} \Delta_i(y^s_{inf}, V_1^{s-1}, V_1^s) < -C \Big).
\end{aligned}
$$

Because $\mathcal{Y}$ is compact and $\frac{1}{n}\sum_{i=1}^{n} \Delta_i$ is a piecewise constant function that jumps at observed values of Y only, there exists some y^s_{sup} and y^s_{inf} such that $\sup_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y^s_{sup}, V_1^{s-1}, V_1^s)$ and $\inf_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y, V_1^{s-1}, V_1^s) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i(y^s_{inf}, V_1^{s-1}, V_1^s)$.

In (A.29) and (A.31), the functions $V_1^{*,s}, \tilde{V}_1^{*,s} \in \bar{\mathcal{M}}_{s-1,n}^*$ and $V_1^{**,s}, \tilde{V}_1^{**,s} \in \bar{\mathcal{M}}_{s,n}^*$ are chosen such that

$$Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y_{sup}^s, V_1^{*,s}, V_1^{**,s}) > c_a 2^{-as} n^{-\kappa_1}\Big) = \max_{V_1^{s-1}, V_1^s} Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y_{sup}^s, V_1^{s-1}, V_1^s) > c_a 2^{-as} n^{-\kappa_1}\Big)$$
$$Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y_{inf}^s, \tilde{V}_1^{*,s}, \tilde{V}_1^{**,s}) < -c_a 2^{-as} n^{-\kappa_1}\Big) = \max_{V_1^{s-1}, V_1^s} Pr\Big(\frac{1}{n}\sum_{i=1}^n \Delta_i(y_{inf}^s, V_1^{s-1}, V_1^s) < -c_a 2^{-as} n^{-\kappa_1}\Big).$$

The following shows T_2 and $T_3 \leq \exp(-cn^c)$, tending to zero at an exponential rate. By Markov inequality, $T_2 \leq$

$$C\sum_{s=1}^{G_n} \Pi_j \exp\big((2^{-s}n^{-\delta_j})^{-\beta_j} n^{\xi_j}\big)(1+2^{-\beta_j}) E\Big[\exp\big(\gamma_{n,s}\frac{1}{n}\sum_{i=1}^n \Delta_i(y_{sup}^s, V_1^{*,s}, V_1^{**,s}) - \gamma_{n,s} c_a 2^{-as} n^{-\kappa_1}\big)\Big]$$
$$\leq C\sum_{s=1}^{G_n} \exp\big(\sum_j 2^{s\beta_j} n^{\delta_j\beta_j+\xi_j} - \gamma_{n,s} c_a 2^{-as} n^{-\kappa_1}\big)\Pi_{i=1}^n E\Big[\exp\big(\gamma_{n,s}\frac{1}{n}\Delta_i(y_{sup}^s, V_1^{*,s}, V_1^{**,s})\big)\Big],$$

where $\gamma_{n,s} = c_\gamma 2^{(2-a)s} n^{-\kappa_1+1-\eta_++2(\delta-\eta)_{min}}$ with a constant $c_\gamma > 0$ small enough. For the last term, use the equality $Ee^x \leq 1 + |x|Ex^2 \leq 1 + CEx^2 \leq \exp(cEx^2)$ by $Ex = 0$ and $|x| \leq C$ for some $C > 0$.

$$E\Big[\exp\big(\gamma_{n,s}\frac{1}{n}\Delta_i(y_{sup}^s, V_1^{*,s}, V_1^{**,s})\big)\Big] \leq \exp\big(C\gamma_{n,s}^2 n^{-2} n^{\eta_+-2(\delta-\eta)_{min}} 2^{-2s}\big),$$

by (ii). To show $|x| \leq C$, by (iii),

$$\big|\gamma_{n,s}\frac{1}{n}\Delta_i(y_{sup}^s, V_1^{*,s}, V_1^{**,s})\big| \leq C\gamma_{n,s}\frac{1}{n} n^{\eta_+-(\delta-\eta)_{min}} 2^{-s}$$
$$\leq Cn^{(\delta-\eta)_{min}-\kappa_1} 2^{-as+s} \leq Cn^{(c_G-a)(\kappa_1-(\delta-\eta)_{min})} \leq C.$$

When $a < 1$, $Cn^{(\delta-\eta)_{min}-\kappa_1} 2^{-as+s} \leq Cn^{(\delta-\eta)_{min}-\kappa_1} 2^{G_n(1-a)}$. The above inequality holds by the chosen G_n, When $a \geq 1$, the above inequality holds for n large enough. Therefore,

$$T_2 \leq C\sum_{s=1}^{G_n} \exp\Big(\sum_j 2^{s\beta_j} n^{\delta_j\beta_j+\xi_j} - \gamma_{n,s} c_a 2^{-as} n^{-\kappa_1} + C\gamma_{n,s}^2 n^{-1+\eta_+-2(\delta-\eta)_{min}} 2^{-2s}\Big)$$
$$= C\sum_{s=1}^{G_n} \exp\Big(\sum_j 2^{s\beta_j} n^{\delta_j\beta_j+\xi_j} - c 2^{2(1-a)s} n^{1-2\kappa_1-\eta_++2(\delta-\eta)_{min}}\Big) \leq C\sum_{s=1}^{G_n} \exp(-c^s n^c) \leq \exp(-cn^c).$$

$\gamma_{n,s}$ is chosen so that the last two terms in the first line is of the same order. And choose a and c_γ to be small enough, so that the sum of the last two terms is negative. Then κ_1 is chosen so that the second term in the second line dominates. Similarly, $T_3 \leq \exp(-cn^c)$.

Because $\bar{\mathcal{M}}^*_{0,n}$ can always be chosen such that it contains only a single element and (i),

$$T_4 = Pr(\sup_{V_1,V_2\in\bar{\mathcal{M}}_n,y\in\mathcal{Y}} \big|\frac{1}{n}\sum_{i=1}^n \Delta_i(y,V_1^0,V_2^0)\big| > n^{-\kappa_1}) \leq \exp(-cn^c).$$

Therefore,

$$\sup_{v\in\mathcal{V},t\in\mathcal{T}} Pr\Big(\sup_{V_1,V_2\in\bar{\mathcal{M}}_n,y\in\mathcal{Y}} \Big|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i\leq y\}}K_h(T_i-t)\Big(K_h(V_1(T_i,S_i)-v)-K_h(V_2(T_i,S_i)-v)\Big)$$
$$-E\Big[\mathbf{1}_{\{Y\leq y\}}K_h(T-t)\Big(K_h(V_1(T,S)-v)-K_h(V_2(T,S)-v)\Big)\Big]\Big| \geq Cn^{-\kappa_1}\Big) \leq \exp(-cn^c)$$

For the uniformity in $(t,v)\in\mathcal{T}\times\mathcal{V}$, for $C_t>0$, choose a grid $\mathcal{T}_n\times\mathcal{V}_n$ with $O(n^{C_t})$ points, such that for each $(t,v)\in\mathcal{T}\times\mathcal{V}$, there exists a grid point $(t^*,v^*)=(t^*(t),v^*(v))\in\mathcal{T}_n\times\mathcal{V}_n$ such that $\|t-t^*\|\leq n^{-cC_t}$ and $\|v-v^*\|\leq n^{-cC_t}$. Choosing C_t large enough implies

$$\sup_{v\in\mathcal{V},t\in\mathcal{T},y\in\mathcal{Y},V\in\bar{\mathcal{M}}_n} \Big|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i\leq y\}}\Big(K_h(T_i-t)K_h(V(T_i,S_i)-v)-K_h(T_i-t^*)K_h(V(T_i,S_i)-v^*)\Big)$$
$$-E\Big[\mathbf{1}_{\{Y\leq y\}}\Big(K_h(T-t)K_h(V(T,S)-v)-K_h(T-t^*)K_h(V(T,S)-v^*)\Big)\Big]\Big| \leq Cn^{cC_t}/h \leq n^{-\kappa_1}$$

for large enough n with probability tending to one. Using the triangle inequality, the statement in this lemma is proved.

A.6.4 Proof of Lemma A.4

The proof is implied by the proof of Lemma A.3, where $\Delta_i(V_1,V_2)\equiv A(y,t,W_i,V_i;S)\big(K_h(V_1(t,S)-V_i)-K_h(V_2(t,S)-V_i)\big)-E_{WV}\Big[A(y,t,W,V;S)\big(K_h(V_1(t,S)-V)-K_h(V_2(t,S)-V)\big)\Big]$. Note that the following still holds the same as the proof of Lemma A.3:
(i) $|\frac{1}{n}\sum_{i=1}^n \Delta_i(V_1,V_2)| \leq C\max_j\|V_{1j}-V_{2j}\|/h_j$. (ii) $E\Delta_i(V_1,V_2)^2 \leq Cn^{\eta_+}\big(\max_j\|V_{1j}-V_{2j}\|/h_j\big)^2$. (iii) $|\Delta_i(V_1,V_2)|\leq Cn^{\eta_+}\max_j\|V_{1j}-V_{2j}\|/h_j$. Therefore, it follows that

$$\sup_{T\in\mathcal{T},S\in\mathcal{S},y\in\mathcal{Y}} Pr\Big(\sup_{V_1,V_2\in\bar{\mathcal{M}}_n} \Big|\frac{1}{n}\sum_{i=1}^n A(y,t,W_i,V_i;S)\big(K_h(V_1(t,S)-V_i)-K_h(V_2(t,S)-V_i)\big)$$
$$-E_{WV}\Big[A(y,t,W,V;S)\big(K_h(V_1(t,S)-V)-K_h(V_2(t,S)-V)\big)\Big]\Big| \geq Cn^{-\kappa_1}\Big) \leq \exp(-cn^c).$$

For uniformity in $S \in \mathcal{S}$, for $C > 0$, choose a grid $\mathcal{S}_n$ with $O(n^C)$ points, such that for each $S \in \mathcal{S}$, there exists a grid point $S^* \in \mathcal{S}_n$ such that $\|S - S^*\| \leq n^{-cC}$.

$$\sup_{S,y,V} \left| \frac{1}{n} \sum_{i=1}^{n} A(y,t,W_i,V_i;S) K_h(V(t,S) - V_i) - A(y,t,W_i,V_i;S^*) K_h(V(t,S^*) - V_i) \right|$$

$$\leq \sup_{S,y,V} \left| \frac{1}{n} \sum_{i=1}^{n} A(y,t,W_i,V_i;S^*) \Big(K_h(V(t,S) - V_i) - K_h(V(t,S^*) - V_i) \Big) \right| \tag{A.32}$$

$$+ \sup_{S,y,V} \left| \frac{1}{n} \sum_{i=1}^{n} \Big(A(y,t,W_i,V_i;S) - A(y,t,W_i,V_i;S^*) \Big) K_h(V(t,S^*) - V_i) \right| \tag{A.33}$$

$$+ \sup_{S,y,V} \left| \frac{1}{n} \sum_{i=1}^{n} \Big(A(y,t,W_i,V_i;S) - A(y,t,W_i,V_i;S^*) \Big) \Big(K_h(V(t,S) - V_i) - K_h(V(t,S^*) - V_i) \Big) \right|.$$

(A.32) $\leq n^{-\kappa_1}$ for large enough n if C_s is chosen large enough. (A.33) $\leq n^{-\kappa_1}$ for large enough n by the smoothness of A.

A.6.5 Proof of Lemma A.5

The proof is implied by the proof of Lemma A.3, where

$$\Delta_i(V_1,V_2) \equiv A(y,T_i,S_i,W_i) \Big(\hat{V}(T_i,S_i) - V(T_i,S_i) \Big) - E_{SW} \Big[A(y,T,S,W) \Big(\hat{V}(T_i,S) - V(T_i,S) \Big) \Big]$$

and $A(y,T_i,S_i,W_i) = \nabla_v F_{Y|TV}(y|t,V(T_i,S_i))W_i$. Note that the following still holds the same as the proof of Lemma A.3: (i) $|\frac{1}{n}\sum_{i=1}^{n} \Delta_i(V_1,V_2)| \leq C \max_j \|V_{1j} - V_{2j}\|$. (ii) $E\Delta_i(V_1,V_2)^2 \leq C \max_j \|V_{1j} - V_{2j}\|^2$. (iii) $|\Delta_i(V_1,V_2)| \leq C \max_j \|V_{1j} - V_{2j}\|$. That is, the proof is essentially the same for the case $\eta_j = 0$.

Appendix B: Supplmentary Appendix to Chapter 2

For notational ease, I define the following: $Q_i \equiv Q(X_i) \equiv Q_\tau(Y|X_i)$, $G_{ij} = G\big(\frac{Q_\tau(Y|X_i)-y_j}{h_0}\big)$, $K_{ij} = K\big(H^{-1}(X_i - X_j)\big)$, $f_i = f(X_i)$. And $\kappa_\nu = \int K(Z)Z^\nu dZ$ and $\kappa_{G2} = \int G'(z)z^2dz$. For the Gaussian kernel $G = \Phi$, $\kappa_{G2} = 1$. $R(K) = \int K^2(Z)dZ$ is the roughness of kernel K. ∂_k denotes the partial derivative with respect to the kth component of X. *s.o.* represents smaller order terms. *w.p.a.*1 means with probability approaching one. For a $q \times 1$ vector a, $\|a'a\| = \|a\|^2$. For a $m \times n$ matrix A, I use Frobenius norm: $\|A\| = trace(A'A)^{1/2}$. Define the operator Γ to a function $g : V \to R$ where V is an open and convex subset of R^q:

$$\Gamma g(X + HZ) \equiv \Gamma g(\bar{X}) = h\sum_{k=1}^{q} \partial_k g(X)Z_k + \frac{h^2}{2}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q}\Big[\partial_{k_1}\partial_{k_2} g(X)\Big]Z_{k_1}Z_{k_2} + \dots$$

$$+ \frac{h^{\nu-1}}{(\nu-1)!}\sum_{k_1=1}^{q}\cdots\sum_{k_{\nu-1}=1}^{q}\Big[\partial_{k_1}\cdots\partial_{k_{\nu-1}} g(X)\Big]Z_{k_1}\cdots Z_{k_{\nu-1}} \tag{B.1}$$

$$+ \frac{h^{\nu}}{\nu!}\sum_{k_1=1}^{q}\cdots\sum_{k_{\nu}=1}^{q}\Big[\partial_{k_1}\cdots\partial_{k_{\nu}} g(\bar{X})\Big]Z_{k_1}\cdots Z_{k_{\nu}}, \tag{B.2}$$

where Z_k is the kth component of the vector Z and $\bar{X}$ is on the line segment of X and $X + HZ$. Hence, Taylor's theorem expands $g(X + HZ) = g(X) + \Gamma g(X + HZ)$ for small H.

A.FX The marginal density of X is denoted as f with convex (possibly unbounded) support $\mathcal{X} \subseteq R^q$ whose interior is nonempty, i.e., $f(X) > 0$ for all $X \in \mathcal{X}_0$, the interior of $\mathcal{X}$. $f(X) = 0$ for all X on the boundary of $\mathcal{X}$. $\sup_{X\in\mathcal{X}} f(X)$ is bounded above. The $q \times 1$ vector $\nabla f(X)$ with its sth component $\partial_{X_s} f(X)$, is p_X times differentiable, and its p_Xth-order derivative $\partial_{X_{k_1}} \dots \partial_{X_{k_{p_X}}} \nabla f(X)$ for $k_1, \dots, k_{p_X} \in \{1, ..., q\}$ is uniformly continuous in X, for all $X \in \mathcal{X}_0$.

A.FY The conditional density function of Y given X is bounded away from zero on a convex and compact support, $\mathcal{Y} \equiv [\underline{y}, \bar{y}] \subset R$, i.e., $f_Y(y|X) > b_1 > 0$, $\forall y \in \mathcal{Y}$ and $\forall X \in \mathcal{X}_0$, the interior of $\mathcal{X}$.

The conditional density function of Y, $f_Y(y|X)$, is differentiable with respect to X, and its p_Yth-order derivatives $\partial_{k_1}...\partial_{k_{p_Y}} f'_Y(y|X)$ for $k_1, ..., k_{p_Y} \in \{1, ..., q\}$ are uniformly continuous in X, for all $X \in \mathcal{X}_0$ and for all $y \in \mathcal{Y}$.

$F_Y(y|X)$ is twice differentiable with respect to y, and its second order derivative $f'_Y(y|X)$ is uniformly continuous in y, $\forall X \in \mathcal{X}_0$ and $\forall y \in \mathcal{Y}$.

$f_Y(y|X)$ and $f'_Y(y|X)$ are bounded above almost surely.

A.Q For $\tau \in \mathcal{T}$, $lim_{X \to \partial\mathcal{X}} Q_\tau(Y|X) f^2(X) = 0$, where $\partial\mathcal{X}$ is the boundary of the support of X.

The $q \times 1$ vector $\nabla Q_\tau(Y|X)$ with its sth component $\partial_{X_s} Q(X)$, is p_Q times differentiable with respect to X, and its p_Qth order derivative $\partial_{X_{k_1}} \dots \partial_{X_{k_{p_Q}}} \nabla Q(X)$ for $k_1, \dots, k_{p_Q} \in \{1, ..., q\}$ is uniformly continuous in X, for all $X \in \bar{\mathcal{X}}$, where $\bar{\mathcal{X}}$ differs from $\mathcal{X}_0$ by a set of measure zero.

A.M The following moments exist: For uniform convergence: $E|Y|^s$ for some $s > 2$.

For trimming, $E\|Q_i \nabla f_i\|^2$, $E\|\nabla f_i\|^2$ and EQ_i^2. For $r_{1,n}$: $E\Big[\|\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\|^2\Big]$.

For the projection of U-statistic in Lemma B.2: $E\Big[\Big\|\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\Big\|^2 \Big[f_i\Big(1 + f_Y(Q_i|X_i) + f'_Y(Q_i|X_i)\Big) + \sum_{k=1}^q \partial_k f_i \sum_{k=1}^q \partial_k \Big(F_Y(Q_i|X_i) + f_Y(Q_i|X_i) + f'_Y(Q_i|X_i)\Big)\Big]\Big]$

and $E\Big[\Big\|\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\Big\|^2 f_i \Big[\sum_{k=1}^q \partial_k F_Y(Q_i|X_i) + f'_Y(Q_i|X_i) + \sum_{k=1}^q \partial_k f'_Y(Q_i|X_i)\Big]$
$\Big[f_Y(Q_i|X_i) \sum_{k=1}^q \partial_k Q_i + f'_Y(Q_i|X_i) + f''_Y(Q_i|X_i) \sum_{k=1}^q \partial_k Q_i\Big]\Big]$.

For t_{1n}: $E\Big[\Big\|\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\Big\|^2 \Big| f_i f'_Y(Q_i|X_i) + \sum_{l=1}^\nu \frac{1}{l!(\nu-l)!} \sum_{k=1}^q \partial_k^l F_Y(Q_i|X_i) \cdot \partial_k^{\nu-l} f_i$
$+ \sum_{l=0}^\nu \frac{1}{l!(\nu-l)!} \sum_{k=1}^q \partial_k^l f'_Y(Q_i|X_i) \cdot \partial_k^{\nu-l} f_i \Big|^2\Big]$.

For t_{2n}: $E\Big[\frac{\nabla f_i \nabla f'_i}{f_Y(Q_i|X_i)}\Big(1 + \sum_{k=1}^q \partial_k Q_i + \sum_{k=1}^q \partial_k^\nu Q_i\Big)\Big]$.

For Lemma C.1: $E\Big[\big(\sum_{k=1}^q \partial_k Q_i\big)^2 \big(f_i + \sum_{k=1}^q \partial_k f_i\big)\Big]$;

for ($I\!I$): $var(Q_i \nabla f_i)$, $var(Q_i \sum_{k=1}^q \partial_k^\nu \nabla f_i)$,

A.T Define $B_n \equiv \{X : f(X) < \delta\}$, where the trimming parameter δ satisfies Assumption A.B. $\int_{B_n} \|Q_i \nabla f_i\| f_i dX_i = o(n^{-1/2})$.

A.L Denote $\frac{\nabla f(X)}{f_Y(Q(X)|X)} \equiv A(X)$ for notational ease. Lipschitz conditions: for some $m(X)$,

$$\|\nabla f(X+V) - \nabla f(X)\| < m(X)\|V\|$$
$$\|\nabla (f\cdot Q)(X+V) - \nabla (f\cdot Q)(X)\| < m(X)\|V\|$$
$$\|A(X+V) - A(X)\| < m(X)\|V\|$$

with the existence of the moments: $E\big[(1+|Q(X_i)|)^2 m(X_i)^2\big]$, $E\big[m(X)\|A(X)\|\big]$, and $E\Big[\|A(X)\|m(X)\Big[f'_Y(Q_i|X_i)^2 + \sum_{k=1}^q \partial_k Q_i\Big(f_Y(Q_i|X_i) + f''_Y(Q_i|X_i)\Big)\sum_{k=1}^q \partial_k\Big(F_Y(Q_i|X_i) + f'_Y(Q_i|X_i)\Big)\Big]\Big]$.

A.K For the $q\times 1$ vector Z, define the product kernel $K(Z) \equiv k(Z_1)k(Z_2)\cdots k(Z_q)$, where Z_s denotes the sth component of $Z = (Z_1, ..., Z_s, ..., Z_q)'$. The kernel is bounded and integrable: $|k(x)| \le \bar{K} < \infty$ and $\int_R |k(z)|dz \le \mu < \infty$.

The symmetric kernel k with convex bounded support has order of ν, i.e., $\kappa_j = \int k(z)z^j dz = 0$ for $j < \nu$ and $\kappa_\nu \in (0,\infty)$. $\int k(z)^2 z^2 dz < \infty$ and $\int k(z)^2 z^4 dz < \infty$. For some $\Lambda_1 < \infty$, for all $z, z' \in R^q$, $|k(z) - k(z')| \le \Lambda_1\|z - z'\|$.

A.G Let $G(z) = \int_{-\infty}^z g(t)dt$, where the second order kernel g with bandwidth h_0 is everywhere positive on a convex support. $g(u)$ is differentiable. For some $\Lambda_1 < \infty$ and $L < \infty$, $|\frac{\partial}{\partial u}g(u)| \le \Lambda_1$, for some $m > 4$, $|\frac{\partial}{\partial u}g(u)| \le \Lambda_1\|u\|^{-m}$, for $\|u\| > L$. Therefore, for any $z < 0$, $G(z/h_0) = o(h_0^2)$. [1]

A.B Let ν, ν_1, and the positive sequences h, h_1, h_0 satisfy $\delta^{-2}h^{-q}(nh_0)^{-1/2} \to 0$, $\delta^{-2}h^{-q}(nh_1^{q+2})^{-1} \to 0$, $\sqrt{n}(h_0^2 + h^\nu + h_1^{\nu_1}) \to 0$. An alternative sufficient condision is that $h \propto n^{-a}$, $h_1 \propto n^{-c}$, $h_0 \propto n^{-d}$, and $\delta \propto n^{-b}$, for some constants, $a, b, c, d > 0$. Choose $\nu > \frac{4q}{3}$, $a \in (\frac{1}{2\nu}, \frac{3}{8q})$, $\nu_1 > \frac{q+2}{2-2aq}$, $c \in (\frac{1}{2\nu_1}, \frac{1-aq}{q+2})$, $d \in (\frac{1}{4}, 1-2aq)$, and $b < min\{\frac{1}{4}(1-2aq-d), \frac{1}{2}(1-aq-c(q+2))\}$.

Assumption A.FX implies the covariate X to be continuous. Assumption A.FY implies $F_Y(y|X)$ is continuous and strictly increasing in y, so the conditional quantile function

[1] I put stronger assumption for $m > 4$, instead of $m > 1$, such that for any $z < 0$, $G(z/h_0) = o(h_0^2)$, $lim_{z\to\infty} g(z)z^2 = 0$, and $g(z/h_0^2)/h_0^2 = o(h_0)$ used in Lemma B.4.

$Q_\tau(Y|X) = F_Y^{-1}(\tau|X)$, the inverse function, is uniquely defined almost surely. Assumption A.Q implies $Q_\tau(Y|X)$ is smooth in X such that $\nabla Q_\tau(Y|X)$ exists almost surely and $\beta(\tau)$ in equation (2.1) is well defined. The νth-order differentiability of these functions will be used in deriving asymptotic bias for the kernel estimation. The support of X cannot be compact. Assumption A.T restricts how fast the unknown functions approach zero which is the same as the stochastic trimming in other content; e.g.,Hardle and Stoker (1989), Lavergne and Vuong (1996). So that the bias induced by the trimming vanishes faster than the parametric rate. Following Powell et al. (1989), the Lipschitz conditions in Assumption A.L impose standard bounded moment and dominance conditions. Assumption A.K restricts the kernel to have bounded support which is used for asymptotic trimming. The kernel has truncated support and is Lipschitz as in Assumption 3 in Hansen (2008). The covariates Z can be normalized so that the bandwidths equal to the same h for all components of Z. Assumption A.G restricts the tail behavior of the kernel g which has unbounded convex support. The commonly used Gaussian kernel $g = \phi$ satisfies this assumption.

B.1 Uniform Convergence of Kernel Estimation

I show the bias of $\hat{f}(x)$ and $\nabla\hat{f}(x)$ in the following Lemmas, where the same technical proof is used repeatedly. The kernel estimation $\hat{f}(x)$ is defined in (2.8).

Lemma B.1. *Suppose Assumptions A.FX, and A.K hold. Let $p_X \geq \nu$. Then $E\big[\hat{f}(x)\big] = f(x) + \frac{h^\nu}{\nu!}\sum_{k=1}^q \partial_k^\nu f(x)\kappa_\nu + o(h^\nu)$ and $E\big[\nabla\hat{f}(x)\big] = \nabla f(x) + \frac{h^\nu}{\nu!}\kappa_\nu \sum_{k=1}^q \partial_k^\nu \nabla f(X_i) + o(h^\nu)$.*

Proof.

$$E[\hat{f}(x)] = \int_{\mathcal{X}} \frac{1}{|H|}K(H^{-1}(X_j - x))f(X_j)dX_j = \int_{\mathcal{Z}} K(Z)f(x + HZ)dZ$$

$$= \int_{\mathcal{Z}} K(Z)\Big[f(X) + \Gamma f(x + HZ)\Big]dZ$$

$$= f(X) + \int_{\mathcal{Z}} K(Z)\Big[\frac{h^\nu}{\nu!}\sum_{k=1}^q \partial_k^\nu f(x)Z_k^\nu\Big]dZ \tag{B.3}$$

$$+ \int_{\mathcal{Z}} K(Z)\Big[\frac{h^\nu}{\nu!}\sum_{k_1=1}^q ... \sum_{k_\nu=1}^q \partial_{k_1}...\partial_{k_\nu}\Big(f(\bar{x}) - f(x)\Big)Z_{k_1}...Z_{k_\nu}\Big]dZ, \tag{B.4}$$

where the second equality is by change of variable $Z = H^{-1}(X_j - x)$ and the third equality is by Taylor's theorem. Since $\partial_{k_1} \cdots \partial_{k_\nu} f(\bar{x})$ is uniformly continuous and $\bar{x} \to x$ as $h \to 0$, it converges to $\partial_{k_1} \cdots \partial_{k_\nu} f(x)$ as $h \to 0$. I can apply dominated convergence theorem (DCT) to the last term containing $\bar{x}$ in $\Gamma f(x + HZ)$. That is, the last term in equation (B.4) is $o(h^\nu)$. Also, by Assumption A.K, $\kappa_j = \int k(z) z^j dz = 0$ for $j < \nu$ implies the forth equality.

Let's focus on the sth component of $\nabla \hat{f}(x)$,

$$
\begin{aligned}
E[\partial_s \hat{f}(x)] &= \int_{\mathcal{X}} \frac{1}{nh^{(q+1)}} \sum_{i=1}^{n} k'\Big(\frac{x_s - X_{js}}{h}\Big) \Pi_{t \neq s} k\Big(\frac{x_t - X_{jt}}{h}\Big) f(X_j) dX_j \\
&= \int_{\mathcal{X}} \frac{1}{h^q} k\Big(\frac{x_s - X_{js}}{h}\Big) \Pi_{t \neq s} k\Big(\frac{x_t - X_{jt}}{h}\Big) \partial_s f(X_j) dX_j = \int_{\mathcal{Z}} K(Z) \partial_s f(x + HZ) dZ \\
&= \int_{\mathcal{Z}} K(Z) \Big(\partial_s f(x) + \Gamma\big[\partial_s f(x + HZ)\big] \Big) dZ = \partial_s f(x) + \frac{h^\nu}{\nu!} \sum_{k=1}^{q} \partial_k^\nu \partial_s f(x) \kappa_\nu + o(h^\nu),
\end{aligned}
$$

The second equality is integration by parts of the element X_{js}. □

Lemma B.2. *Suppose Assumptions A.FX, A.FY, and A.K hold. Let $p_X, p_Y \geq \nu$. $\mathcal{S} = \{X : f(X) \geq \delta\}$ is a compact subset of $\mathcal{X}$.*

$$
\sup_{X \in S, y \in \mathcal{Y}} |\hat{f}_Y(y|X) - f_Y(y|X)| = O_p\Big(\frac{1}{\delta}\Big(\big(\frac{\log n}{nh_0 h^q}\big)^{1/2} + h_0^2 + h^\nu \Big)\Big)
$$

and

$$
\sup_{X \in S, y \in R} |\hat{F}_Y(y|X) - F_Y(y|X)| = O_p\Big(\frac{1}{\delta}\Big(\big(\frac{\log n}{nh^q}\big)^{1/2} + h_0^2 + h^\nu \Big)\Big).
$$

Proof of Lemma B.2. I modify the uniform convergence results for kernel estimation in Hansen (2008). Following the proof of Theorem 8 in Hansen (2008), $\hat{F}_Y(y|X) \equiv \hat{g}(y, X)/\hat{f}(X) = \frac{\hat{g}(y,X)/f(X)}{\hat{f}(X)/f(X)}$ for $X \in \mathcal{S}$, where $\hat{g}(y, X) \equiv \frac{1}{nh^q} \sum_{i=1}^{n} K(H^{-1}(X_i - X)) G(\frac{y - y_i}{h_0})$. Since $G(\frac{y-y_i}{h_0})$ is bounded between 0 and 1 for all $y \in R$, Hansen's proof of Theorem 2 gives $\sup_{y \in R} \sup_{X \in S} |\hat{g}(y, X) - E\hat{g}(y, X)| = O_p\Big(\Big(\frac{\log n}{nh^q}\Big)^{1/2}\Big)$. By the law of iterated expectations, for any $y \in \mathcal{R}$ and $X \in \mathcal{S}$,

$$
\begin{aligned}
E\hat{g}(y, X) &= \frac{1}{h^q} E\Big[K(H^{-1}(X_i - X)) E\Big(G(\frac{y - y_i}{h_0}) \Big| X_i\Big)\Big] \\
&= \frac{1}{h^q} E\Big[K(H^{-1}(X_i - X)) F(y|X_i) + \frac{h_0^2}{2} \kappa_{G2} f'(y|X_i) + o(h_0^2)\Big] = F(y|X) f(X) + O(h^\nu + h_0^2).
\end{aligned}
$$

Thus, $\sup_{X\in S,y\in R}|\hat{g}(y,X)-g(y,X)| = O_p\Big(\big(\frac{\log n}{nh^q}\big)^{1/2}+h_0^2+h^\nu\Big)\equiv O_p(a^\dagger)$, where $g(y,X)=F(y|X)f(X)$. Theorem 6 in Hansen (2008) gives

$$\sup_{X\in S}|\hat{f}(X)-f(X)| = O_p(\big(\frac{\log n}{nh^q}\big)^{1/2}+h^\nu)\equiv O_p(a^*). \tag{B.5}$$

Therefore, uniformly in $y\in R$ and $X\in\mathcal{S}$, $\hat{F}(y|X)=\frac{\hat{g}(y,X)/f(X)}{\hat{f}(X)/f(X)}=\frac{F(y|X)+O_p(a^\dagger\delta^{-1})}{1+O_p(a^*\delta^{-1})}=F(y|X)+O_p(a^\dagger\delta^{-1})$.

Similarly for the conditional pdf, $\hat{f}(y|X)\equiv\frac{\hat{g}(y,X)}{\hat{f}(X)}$, where $\hat{g}(y,X)\equiv\frac{1}{nh_0h^q}\sum_{i=1}^n K(H^{-1}(X_i-X))g(\frac{y-y_i}{h_0})$, $\sup_{X\in S,y\in\mathcal{Y}}\Big|\hat{g}(y,X)-E\hat{g}(y,X)\Big| = O_p\Big(\big(\frac{\log n}{nh^qh_0}\big)^{1/2}\Big)$ and $bias(\hat{g}(y,X))=O(h^\nu+h_0^2)$. □

B.1.1 Proof of Proposition 2

Lemma B.3 (Uniform convergence rate)**.** *Suppose Assumptions A.FX, A.FY, and A.K hold. Let $p_X,p_Y\geq\nu$. $\mathcal{S}=\{X: f(X)\geq\delta\}$ is a compact subset of $\mathcal{X}$. Choose the order of the kernel ν, the bandwidths h,h_0, and the trimming parameter δ such that $\delta,h,h_0\to 0$ as $n\to 0$ and $(nh^q)^{1/2}(h_0^2+h^\nu)=o(1)$. Then $\sup_{X\in\mathcal{S}}|\hat{Q}_\tau(Y|X)-Q_\tau(Y|X)| = O_p\Big(\frac{1}{b_1\delta}\Big(\frac{\log n}{nh^q}\Big)^{1/2}\Big)$, for $\tau\in\mathcal{T}\equiv[\epsilon,1-\epsilon]$ for some $\epsilon>0$.*

Proof. Denote $Q=Q_\tau(Y|X)$. By Taylor's theorem, for any $X\in\mathcal{S}$, $F(\hat{Q}|X)=F(Q|X)+f(\bar{Q}|X)(\hat{Q}-Q)$, where $\bar{Q}$ is on the line segment between Q and $\hat{Q}$. Therefore,

$$|\hat{Q}-Q| = \Big|\frac{F(\hat{Q}|X)-\tau}{f(\bar{Q}|X)}\Big| \leq \frac{1}{b_1}\sup_{y\in\mathcal{Y}}|F(y|X)-\hat{F}(y|X)| \qquad w.p.a.1.$$

The inequality comes from (1) $\hat{F}(\hat{Q}|X)=\tau=F(Q|X)$ by construction; (2) It can be shown that $\hat{Q}=\hat{Q}_\tau(Y|X)\in\mathcal{Y}_0$ for n large enough, $\forall X\in\mathcal{S}$, $\tau\in\mathcal{T}$ with high probability, [2] and hence $\bar{Q}\in\mathcal{Y}_0$. And $f(y|X)\geq b_1>0$ $w.p.a.1$, $\forall y\in\mathcal{Y}$, $\forall X\in\mathcal{X}_0$ by Assumption A.FY. Then

$$\sup_{X\in\mathcal{S}}|\hat{Q}-Q| \leq \frac{1}{b_1}\sup_{X\in\mathcal{S}}\sup_{y\in\mathcal{Y}}|F(y|X)-\hat{F}(y|X)| = O_p\Big(\frac{1}{b_1\delta}\Big(\frac{\log n}{nh^q}\Big)^{1/2}\Big).$$

[2] Choose $\epsilon>0$ such that $\epsilon<min\{\tau,1-\tau\}$. By the weak uniform convergence of $\hat{F}_Y(y|X)$ in (**??**): i.e., $|\hat{F}_Y(\bar{y}|X)-F_Y(\bar{y}|X)|<\epsilon$ $w.p.a.1$ for any $X\in\mathcal{S}$. Note that $F_Y(\bar{y}|X)=1$ by definition. Since $\hat{F}_Y(y|X)$ is strictly increasing in y by construction, $\hat{F}_Y(\hat{Q}|X)=\tau<1-\epsilon<\hat{F}_Y(\bar{y}|X)$ implies $\hat{Q}<\bar{y}$ for n large enough with high probability. Similarly, I can conclude that $\hat{Q}\in\mathcal{Y}_0$ $w.p.a.1$.

□

Since $\hat{F}_Y(y|X)$ defined in (2.7) is smooth in y, expand $\hat{F}_Y(\hat{Q}|X)$ around Q by Taylor's theorem: $\hat{F}_Y(\hat{Q}|X) = \hat{F}_Y(Q|X) + \hat{f}_Y(\bar{Q}|X)(\hat{Q} - Q)$, where $\bar{Q}$ is on the line segment between Q and $\hat{Q}$. Define

$$\tau - \hat{F}_Y(Q|X) = \tau - \frac{\frac{1}{(n-1)|H|}\sum_{j=1}^n K_j G_j}{\frac{1}{(n-1)|H|}\sum_{j=1}^n K_j} \equiv A\frac{1}{f(X)} + A\Big(\frac{1}{\hat{f}(X)} - \frac{1}{f(X)}\Big),$$

where $A \equiv \frac{1}{(n-1)|H|}\sum_{j=1}^n K_j(\tau - G_j)$. So

$$\begin{aligned}\hat{Q} - Q &= \frac{\tau - \hat{F}_Y(Q|X)}{f_Y(Q|X)} + (\tau - \hat{F}_Y(Q|X))\Big(\frac{1}{\hat{f}_Y(\bar{Q}|X)} - \frac{1}{f_Y(Q|X)}\Big) \\ =&\frac{A}{f(X)f(Q|X)} + \frac{A}{f(Q|X)}\underbrace{\Big(\frac{1}{\hat{f}(X)} - \frac{1}{f(X)}\Big)}_{\equiv B} + A\underbrace{\frac{1}{f(X)}\Big(\frac{1}{\hat{f}(\bar{Q}|X)} - \frac{1}{f(Q|X)}\Big)}_{\equiv C} + f(X)ABC.\end{aligned} \tag{B.6}$$

By assuming $(nh^q)^{1/2}(h^\nu + h_0^2) = o(1)$, (B.5) and Lemma B.2, for any $\epsilon > 0$, there is a constant c_f such that with high probability, for n large enough, $|\hat{f}(X) - f(X)|\mathbf{1}_{\{f(X)\geq\delta\}} \leq c_f(n^{1-\epsilon}h^q)^{-1/2} \equiv c_{1n}$ and $|\hat{f}_Y(y|X) - f_Y(y|X)|\mathbf{1}_{\{f(X)\geq\delta\}} \leq c_f\delta^{-1}(n^{1-\epsilon}h_0h^q)^{-1/2} \equiv c_{2n}$. So for $f(X) \geq \delta$, $f(X) - c_{1n} \leq \hat{f}(X) \leq f(X) + c_{1n}$ which implies $\hat{f}(X) \geq \delta - c_{1n}$. Similarly, $\hat{f}_Y(y|X) \geq b_1 - c_{2n}$ for $f_Y(y|X) \geq b_1$ and $f(X) \geq \delta$. So with high probability

$$\sup_{X\in\mathcal{S}}|B| = \sup_{X\in\mathcal{S}}\frac{1}{f(X)\hat{f}(X)}|f(X) - \hat{f}(X)| \leq \frac{1}{\delta(\delta - c_{1n})}\sup_{X\in\mathcal{S}}|f(X) - \hat{f}(X)|.$$

By $\delta^2\sqrt{nh^q} \to \infty$, $\sup_{X_i\in\mathcal{S}}|B| = O_p\Big(\frac{1}{\delta(\delta-c_{1n})}\Big(\frac{\log n}{nh^q}\Big)^{1/2}\Big) = o_p(1)$.

For C, first,

$$\sup_{X\in\mathcal{S}}|f(\bar{Q}|X) - \hat{f}(\bar{Q}|X)| \leq \sup_{X\in\mathcal{S},y\in\mathcal{Y}}|f(y|X) - \hat{f}(y|X)| = O_p\Big(\frac{1}{\delta}\Big(\frac{\log n}{nh_0h^q}\Big)^{1/2}\Big)$$

by $\bar{Q} \in \mathcal{Y}_0$ *w.p.a.*1. Second, use Taylor's theorem with $\tilde{Q}$ on the line segment between $\bar{Q}$ and Q,

$$\sup_{X\in\mathcal{S}}|f(Q|X) - f(\bar{Q}|X)| = \sup_{X\in\mathcal{S}}|f'(\tilde{Q}|X)(\bar{Q} - Q)| \leq b_2\sup_{X\in\mathcal{S}}|\hat{Q} - Q| = O_p\Big(\frac{b_2}{b_1\delta}\Big(\frac{\log n}{nh^q}\Big)^{1/2}\Big),$$

by $f'(y|X) \leq b_2 < \infty$, $\forall y \in \mathcal{Y}$, $\forall X \in \mathcal{X}_0$ in Assumption A.FY and Lemma A.3. By triangle inequality,

$$\begin{aligned}
&\sup_{X_i \in \mathcal{S}} |C| = \sup_{X \in \mathcal{S}} \Big| \frac{1}{\hat{f}(\bar{Q}|X)} - \frac{1}{f(Q|X)} \Big| \frac{1}{f_i} \\
&\leq \frac{1}{b_1(b_1 - c_{2n})\delta} \Big(\sup_{X_i \in \mathcal{S}} |f(\bar{Q}|X) - \hat{f}(\bar{Q}|X)| + \sup_{X \in \mathcal{S}} |f(Q|X) - f(\bar{Q}|X)| \Big)
\end{aligned}$$

By Lemma B.2 and $\delta^{-2}(nh_0h^q)^{-1/2} = o(1)$, $\sup_{X_i \in \mathcal{S}} |C| = O_p\Big(\frac{1}{b_1(b_1 - c_{2n})} \frac{1}{\delta^2} \Big(\frac{\log n}{nh_0h^q} \Big)^{1/2} \Big) = o_p(1)$. Therefore, (B.6) implies $\hat{Q} - Q = \frac{A_i}{f(X)f(Q|X)} + R_n(X)$, where the remaining term [3]

$$\sup_{X \in \mathcal{S}} |R_n(X)| = \sup_{X \in \mathcal{S}} \Big| AC \Big| + s.o. = O_p\Big(\Big(\frac{\log n}{nh^q} \Big)^{1/2} \frac{1}{\delta^2} \Big(\frac{\log n}{nh_0h^q} \Big)^{1/2} \Big) = O_p\Big(\frac{1}{\delta^2} \Big(\frac{\log n}{nh^q\sqrt{h_0}} \Big) \Big).$$

B.2 Proof of Theorem 2.1

B.2.1 Sketch of the proof of the influence functions

Following Hardle and Stoker (1989), the asymptotic theorem will be first derived for $\tilde{\beta} = -\frac{2}{n} \sum_{i=1}^n \hat{Q}_i \nabla \hat{f}_i \mathbf{1}_{\{f(X) \geq \delta\}}$, trimmed based on the true density. Then I will show that $\sqrt{n}(\tilde{\beta} - \hat{\beta}) = o_p(1)$.

Following the idea of the proof in Powell et al. (1989) and Chaudhuri et al. (1997), $\tilde{\beta}$ can be decomposed as $-\frac{2}{n} \sum_{i=1}^n \hat{Q}_\tau(X_i) \nabla \hat{f}_i \mathbf{1}_{X_i} =$

$$\underbrace{-\frac{2}{n} \sum_{i=1}^n (\hat{Q}_i - Q_i) \nabla f_i \mathbf{1}_{X_i}}_{(I)} \underbrace{-\frac{2}{n} \sum_{i=1}^n Q_i \nabla \hat{f}_i \mathbf{1}_{X_i}}_{(II)} \underbrace{-\frac{2}{n} \sum_{i=1}^n \big(\hat{Q}_i - Q_i \big) \big(\nabla \hat{f}_i - \nabla f_i \big) \mathbf{1}_{X_i}}_{(III)},$$

where $\mathbf{1}_{X_i} \equiv \mathbf{1}_{\{X_i \in S\}} = \mathbf{1}_{\{f(X_i) \geq \delta\}}$. The asymptotic properties for (I) and (II) can be derived by the U-statistics theory. The third term (III) will be made smaller order term by choosing h, h_0, h_1, ν, and ν_1.

[3] By the proof in Appendix A.1, $A_i = \tau\Big(\hat{f}(X_i) - f(X_i) \Big) - \Big(\hat{g}(Q, X) - \tau f(X) \Big)$. So $\sup_{X \in \mathcal{S}} |A| = O_p\Big(\Big(\frac{\log n}{nh^q} \Big)^{1/2} \Big)$.

By the uniform linear representation in equation (2.10), the first term (I) becomes

$$-\frac{2}{n}\sum_{i=1}^{n}(\hat{Q}_i - Q_i)\nabla f_i \mathbf{1}_{X_i} = \underbrace{-\frac{2}{n}\frac{1}{(n-1)|H|}\sum_{i=1}^{n}\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\sum_{j\neq i}K_{ij}(\tau - G_{ij})\mathbf{1}_{X_i}}_{\equiv U_n}$$
$$-\frac{2}{n}\sum_{i=1}^{n}\nabla f_i \mathbf{1}_{X_i}\cdot R_n(X_i) + s.o.$$

where the second term $O_p\big(\frac{1}{n}\sum_{i=1}^{n}|\nabla f_i|\mathbf{1}_{X_i}\cdot|R_n(X_i)|\big) = O_p\big(\sup_{X_i\in\mathcal{S}}|R_n(X_i)|\big)$. The second-order U-statistic U_n can be rewritten as

$$U_n \equiv \frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j\neq i}\underbrace{\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)|H|}K_{ij}(G_{ij}-\tau)\mathbf{1}_{X_i}}_{\equiv\eta_{nij}} = \frac{1}{\binom{n}{2}}\sum_{1\leq i<j\leq n}(\eta_{nij}+\eta_{nji}) \equiv \frac{1}{\binom{n}{2}}\sum_{1\leq i<j\leq n}\xi_{nij}, \tag{B.7}$$

where $\xi_{nij} \equiv \eta_{nij} + \eta_{nji}$ is symmetric in i and j by construction. Note that η_{nij} varies with n through h, h_0, and δ.

The Hoeffding projection of U_n is $\hat{U}_n = \theta_n + \frac{2}{n}\sum_{i=1}^{n}[r_n(Z_i) - \theta_n]$, where $Z_i \equiv (Y_i, X_i')'$, $r_n(Z_i) = E[\xi_{nij}|Z_i] = E[\eta_{nij}|Z_i] + E[\eta_{nji}|Z_i]$ and $\theta_n = E[r_n(Z_i)] = E[\xi_{nij}] = 2E[\eta_{nij}] = E[U_n]$. Define $r_{1n}(Z_i) = E[\eta_{nij}|i]$ and $r_{2n}(Z_i) = E[\eta_{nji}|i]$. I abuse these generic notations for analysing the U-statistics for (II) and (III).

I show the asymptotic equivalence of U_n and its projection $\hat{U}_n$ in Lemma B.5 by Lemma 3.1 in PSS: if $E[||\xi_{nij}||^2] = o(n)$, then $\sqrt{n}(U_n - \hat{U}_n) = o_p(1)$. Then

$$\sqrt{n}((I) - \theta_n) = \sqrt{n}(U_n - \theta_n) + o_p(1) = \sqrt{n}(\hat{U}_n - \theta_n) + o_p(1) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}(r_n(Z_i) - \theta_n) + o_p(1)$$
$$= \frac{2}{\sqrt{n}}\sum_{i=1}^{n}(r(Z_i) - E[r(Z_i)]) + o_p(1), \tag{B.8}$$

where the first equality is controlled by $\sqrt{n}R_n(X_i) = o_p(1)$. The forth equality is the hard part described in the following. I am going to find $r(Z_i)$ independent of n such that $r_n(Z_i) = r(Z_i) + t_n(Z_i)$ and $\frac{2}{\sqrt{n}}\sum_{i=1}^{n}(r_n(Z_i) - E[r_n(Z_i)]) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}(r(Z_i) - E[r(Z_i)]) + o_p(1)$. That is, I will show $T_n := \frac{2}{\sqrt{n}}\sum_{i=1}^{n}(t_n(Z_i) - E[t_n(Z_i)]) = o_p(1)$ by showing $E[T_n^2]$ converging to zero.

Theorem B.1. *Let the bandwidths, h, h_0, and the order of the kernel ν satisfy Assumption A.B. Then $(I) = -\frac{2}{n}\sum_{i=1}^n(\hat{Q}_i - Q_i)\nabla f_i \mathbf{1}_{X_i} = \frac{2}{n}\sum_{i=1}^n r_I(Z_i) + o_p(n^{-1/2})$, where the influence function is*

$$2\, r_I(Z_i) \equiv \frac{2\nabla f_i}{f_Y(Q_i|X_i)}\Big(\mathbf{1}_{\{y_i \le Q_i\}} - \tau\Big). \tag{B.9}$$

The bias is made $o_p(n^{-1/2})$ where

$$\begin{aligned}
\theta_n &= -2E[(\hat{Q}_i - Q_i)\nabla f_i \mathbf{1}_{X_i}] \\
&= 2E\Big[\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\Big\{\frac{h_0^2}{2}\kappa_{G2} f_Y'(Q_i|X_i) f_i + h^\nu \kappa_\nu \sum_{l=1}^{\nu}\frac{1}{l!(\nu-l)!}\sum_{k=1}^{q}\partial_k^l F_Y(Q_i|X_i)\cdot\partial_k^{\nu-l} f(X_i)\Big\} \\
&\quad + h^\nu h_0^2 R_I(X_i)\Big] + o(h^\nu + h_0^2) \qquad\qquad (B.10)\\
&= O(h_0^2 + h^\nu),
\end{aligned}$$

where $\partial_k^0 f(X) = f(X)$ and $R_I(X_i) \equiv \frac{\kappa_{G2}}{2}\kappa_\nu \sum_{l=0}^{\nu}\frac{1}{l!(\nu-l)!}\sum_{k=1}^{q}\partial_k^l f_Y'(Q_i|X_i)\cdot\partial_k^{\nu-l} f(X_i)$ by Lemma B.4.

As noted by PSS, the influence function does not depend on the kernel, but the bias does.

Note that the second term ($I\!I$) is similar to the average derivate in mean regression in PSS, where Q_i in ($I\!I$) is replaced by y_i. So following PSS, I have the following expression for ($I\!I$):

$$-\frac{2}{n}\sum_{i=1}^n Q_i \nabla \hat{f}_i \mathbf{1}_{X_i} = -\frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j=1,j\neq i}^n Q_i \mathbf{1}_{X_i}\nabla K_{ij}\frac{1}{h^{q+1}} \tag{B.11}$$

$$= \frac{1}{\binom{n}{2}}\sum_{i=1}^{n-1}\sum_{j>i}\underbrace{\frac{-1}{h^{q+1}}\nabla K_{ij}(Q_i\mathbf{1}_{X_i} - Q_j\mathbf{1}_{X_j})}_{\equiv \xi_{nij}}. \tag{B.12}$$

The second equality is because of the symmetric kernel, $k'(-u) = -k'(u)$, i.e., $\nabla K_{ij} = -\nabla K_{ji}$. So ξ_{nij} is symmetric in i and j.

Theorem B.2. *Let the bandwidth, h_1, and the order of the kernel ν_1 satisfy Assumption A.B. Then as $n \to \infty$, $(I\!I) = -\frac{2}{n}\sum_{i=1}^n Q_i \nabla \hat{f}_i \mathbf{1}_{X_i} = \frac{2}{n}\sum_{i=1}^n r_{I\!I}(Z_i) + o_p(n^{-1/2})$, where the*

influence function for (Ⅱ) *is*

$$2\, r_{\text{Ⅱ}}(Z_i) = 2 f_i \nabla Q_i - 2E\big[\nabla(Q_i f_i)\big]. \tag{B.13}$$

The bias $E\Big[-\frac{2}{n}\sum_{i=1}^{n} Q_i \nabla \hat{f}_i \mathbf{1}_{X_i}\Big] + 2E[Q_i \nabla f_i] =$

$$-2E\Big[Q_i \frac{h_1^{\nu_1}}{\nu_1!}\kappa_{\nu_1} \sum_{k=1}^{q} \partial_k^{\nu_1} \nabla f_i\Big] + o(h_1^{\nu_1} + \delta) = O_p(h_1^{\nu_1}) \tag{B.14}$$

is made $o_p(n^{-1/2})$.

Finally, the third term (Ⅲ) is

$$\sup_{X_i \in S} \Big| -\frac{2}{n}\sum_{i=1}^{n} \big(\hat{Q}_i - Q_i\big)\big(\nabla \hat{f}_i - \nabla f_i\big)\Big| = O_p\Big(\Big[\frac{1}{\delta}\Big(\frac{\log n}{n h^q}\Big)^{1/2}\Big]\Big(\frac{\log n}{n h_1^{q+2}}\Big)^{1/2}\Big) \tag{B.15}$$

by the uniform convergence result in Theorem 6 in Hansen (2008). The third term (Ⅲ) will be made $o_p(n^{-1/2})$ by choosing h, h_0, h_1, δ, ν, and ν_1 according to Assumption A.B such that $\sqrt{n}$(Ⅲ) $= O_p(\delta^{-1}(n h^q h_1^{q+2})^{-1/2}) \to 0$.

Combining the results in Theorem B.1, B.2, and equation (B.15), I have the influence function for $\tilde{\beta}$. In Appendix, I show $\sqrt{n}(\tilde{\beta} - \hat{\beta}) = o_p(1)$. Therefore, I derive the influence function for $\hat{\beta}$:

$$\begin{aligned}
\hat{\beta} - \beta &= \frac{-2}{n}\sum_{i=1}^{n} \hat{Q}_\tau(X_i) \nabla \hat{f}(X_i) \mathbf{1}_{\{\hat{f}(X_i) \geq \delta\}} + 2E[Q_i \nabla f_i] \\
&= \frac{2}{n}\sum_{i=1}^{n} \frac{\nabla f_i}{f_Y(Q_i|X_i)}\Big(\mathbf{1}_{\{y_i \leq Q_i\}} - \tau\Big) + \frac{2}{n}\sum_{i=1}^{n} f_i \nabla Q_i - 2E[f_i \nabla Q_i] + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} r_\beta(Z_i) - \beta + o_p(n^{-1/2}),
\end{aligned}$$

where the influence function $r_\beta(Z_i) \equiv 2\Big(r_{\text{Ⅰ}}(Z_i) + r_{\text{Ⅱ}}(Z_i)\Big)$. By Linderberg-Levy central limit theorem, I derive the first part of Theorem 2.1. From (B.10) and (B.14), the bias $E[\hat{\beta} - \beta] =$

$$\begin{aligned}
&2E\Big[\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\Big\{\frac{h_0^2}{2}\kappa_{G2} f_Y'(Q_i|X_i) f_i + h^\nu \kappa_\nu \sum_{l=1}^{\nu} \frac{1}{l!(\nu - l)!}\sum_{k=1}^{q} \partial_k^l F_Y(Q_i|X_i) \cdot \partial_k^{\nu-l} f(X_i)\Big\}\Big] \\
&\quad - 2E\Big[Q_i \frac{h^{\nu_1}}{\nu_1!}\kappa_{\nu_1}\sum_{k=1}^{q} \partial_k^{\nu_1} \nabla f_i\Big] + o(h^\nu + h_0^2 + h_1^{\nu_1} + \delta) = O(h^\nu + h_0^2 + h_1^{\nu_1}).
\end{aligned}$$

B.2.2 Proof of Theorem B.1

I first apply Taylor's theorem to expand $Q_j \equiv Q_\tau(Y|X_j)$ around X_i, which will be used in the following proofs. Change variables: $X_j = X_i + HZ$ for $X_i, X_j \in \mathcal{X}_0$.

$$Q_j \equiv Q_\tau(Y|X_j) = Q_\tau(Y|X_i + HZ) = Q_i + h\sum_{k=1}^{q} \partial_k \bar{Q}_i Z_k \tag{B.16}$$

$$= Q_i + \Gamma Q_j \tag{B.17}$$

where $\bar{X}_i$ is on the line segment between X_i and $X_i + HZ$ and $\bar{Q}_i = Q_\tau(Y|\bar{X}_i)$ is well-defined since $\mathcal{X}$ is convex. Equation (B.16) is the Taylor's theorem to the first order, and equation (B.17) is to the νth order as defined in equation (B.2). Therefore, for any X_i,

$$\frac{1}{|H|}\int K_{ij}Q_j dX_j = \int K(Z)Q(X_i + HZ)dZ$$
$$= \int K(Z)(Q_i + \Gamma Q_j)dZ = Q_i + \frac{h^\nu}{\nu!}\sum_{k=1}^{q}\partial_k^\nu Q(X_i)\kappa_\nu + o(h^\nu),$$

by the dominated convergence theorem (DCT) and the uniform continuity of $\partial_{k_1}...\partial_{k_\nu}Q(X)$ for $k_1, ..., k_\nu \in \{1, ..., q\}$ by Assumption A.Q and $p_Q \geq \nu$.

Lemma B.4.

$$r_{1n}(Z_i) = E[\eta_{nij}|i] = \frac{\nabla f_i \mathbf{1}_{X_i}}{f_i f_Y(Q_i|X_i)} E\left[\frac{1}{|H|}K_{ij}(G_{ij} - \tau)\Big|i\right]$$
$$= \frac{\nabla f_i \mathbf{1}_{X_i}}{f_i f_Y(Q_i|X_i)}\Big\{\frac{h_0^2}{2}\kappa_{G2}f_Y'(Q_i|X_i)f_i + h^\nu\kappa_\nu\sum_{l=1}^{\nu}\frac{1}{l!(\nu-l)!}\sum_{k=1}^{q}\partial_k^l F_Y(Q_i|X_i)\cdot\partial_k^{\nu-l}f(X_i)\Big\}$$
$$+ h^\nu h_0^2 R_I(X_i) + o(h^\nu + h_0^2),$$

where $\partial_k^0 f(X) = f(X)$ *for notational simplicity, and* $R_I(X_i) \equiv \frac{\kappa_{G2}}{2}\kappa_\nu\sum_{l=0}^{\nu}\frac{1}{l!(\nu-l)!}$ $\sum_{k=1}^{q}\partial_k^l f_Y'(Q_i|X_i)\cdot\partial_k^{\nu-l}f(X_i)$. *Therefore, I have* $\theta_n = 2E[E[\eta_{nij}|i]] =$

$$2E\Bigg[\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\Big\{\frac{h_0^2}{2}\kappa_{G2}f_Y'(Q_i|X_i)f_i$$
$$+ h^\nu\kappa_\nu\sum_{l=1}^{\nu}\frac{1}{l!(\nu-l)!}\sum_{k=1}^{q}\partial_k^l F_Y(Q_i|X_i)\cdot\partial_k^{\nu-l}f(X_i)\Big\} + h^\nu h_0^2 R_I(X_i)\Bigg] + o(h^\nu + h_0^2).$$

Proof. By the law of iterated expectations,

$$E\left[\frac{1}{|H|}K_{ij}(G_{ij}-\tau)\Big|i\right]=\int_{\mathcal{X}}\frac{1}{|H|}K_{ij}\left[\int_{\mathcal{Y}}\left(G\Big(\frac{Q_i-y_j}{h_0}\Big)-\tau\right)f_Y(y_j|X_j)dy_j\right]f(X_j)dX_j.$$

The conditional expectation of Y_j given X_j is

$$\begin{aligned}
&E\big[G_{ij}-\tau\big|X_i,X_j\big]=\int_{\mathcal{Y}}\left(G\Big(\frac{Q_i-y_j}{h_0}\Big)-\tau\right)f_Y(y_j|X_j)dy_j\\
&=\Big(G\Big(\frac{Q_i-y_j}{h_0}\Big)-\tau\Big)F_Y(y_j|X_j)\Big|_{\mathcal{Y}}+\int_{\mathcal{Y}}\frac{1}{h_0}G'\Big(\frac{Q_i-y_j}{h_0}\Big)F_Y(y_j|X_j)dy_j\\
&=G\Big(\frac{Q_i-\bar{y}}{h_0}\Big)-\tau+\int_{\mathcal{Z}}G'(z)F_Y(Q_i-h_0z|X_j)dz\\
&=G\Big(\frac{Q_i-\bar{y}}{h_0}\Big)-\tau+\int_{-\infty}^{\infty}G'(z)\Big(F_Y(Q_i|X_j)-h_0zf_Y(Q_i|X_j)+\frac{h_0^2}{2}z^2f'_Y(\bar{Q}_i|X_j)\Big)dz\\
&\quad-\int_{-\infty}^{\infty}(1-\mathbf{1}_{\{z\in\mathcal{Z}\}})G'(z)\Big(F_Y(Q_i|X_j)-h_0zf_Y(Q_i|X_j)+\frac{h_0^2}{2}z^2f'_Y(\bar{Q}_i|X_j)\Big)dz \qquad \text{(B.18)}\\
&=-\tau+F_Y(Q_i|X_j)+\frac{h_0^2}{2}f'_Y(Q_i|X_j)\kappa_{G2}+o(h_0^2), \qquad \text{(B.19)}
\end{aligned}$$

where the compact support of Y is $\mathcal{Y}\equiv[\underline{y},\bar{y}]$ given X_j, the second equality is by integration by parts, the third equality is by change of variables $\frac{Q_i-y_j}{h_0}=z$ with support $\mathcal{Z}\equiv\left[\frac{Q_i-\bar{y}}{h_0},\frac{Q_i-\underline{y}}{h_0}\right]$. Since G' is chosen as a second-order kernel, the forth equality is the second-order expansion around Q_i by Taylor's theorem, where $\bar{Q}_i$ is on the line segment between Q_i and Q_i-h_0z. The first part of (B.19) comes from similar argument of dominated convergence theorem as in Lemma B.1. Using Assumption A.G, the first term of (B.18) is $-F_Y(Q_i|X_j)\Big[1-G\Big(\frac{Q_i-\underline{y}}{h_0}\Big)+G\Big(\frac{Q_i-\bar{y}}{h_0}\Big)\Big]=o(h_0^2)$, and the second term is $o(h_0^2)$ by integration by parts. The third term of (B.18) is $o(h_0^2)$ by the uniform continuity of $f'_Y(y|X)$ in y and dominated convergence theorem.

Again, by change of variables: $Z = H^{-1}(X_j - X_i)$, $X_j = X_i + HZ$.

$$E\left[\frac{1}{|H|}K_{ij}(G_{ij}-\tau)\Big|i\right] = \int_{\mathcal{X}} \frac{1}{|H|}K_{ij}\left[-\tau + F_Y(Q_i|X_j) + \frac{h_0^2}{2}f'_Y(Q_i|X_j)\kappa_{G2} + o(h_0^2)\right] f(X_j)dX_j$$

$$= \int_{\mathcal{Z}} K(Z)\left[-\tau + F_Y(Q_i|X_i+HZ) + \frac{h_0^2}{2}f'_Y(Q_i|X_i+HZ)\kappa_{G2} + o(h_0^2)\right] f(X_i+HZ)dZ$$

$$= \int_{\mathcal{Z}} K(Z)\Big[-\tau + F_Y(Q_i|X_i) + \Gamma F_Y(Q_i|\bar{X}_i) + \frac{h_0^2}{2}\kappa_{G2}\Big(f'_Y(Q_i|X_i)$$

$$+ \Gamma f'_Y(Q_i|\bar{X}_i)\Big) + o(h_0^2)\Big]\Big[f(X_i) + \Gamma f(\bar{X}_i)\Big]dZ$$

$$= \int K(Z)\Big[\Gamma F_Y(Q_i|\bar{X}_i)f_i + \Gamma F_Y(Q_i|\bar{X}_i)\cdot\Gamma f(\bar{X}_i) + \frac{h_0^2}{2}\kappa_{G2}f'_Y(Q_i|X_i)\cdot\Gamma f(\bar{X}_i)$$

$$+ \frac{h_0^2}{2}\kappa_{G2}\Gamma f'_Y(Q_i|\bar{X}_i)f_i + \frac{h_0^2}{2}\kappa_{G2}\Gamma f'_Y(Q_i|\bar{X}_i)\cdot\Gamma f(\bar{X}_i)\Big]dZ + \frac{h_0^2}{2}\kappa_{G2}f'_Y(Q_i|X_i)f_i + o(h_0^2)$$

$$= \frac{h_0^2}{2}\kappa_{G2}f'_Y(Q_i|X_i)f_i + \frac{h^\nu}{\nu!}\kappa_\nu\sum_{k=1}^q \partial_k^\nu F_Y(Q_i|X_i)f_i + \int K(Z)\Gamma F_Y(Q_i|\bar{X}_i)\cdot\Gamma f(\bar{X}_i)dZ$$

$$+ \frac{h_0^2}{2}\kappa_{G2}f'_Y(Q_i|X_i)\cdot\frac{h^\nu}{\nu!}\kappa_\nu\sum_{k=1}^q\partial_k^\nu f_i + \frac{h_0^2}{2}\kappa_{G2}f_i\cdot\frac{h^\nu}{\nu!}\kappa_\nu\sum_{k=1}^q\partial_k^\nu f'_Y(Q_i|X_i)$$

$$+ \frac{h_0^2}{2}\kappa_{G2}\int K(Z)\Gamma f'_Y(Q_i|\bar{X}_i)\cdot\Gamma f(\bar{X}_i)dZ + o(h^\nu + h_0^2)$$

$$= \frac{h_0^2}{2}\kappa_{G2}f'_Y(Q_i|X_i)f_i + \frac{h^\nu}{\nu!}\kappa_\nu\sum_{k=1}^q\partial_k^\nu F_Y(Q_i|X_i)f_i$$

$$+ h^\nu\kappa_\nu\sum_{l=1}^{\nu-1}\frac{1}{l!(\nu-l)!}\sum_{k=1}^q\partial_k^l F_Y(Q_i|X_i)\cdot\partial_k^{\nu-l}f(X_i) + h^\nu h_0^2 R_I(X_i) + o(h_0^2+h^\nu),$$

where $\bar{X}_i$ is on the line segment between X_i and $X_i + HZ$, using again DCT and the uniform continuity of $\partial_{k_1}...\partial_{k_\nu}f(X)$ and $\partial_{k_1}...\partial_{k_\nu}f'_Y(y|X)$ for $k_1,...,k_\nu \in \{1,...,q\}$ in X.

Since $\mathbf{1}_{\{X_i\notin S\}} = o(1)$ by $\delta \to 0$ and the moments exist by Assumption A.M, using DCT:

$$E\Bigg[\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\Big\{\frac{h_0^2}{2}\kappa_{G2}f'_Y(Q_i|X_i)f_i + h^\nu\kappa_\nu\sum_{l=1}^\nu\frac{1}{l!(\nu-l)!}\sum_{k=1}^q\partial_k^l F_Y(Q_i|X_i)\cdot\partial_k^{\nu-l}f(X_i)$$

$$+ h^\nu h_0^2 R_I(X_i)\Big\}\mathbf{1}_{\{X_i\notin S\}}\Bigg] = o(h^\nu + h_0^2).$$

□

Lemma B.5. *h is chosen so that $(nh^q)^{-1} = o(1)$ as $n\to\infty$ so that $E[||\xi_{nij}||^2] = o(n)$. Then by Lemma 3.1 in PSS, $\sqrt{n}(U_n - \hat{U}_n) = o_p(1)$.*

Proof. $E[||\xi_{nij}||^2] = E[||\eta_{nij} + \eta_{nji}||^2] = 2E[\eta'_{nij}\eta_{nij}] + 2E[\eta'_{nij}\eta_{nji}]$. For the first term,

$$E[\eta'_{nij}\eta_{nij}|Z_i] = \frac{\mathbf{1}_{X_i}\nabla f_i'\nabla f_i}{f_i^2 f_Y^2(Q_i|X_i)} E\Big[\frac{1}{|H|^2}K_{ij}^2 E\Big[\Big(G_{ij}-\tau\Big)^2\Big|Z_i, X_j\Big]\Big|Z_i\Big]$$

by the law of iterated expectations. Using integration by parts, change of variable ($y_j = Q_i - h_0 z$ and $\mathcal{Z} \equiv \Big[\frac{Q_i-\bar{y}}{h_0}, \frac{Q_i-\underline{y}}{h_0}\Big]$), and Taylor's theorem, I first calculate

$$\begin{aligned}
&E\Big[\Big(G_{ij}-\tau\Big)^2\Big|Z_i, X_j\Big] = \Big[G\Big(\frac{Q_i-y_j}{h_0}\Big)-\tau\Big]^2 F_Y(y_j|x_j)\Big|_{\mathcal{Y}} \\
&+\int_{\mathcal{Y}} 2\Big[G\Big(\frac{Q_i-y_j}{h_0}\Big)-\tau\Big]g\Big(\frac{Q_i-y_j}{h_0}\Big)\frac{1}{h_0}F_Y(y_j|X_j)dy_j \\
&= \tau^2 + 2\int_{\mathcal{Z}}\big[G(z)-\tau\big]g(z)F_Y(Q_i-h_0z|X_j)dz + o(h_0^2) \\
&= \tau^2 + 2\Big\{F_Y(Q_i|X_j)\int_{-\infty}^{\infty}\mathbf{1}_{\{z\in\mathcal{Z}\}}\big[G(z)-\frac{1}{2}+\frac{1}{2}-\tau\big]g(z)dz \\
&+\int_{-\infty}^{\infty}\mathbf{1}_{\{z\in\mathcal{Z}\}}\big[G(z)-\tau\big]g(z)\Big(-h_0zf_Y(Q_i|X_j)+\frac{h_0^2z^2}{2}f_Y'(\bar{Q}_i|X_j)\Big)dz\Big\} + o(h_0^2) \\
&= \tau^2 + (1-2\tau)F_Y(Q_i|X_j) - 2h_0f_Y(Q_i|X_j)\int_{-\infty}^{\infty}\mathbf{1}_{\{z\in\mathcal{Z}\}}\big[G(z)-\frac{1}{2}\big]g(z)zdz \\
&+h_0^2\int_{-\infty}^{\infty}\mathbf{1}_{\{z\in\mathcal{Z}\}}\big[G(z)-\frac{1}{2}+\frac{1}{2}-\tau\big]g(z)z^2dzf_Y'(Q_i|X_j) + o(h_0^2) \qquad \text{(B.20)} \\
&= \tau^2 + (1-2\tau)F_Y(Q_i|X_j) + C\cdot h_0\cdot f_Y(Q_i|X_j) + C\cdot h_0^2\cdot f_Y'(Q_i|X_j) + o(h_0^2),
\end{aligned}$$

where C denotes a generic constant. This is because $(G(z)-1/2)g(z)$ is an odd function for an symmetric kernel g. Again by Lemma B.1, $\bar{Q}_i$ is "between" Q_i and $Q_i - h_0 z$. In equation (B.20), $(1/2-\tau)\int z^t g(z)dz$ is a finite constant for the Gaussian kernel $g = \phi$. Then $\int_{-\infty}^{\infty} z^t g(z)[G(z)-1/2]dz$ is zero for even t, since $z^t g(z)[G(z)-1/2]$ is an odd function. When t is odd, $\int_{-\infty}^{\infty} z^t g(z)[G(z)-1/2]dz = 2\int_0^{\infty} z^t g(z)[G(z)-1/2]dz \le \int_0^{\infty} z^t\phi(z)dz < \infty$.

Also, $\mathbf{1}_{\{\mathbf{z}\notin\mathcal{Z}\}} = o(1)$, then use DCT.

$$
\begin{aligned}
&E\Big[\frac{1}{|H|^2}K_{ij}^2 E\Big[\Big(G_{ij}-\tau\Big)^2\Big|Z_i,X_j\Big]\Big|Z_i\Big] \\
&= E\Big[\frac{1}{|H|^2}K_{ij}^2\Big(\tau^2+(1-2\tau)F_Y(Q_i|X_j)+C\cdot h_0\cdot f_Y(Q_i|X_j)+C\cdot h_0^2\cdot f_Y'(Q_i|X_j)+o(h_0^2)\Big)\Big|Z_i\Big] \\
&= \frac{1}{|H|}\int_{\mathcal{Z}} K(Z)^2\Big[\tau^2+(1-2\tau)\Big(\tau+h\sum_{k=1}^q \partial_k F_Y(Q_i|\bar{X}_i)Z_k\Big) \\
&\quad + C\cdot h_0\cdot\Big(f_Y(Q_i|X_i)+h\sum_{k=1}^q \partial_k f_Y(Q_i|\bar{X}_i)Z_k\Big) \\
&\quad + C\cdot h_0^2\cdot\Big(f_Y'(Q_i|X_i)+h\sum_{k=1}^q \partial_k f_Y'(Q_i|\bar{X}_i)Z_k\Big)+o(h_0^2)\Big]\Big[f_i+h\sum_{k=1}^q \partial_k f(\bar{X}_i)Z_k\Big]dZ,
\end{aligned}
$$

where $\bar{X}_i$ is on the line segment between X_i and $X_j = X_i + HZ$, by Taylor's theorem to the first order. The trimming and change of variables won't affect the results, since K has bounded support. Therefore, $E[\eta_{nij}'\eta_{nij}] = O(\frac{1}{|H|}) = O(\frac{1}{h^q})$, given $\int k(z)^2 z dz = 0$, $\int k(z)^2 z^2 dz < \infty$, and Assumption A.M.

For the second term, by the law of iterated expectations and the independence of Y_i and Y_j,

$$
\begin{aligned}
E[\eta_{nij}'\eta_{nji}] &= E\left[\frac{\nabla f_i'\mathbf{1}_{X_i}}{f_i f_Y(Q_i|X_i)|H|}\frac{\nabla f_{X_j}\mathbf{1}_{X_j}}{f_{X_j} f_Y(Q_j|X_j)|H|}K_{ij}^2 E\Big[(G_{ij}-\tau)(G_{ji}-\tau)\Big|X_i,X_j\Big]\right] \\
&= E\left[\frac{\nabla f_i'\mathbf{1}_{X_i}}{f_i f_Y(Q_i|X_i)|H|}\int_{\mathcal{X}}\frac{\nabla f_{X_j}\mathbf{1}_{X_j}}{f_Y(Q_j|X_j)|H|}K_{ij}^2 E\Big[(G_{ij}-\tau)\Big|X_i,X_j\Big]E\Big[(G_{ji}-\tau)\Big|X_i,X_j\Big]dX_j\right].
\end{aligned}
$$

Since there is K_{ij}^2 instead of K_{ij} in $E[\eta_{nij}'\eta_{nji}]$, it suffices to expand up to the first order. By equation (B.19) in the proof of Lemma B.4,

$$
\begin{aligned}
E\big[G_{ij}-\tau\big|X_i,X_j\big] &= -\tau + F_Y(Q_i|X_j) + \frac{h_0^2}{2}f_Y'(Q_i|X_j)\kappa_{G2} + o(h_0^2) \\
&= h\sum_{k=1}^q \partial_k F_Y(Q_i|\bar{X}_i)Z_k + \frac{h_0^2}{2}\kappa_{G2}f_Y'(Q_i|X_i) + \frac{h_0^2}{2}\kappa_{G2}h\sum_{k=1}^q \partial_k f_Y'(Q_i|\bar{X}_i)Z_k + o(h_0^2),
\end{aligned}
$$

where $X_j = X_i + HZ$ and $\bar{X}_i$ is on the line segment between X_i and $X_i + HZ$. Similarly,

$$
\begin{aligned}
&E\big[G_{ji} - \tau\big|X_j, X_i\big] = -\tau + F_Y(Q_j|X_i) + \frac{h_0^2}{2} f_Y'(Q_j|X_i)\kappa_{G2} + o(h_0^2) \\
&= -\tau + F_Y(Q_i|X_i) + f_Y(\tilde{Q}_i|X_i)(Q_j - Q_i) + \frac{h_0^2}{2} f_Y'(Q_i|X_i)\kappa_{G2} + \frac{h_0^2}{2}\kappa_{G2} f_Y''(\tilde{Q}_i|X_i)(Q_j - Q_i) + o(h_0^2) \\
&= f_Y(\tilde{Q}_i|X_i) h \sum_{k=1}^{q} \partial_k \bar{Q}_i Z_k + \frac{h_0^2}{2} f_Y'(Q_i|X_i)\kappa_{G2} + \frac{h_0^2}{2}\kappa_{G2} f_Y''(\tilde{Q}_i|X_i) h \sum_{k=1}^{q} \partial_k \bar{Q}_i Z_k + o(h_0^2),
\end{aligned}
$$

where $\tilde{Q}_i$ is between Q_j and Q_i, the last equality is by equation (B.16), and $\bar{Q}_i = Q_\tau(Y|\bar{X}_i)$ with $\bar{X}_i$ on the line segment of X_i and $X_i + HZ$. So

$$
\begin{aligned}
&E[\eta_{nij}'\eta_{nji}] = E\left[\frac{\nabla f_i' \mathbf{1}_{X_i}}{f_i f_Y(Q_i|X_i)|H|} \int_S \frac{\nabla f_{X_j}}{f_Y(Q_j|X_j)|H|} K_{ij}^2 E\Big[G_{ij} - \tau\Big|X_i, X_j\Big] E\Big[G_{ji} - \tau\Big|X_i, X_j\Big] dX_j\right] \\
&= E\Big[\frac{\nabla f_i' \mathbf{1}_{X_i}}{f_i f_Y(Q_i|X_i)|H|}\Big(\int_{\mathcal{Z}} \Big(\frac{\nabla f_i}{f_Y(Q_i|X_i)}\Big) K(Z)^2 \\
&\Big(h\sum_{k=1}^{q} \partial_k F_Y(Q_i|\bar{X}_i) Z_k + \frac{h_0^2}{2}\kappa_{G2} f_Y'(Q_i|X_i) + \frac{h_0^2}{2}\kappa_{G2} h \sum_{k=1}^{q} \partial_k f_Y'(Q_i|\bar{X}_i) Z_k + o(h_0^2)\Big) \\
&\Big(h f_Y(\tilde{Q}_i|X_i) \sum_{k=1}^{q} \partial_k \bar{Q}_i Z_k + \frac{h_0^2}{2} f_Y'(Q_i|X_i)\kappa_{G2} + \frac{h_0^2}{2}\kappa_{G2} f_Y''(\tilde{Q}_i|X_i) h \sum_{k=1}^{q} \partial_k \bar{Q}_i Z_k + o(h_0^2)\Big) dZ + Rem\Big)\Big] \\
&= o\Big(\frac{1}{|H|}\Big).
\end{aligned}
\tag{B.21}
$$

For the remaining term Rem, the Lipschitz condition in Assumption A.L gives

$$
\begin{aligned}
\|Rem\| &\le \int \Big\|\frac{\nabla f(\cdot)}{f_Y(Q(\cdot)|\cdot)}(X_i + HZ) - \frac{\nabla f(\cdot)}{f_Y(Q(\cdot)|\cdot)}(X_i)\Big\| K(Z)^2 |(...)(...)| dZ \\
&\le h\ m(X) \int \|Z\| K(Z)^2 |(...)(...)| dZ.
\end{aligned}
$$

By the Schwartz Inequality for vectors: $E|b'c| \le E(\|b\| \cdot \|c\|)$ and the moment assumption in A.L, the Rem term is of smaller order in (B.21). So the second term $E[\eta_{nij}'\eta_{nji}]$ is of smaller order than the first term $E[\eta_{nij}'\eta_{nij}]$. Therefore, h will be chosen so that $O(\frac{1}{nh^q}) = o(1)$. □

From Lemma B.4, $r_{1n}(Z_i) = O(h_0^2 + h^\nu)$. So I guess $r_1(Z_i) = 0$ and verify $\frac{2}{\sqrt{n}} \sum_{i=1}^{n} (r_{1n}(Z_i) - Er_{1n}(Z_i)) \to 0$ which is implied by $E[||r_{1n}||^2] \to 0$. By Lemma B.4, $||r_{1n}|| \le \left\|\frac{\nabla f_i}{f_i f_Y(Q_i|X_i)}\right\| \cdot \left|E\left[\frac{1}{|H|} K_{ij}(G_{ij} - \tau)\Big|i\right]\right|$. The bounded moments are assumed in Assumption A.M, so I have the desired results.

By definition,

$$r_{2n}(Z_i) = E[\eta_{nji}|Z_i] = \int_S \frac{\nabla f_{X_j}}{f_Y(Q_j|X_j)} \frac{1}{|H|} K_{ji}\Big[G\Big(\frac{Q_j - y_i}{h_0}\Big) - \tau\Big] dX_j.$$

From this expression, I might guess $r_2(Z_i) = \frac{\nabla f_i}{f_Y(Q_i|X_i)}\Big[\mathbf{1}_{\{y_i \le Q_i\}} - \tau\Big]$. So $E[r_2(Z_i)] = 0$ and $t_{2n}(Z_i) = r_{2n}(Z_i) - r_2(Z_i)$. To verify the guess, I need to show $T_n \equiv \frac{2}{\sqrt{n}} \sum_{i=1}^n [t_{2n}(Z_i) - Et_{2n}(Z_i)] \xrightarrow{p} 0$ by showing $E[||t_{2n} t_{2n}'||] = E[||t_{2n}(Z_i)||^2] = o(1)$.

$$\begin{aligned}
&t_{2n}(Z_i) = r_{2n}(Z_i) - r_2(Z_i) \\
&= \int_S \frac{\nabla f_{X_j}}{f_Y(Q_j|X_j)} \frac{1}{|H|} K_{ji}\Big[G\Big(\frac{Q_j - y_i}{h_0}\Big) - \tau\Big] dX_j - \frac{\nabla f_i}{f_Y(Q_i|X_i)}\Big[\mathbf{1}_{\{y_i \le Q_i\}} - \tau\Big] \\
&= \int \Big[\frac{\nabla f(\cdot)}{f_Y(Q(\cdot)|\cdot)}(X_i + HZ)\Big(G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \tau\Big) - \frac{\nabla f_i}{f_Y(Q_i|X_i)}\Big(\mathbf{1}_{\{y_i \le Q_i\}} - \tau\Big)\Big] K(Z) dZ \qquad \text{(B.22)} \\
&\quad - \int (1 - \mathbf{1}_{X_j}) \frac{\nabla f(\cdot)}{f_Y(Q(\cdot)|\cdot)}(X_i + HZ) K(Z)\Big[G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \tau\Big] dZ. \qquad \text{(B.23)}
\end{aligned}$$

(B.23) is exactly zero for small enough h, since $K(Z)$ has bounded support. Denote $\frac{\nabla f(Z)}{f_Y(Q(Z)|Z)} \equiv A(Z)$ for notational ease. The first part of $t_{2n}(Z_i)$, (B.22), is

$$\begin{aligned}
&\int \Big[A(X_i + HZ)\Big(G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \tau\Big) - A(X_i)\Big(G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \tau\Big) \\
&+ A(X_i)\Big(G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \tau\Big) - A(X_i)\Big(\mathbf{1}_{\{y_i \le Q_i\}} - \tau\Big)\Big] K(Z) dZ \\
&= \int \Big(A(X_i + HZ) - A(X_i)\Big)\Big(G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \tau\Big) K(Z) dZ \qquad \text{(B.24)} \\
&+ \int A(X_i)\Big(G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \mathbf{1}_{\{y_i \le Q_i\}}\Big) K(Z) dZ. \qquad \text{(B.25)}
\end{aligned}$$

By Assumption A.L, $||(B.24)|| \le 2h\, m(X) \int ||Z|| \cdot |K(Z)| dZ$. So the second moment of (B.24) is bounded by $4h^2 E[m(X)^2]\big(\int ||Z|| \cdot |K(Z)| dZ\big)^2 = O(h^2) = o(1)$. The second moment of (B.25) is

$$\begin{aligned}
E\Big[A(X_i)A(X_i)' \int \int \Big(G\Big(\frac{Q(X_i + HZ) - y_i}{h_0}\Big) - \mathbf{1}_{\{y_i \le Q_i\}}\Big) \qquad \text{(B.26)} \\
\Big(G\Big(\frac{Q(X_i + HV) - y_i}{h_0}\Big) - \mathbf{1}_{\{y_i \le Q_i\}}\Big) K(Z) K(V) dZ dV\Big]. \qquad \text{(B.27)}
\end{aligned}$$

By the law of iterated expectations, I first calculate its conditional expectation of y_i:

$$\begin{aligned}
&\int \Big(G\Big(\frac{Q(X_i+HZ)-y_i}{h_0}\Big)-\mathbf{1}_{\{y_i\le Q_i\}}\Big)\Big(G\Big(\frac{Q(X_i+HV)-y_i}{h_0}\Big)-\mathbf{1}_{\{y_i\le Q_i\}}\Big)f_Y(y_i|X_i)dy_i\\
&=\int G\Big(\frac{Q_Z-y_i}{h_0}\Big)G\Big(\frac{Q_V-y_i}{h_0}\Big)f_Y(y_i|X_i)dy_i-\int \mathbf{1}_{\{y_i\le Q_i\}}G\Big(\frac{Q_V-y_i}{h_0}\Big)f_Y(y_i|X_i)dy_i\\
&-\int \mathbf{1}_{\{y_i\le Q_i\}}G\Big(\frac{Q_Z-y_i}{h_0}\Big)f_Y(y_i|X_i)dy_i+\tau, \qquad \text{(B.28)}
\end{aligned}$$

where $Q_Z=Q(X_i+HZ)$ and $Q_V=Q(X_i+HV)$ for notational ease. The second and third terms can be computed

$$\begin{aligned}
&\int_{\underline{y}}^{Q_i} G\Big(\frac{Q_V-y}{h_0}\Big)f(y|X_i)dy=G\Big(\frac{Q_V-y}{h_0}\Big)F_Y(y|X_i)\Big|_{\underline{y}}^{Q_i}+\int_{\underline{y}}^{Q_i}\frac{1}{h_0}g\Big(\frac{Q_V-y}{h_0}\Big)F(y|X_i)dy\\
&=\ G\Big(\frac{Q_V-Q_i}{h_0}\Big)\tau+\int_{(\underline{y}-Q_V)/h_0}^{(Q_i-Q_V)/h_0} g(t)\Big(F_Y(Q_V|X_i)+h_0tf_Y(\bar{Q}_V|X_i)\Big)dt\\
&=\ G\Big(\frac{Q_V-Q_i}{h_0}\Big)\tau+F_Y(Q_V|X_i)\Big[G\Big(\frac{Q_i-Q_V}{h_0}\Big)-G\Big(\frac{\underline{y}-Q_V}{h_0}\Big)\Big]+h_0\int_{(\underline{y}-Q_V)/h_0}^{(Q_i-Q_V)/h_0} g(t)tf_Y(\bar{Q}_V|X_i)dt\\
&=\ \tau+(Q_V-Q_i)f_Y(\tilde{Q}_V|X_i)G\Big(\frac{Q_i-Q_V}{h_0}\Big)+h_0\int_{(\underline{y}-Q_V)/h_0}^{(Q_i-Q_V)/h_0} g(t)tf_Y(\bar{Q}_V|X_i)dt+o(h_0^2),
\end{aligned}$$

where $\bar{Q}_V$ is between Q_V and Q_V+h_0t. The last equality is because $F_Y(Q_V|X)=F_Y(Q(X)+\Gamma Q_V|X)=\tau+\Gamma Q_Vf_Y(\tilde{Q}_V|X)$, where $\tilde{Q}_V$ is between $Q(X)$ and $Q(X+HV)$. By Assumption A.G, $G\Big(\frac{\underline{y}-Q_V}{h_0}\Big)=o(h_0^2)$.

Therefore, the second term of (B.28) contributes in (B.27) by

$$\begin{aligned}
&\int\Big[\int \mathbf{1}_{\{y_i\le Q_i\}}G\Big(\frac{Q_V-y_i}{h_0}\Big)f_Y(y_i|X_i)dy_i\Big]K(V)dV\\
&=\ \tau+\int\Big[(Q_V-Q_i)f_Y(\tilde{Q}_V|X_i)G\Big(\frac{Q_i-Q_V}{h_0}\Big)\\
&+h_0\int_{(\underline{y}-Q_V)/h_0}^{(Q_i-Q_V)/h_0} g(t)tf_Y(\bar{Q}_V|X_i)dt\Big]K(V)dV+o(h_0^2)\\
&=\ \tau+h\ C\cdot f_Y(Q(X_i)|X_i)\sum_{k=1}^{q}\partial_kQ(X_i)+C\cdot h_0f_Y(Q(X_i)|X_i)+s.o.
\end{aligned}$$

where C denotes a generic constant and the $s.o.$ is by the DCT.

The first term of (B.28)

$$\int G\Big(\frac{Q_Z - y_i}{h_0}\Big)G\Big(\frac{Q_V - y_i}{h_0}\Big)f_Y(y_i|X_i)dy_i$$
$$= G\Big(\frac{Q_Z - y_i}{h_0}\Big)G\Big(\frac{Q_V - y_i}{h_0}\Big)F_Y(y_i|X_i)\Big|_{\underline{y}}^{\bar{y}} - \int \frac{\partial}{\partial y_i}\Big[G\Big(\frac{Q_Z - y_i}{h_0}\Big)G\Big(\frac{Q_V - y_i}{h_0}\Big)\Big]F_Y(y_i|X_i)dy_i. \tag{B.29}$$

The first term of (B.29) is $o(h_0^4)$ by Assumption A.G. Let $\Delta \equiv Q_V - Q_Z$. The second term of (B.29) is

$$\int_{(Q_V-\bar{y})/h_0}^{(Q_V-\underline{y})/h_0} G(t)\Big[g\Big(\frac{\Delta}{h_0} - t\Big)F_Y(Q_V - h_0 t|X)\Big]dt + \int_{(Q_Z-\bar{y})/h_0}^{(Q_Z-\underline{y})/h_0} G(t)\Big[g\Big(\frac{\Delta}{h_0} + t\Big)F_Y(Q_Z - h_0 t|X)\Big]dt.$$

Observe that integration by parts gives

$$\int_{-\infty}^{\infty} g\Big(\frac{\Delta}{h_0} + t\Big)G(t)dt = G\Big(\frac{\Delta}{h_0} + t\Big)G(t)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} G\Big(\frac{\Delta}{h_0} + t\Big)g(t)dt$$
$$= 1 - \int_{-\infty}^{\infty} G(t)g\Big(t - \frac{\Delta}{h_0}\Big)dt.$$

Thus, $\int_{-\infty}^{\infty} G(t)\Big[g\Big(t + \frac{\Delta}{h_0}\Big) + g\Big(t - \frac{\Delta}{h_0}\Big)\Big]dt = 1$. So $\int_{-\infty}^{\infty} G(t)g\Big(t + \frac{\Delta}{h_0}\Big)dt \in [0, 1]$. Note that $f_Y(y|X_i)$ is bounded and $\int_{-\infty}^{\infty}\Big[g\Big(\frac{\Delta}{h_0} + t\Big) + g\Big(-\frac{\Delta}{h_0} + t\Big)\Big]G(t)tdt < \infty$. [4] $F_Y(Q_V - h_0 t|X) = F_Y(Q(X) + \Gamma Q_V - h_0 t|X) = \tau + (\Gamma Q_V - h_0 t)f_Y(\bar{Q}_V|X)$, where $\bar{Q}_V$ is between $Q(X)$ and $Q(X) + \Gamma Q_V - h_0 t$ controlled by h and h_0. Therefore, the first term of (B.28) contributes in the second moment of (B.25) by

$$\int\int\Big[\int G\Big(\frac{Q_Z - y_i}{h_0}\Big)G\Big(\frac{Q_V - y_i}{h_0}\Big)f_Y(y_i|X_i)dy_i\Big]K(Z)K(V)dZdV$$
$$= \tau + \int\Big[\int G(t)g(\Delta/h_0 - t)(\Gamma Q_V - h_0 t)f_Y(\bar{Q}_V|X)dt\Big]K(V)dV$$
$$+ \int\Big[\int G(t)g(\Delta/h_0 + t)(\Gamma Q_Z - h_0 t)f_Y(\bar{Q}_Z|X)dt\Big]K(Z)dZ$$
$$= \tau + C \cdot h^\nu f_Y(Q(X_i)|X_i)\sum_{k=1}^{q}\partial_k^\nu Q(X_i) + C \cdot h_0 f_Y(Q(X_i)|X_i) + s.o.,$$

where C denotes a generic constant and the $s.o.$ is by the DCT.

[4] Observe that $\Big[g\Big(\frac{\Delta}{h_0} + t\Big) + g\Big(-\frac{\Delta}{h_0} + t\Big)\Big](G(t) - 1/2)t$ is an even function and $\big|G(t) - 1/2\big| < 1/2$. Use the fact that $2\int_0^{\infty}\Big[g\Big(\frac{\Delta}{h_0} + t\Big) + g\Big(-\frac{\Delta}{h_0} + t\Big)\Big]tdt < \infty$.

Since $E[A(X)A(X)'f_Y(Q(X)|X)]$ is assumed to exist, the second moment of (B.25) is $O(h^\nu + h_0) = o(1)$.

$E\|(B.22)\|^2 \le E\|(B.24)\|^2 + E\|(B.25)\|^2 + 2E\big(\|(B.24)\| \cdot \|(B.25)\|\big)$. The last term is bounded by $4h \int \|Z\| \cdot |K(Z)| dZ E[m(X)\|A(X)\|] = O(h) = o(1)$. Therefore, the second moment of t_{2n} is $o(1)$.

B.2.3 Proof of Theorem B.2

In this section, h and ν should be h_1 and ν_1 for estimating ∇f. I omit the subscript 1 for notational ease.

Lemma B.6. *h is chosen so that $(nh^q)^{-1} = o(1)$ as $n \to \infty$, then $E[\|\xi_{nij}\|^2] = o(n)$.*

Proof. First note that by change of variable $X_j = X_i + HZ$,

$$\begin{aligned} Q_i \mathbf{1}_{X_i} - Q_j \mathbf{1}_{X_j} &= \Big(Q_i - Q(X_i + HZ)\mathbf{1}_{\{X_i + HZ \in S\}}\Big)\mathbf{1}_{X_i} \\ &= \Big(- h \sum_{k=1}^{q} \partial_k Q(\bar{X}_i) Z_k + Q(X_i + HZ)\mathbf{1}_{\{X_i + HZ \notin S\}}\Big)\mathbf{1}_{X_i}, \end{aligned}$$

by (B.16). So

$$\begin{aligned} &E[\|\xi_{nij}\|^2] = \int\int \frac{1}{h^{2q+2}} \|\nabla K_{ij}\|^2 (Q_i \mathbf{1}_{X_i} - Q_j \mathbf{1}_{X_j})^2 f(X_i) f(X_j) dX_i dX_j \\ &= \frac{1}{h^{q+2}} \int\int \|\nabla K(Z)\|^2 \Big(- h \sum_{k=1}^{q} \partial_k Q(\bar{X}_i) Z_k + Q(X_i + HZ)\mathbf{1}_{\{X_i + HZ \notin S\}}\Big)^2 \\ &\quad f(X_i + HZ) dZ \; f(X_i)\mathbf{1}_{X_i} dX_i \\ &= \frac{1}{h^q} \int\int \|\nabla K(Z)\|^2 \Big(\sum_{k=1}^{q} \partial_k Q(\bar{X}_i) Z_k \Big)^2 \Big(f(X_i) + h \sum_{k=1}^{q} \partial_k f(\bar{X}_i) Z_k \Big) dZ f(X_i)\mathbf{1}_{X_i} dX_i = O(h^{-q}) \end{aligned}$$

where $\bar{X}_i$ is on the line segment of X_i and $X_j = X_i + HZ$ by equation (B.16). The third equality is because the integration over $Q(X_i + HZ)\mathbf{1}_{\{X_i + HZ \notin S\}}$ is zero for small enough h by the bounded-support $K(Z)$. □ Note that h converges faster than that in PSS where $h^{-(q+2)} = o(n)$. This is because here I do not have the randomness from Y in $(I\!I)$. Then by Lemma 3.1 in PSS, $\sqrt{n}[(I\!I) - E(I\!I)] = \sqrt{n}\frac{2}{n} \sum_{i=1}^{n} [r_n(Z_i) - \theta_n] + o_p(1)$, where

$\theta_n = E[r_n(Z_i)] = E\xi_{nij} = E(\text{I\!I})$ and

$$r_n(Z_i) = E[\xi_{nij}|Z_i] = \int_{\mathcal{X}} \frac{-1}{h^{q+1}} \nabla K_{ij} f(X_j) dX_j Q_i \mathbf{1}_{X_i} + \int_{\mathcal{X}} \frac{1}{h^{q+1}} \nabla K_{ij} Q_j f(X_j) \mathbf{1}_{X_j} dX_j$$
$$= -Q_i \mathbf{1}_{X_i} \Big(\frac{-1}{h^q} K_{ij} f(X_j) \Big|_{\mathcal{X}} + \frac{1}{h^q} \int_{\mathcal{X}} K_{ij} \nabla f(X_j) dX_j \Big) + \frac{1}{h^q} \Big(- K_{ij} f(X_j) Q_j \Big|_S + \int_S K_{ij} \nabla (f_j Q_j) dX_j \Big)$$
$$= -Q_i \mathbf{1}_{X_i} \int K(Z) \nabla f(X_i + HZ) dZ + \int K(Z) \nabla (fQ)(X_i + HZ) \mathbf{1}_{\{X_i + HZ \in S\}} dZ.$$

The third equality is by integration by parts and the forth equality is change of variable: $X_j = X_i + HZ$. Then I can guess $r(Z_i) = -Q_i \nabla f_i + \nabla(Q_i f_i) = f_i \nabla Q_i$ whose mean $E[r(Z_i)] = E[f_i \nabla Q_i]$.

To verify the guess, I need to show that $T_n \equiv \frac{2}{\sqrt{n}} \sum_{i=1}^n [t_n(Z_i) - E t_n(Z_i)] \xrightarrow{p} 0$ where

$$t_n(Z_i) = r_n(Z_i) - r(Z_i)$$
$$= \int K(Z) \Big[- Q_i \nabla f(X_i + HZ) + Q_i \nabla f(X_i) + \nabla(fQ)(X_i + HZ) - \nabla(fQ)(X_i) \Big] dZ \tag{B.30}$$
$$- \frac{1}{h^{q+1}} \int \nabla K_{ij} f_j \Big[- Q_i (1 - \mathbf{1}_{X_i}) + Q_j (1 - \mathbf{1}_{\{X_i + HZ \in S\}}) \Big] dX_j \tag{B.31}$$

By Assumption A.L,

$$\|(B.30)\| \leq h(1 + |Q_i|) |m(X_i)| \int |K(Z)| \cdot \|Z\| dZ. \tag{B.32}$$

Therefore, the second moment of (B.30) is bounded above by $h^2 E\big((1 + |Q(X_i)|)^2 |m(X_i)|^2\big)$ $[\int |K(Z)| \cdot \|Z\| dZ]^2 = O(h^2) = o(1)$. The first term in (B.31) is $Q_i(1 - \mathbf{1}_{X_i}) \int K(Z) \nabla f(X_i + HZ) dZ = Q_i (1 - \mathbf{1}_{X_i}) \Big(\nabla f(X_i) + C \cdot h^\nu \cdot \sum_{k=1}^q \partial_k^\nu \nabla f_i) \Big)$ with some constant C and by DCT. So the second moment vanishes by $\delta \to 0$ and the existence of the second moments. The second term in (B.31) is $\frac{1}{h^q} K_{ij} f_j Q_j (1 - \mathbf{1}_{X_j}) |_{\mathcal{X}} - \int K(Z) \nabla(fQ)(X_i + HZ)(1 - \mathbf{1}_Z) dZ = 0$ for small enough H, because kernel K has bounded support. Then $E[t_n(Z_i) t_n(Z_i)'] \to 0$. Therefore, $\sqrt{n}[(\text{I\!I}) - E(\text{I\!I})] = \sqrt{n} \frac{2}{n} \sum_{i=1}^n [r_n(Z_i) - E(\text{I\!I})] + o_p(1) = \sqrt{n} \frac{2}{n} \sum_{i=1}^n [r(Z_i) - E(r(Z_i))] + o_p(1) =$ $\sqrt{n} \frac{2}{n} \sum_{i=1}^n [f_i \nabla Q_i - E(f_i \nabla Q_i)] + o_p(1)$.

The bias

$$E\Big[-\frac{2}{n}\sum_{i=1}^{n} Q_i\nabla\hat{f}_i\mathbf{1}_{X_i}\Big] + 2E[Q_i\nabla f_i] = -2E\Big[Q_i\nabla\hat{f}_i\mathbf{1}_{X_i}\Big] + 2E\Big[Q_i\nabla f_i\mathbf{1}_{X_i}\Big]$$
$$+ 2E\Big[Q_i\nabla f_i(1-\mathbf{1}_{X_i})\Big] = -2E\Big[Q_iE(\nabla\hat{f}_i - \nabla f_i|X_i)\mathbf{1}_{X_i}\Big] + o(\delta)$$
$$= -2E\Big[Q_i\frac{h^\nu}{\nu!}\kappa_\nu\sum_{k=1}^{q}\partial_k^\nu\nabla f(X_i)\Big] + 2E\Big[Q_i\frac{h^\nu}{\nu!}\kappa_\nu\sum_{k=1}^{q}\partial_k^\nu\nabla f(X_i)(1-\mathbf{1}_{X_i})\Big] + o(h^\nu+\delta) = O(h^\nu),$$

by Lemma B.1. Therefore,

$$\sqrt{n}\Big[-\frac{2}{n}\sum_{i=1}^{n} Q_i\nabla\hat{f}_i\mathbf{1}_{X_i} + 2E[Q_i\nabla f_i] + O(h^\nu)\Big] = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\Big[f_i\nabla Q_i - E(f_i\nabla Q_i)\Big] + o_p(1),$$

by choosing h and ν such that $\sqrt{n}h^\nu \to 0$.

B.2.4 Scaled AQD

The scaled AQD $\beta(\tau)^* = \beta(\tau)/\alpha$, where the scaling parameter $\alpha \equiv E[f(X)]$ can be estimated by $\hat{\alpha} = n^{-1}\sum_{i=1}^{n}\hat{f}(X_i)$. Observe that $\hat{\alpha}$ can be rearranged as an U-statistic. So I follow the similar steps in previous proofs to derive the influence function for $\hat{\alpha}$. Define $\xi_{nij} \equiv \frac{1}{|H_1|}K(H^{-1}(X_i - X_j))$, which is symmetric in i and j. Then it can be shown that $E[||\xi_{nij}||^2] = o(n)$ if $(nh^q)^{-1} = o(1)$. By Lemma B.1, $r_n(Z_i) = E[\xi_{nij}|Z_i] = r(Z_i) + t_n(Z_i)$, where $r(Z_i) = f(X_i)$ and $t_n(Z_i) = o(1)$. Therefore, I derive the second part of Theorem 2.1 $\hat{\alpha} - \alpha = \frac{1}{n}\sum_{i=1}^{n} r_\alpha(Z_i) + o_p(n^{-1/2})$, where the influence function $r_\alpha(Z_i) = 2[f(X_i) - Ef(X)]$. [5]

My interest is

$$\begin{aligned}\sqrt{n}\Big(\hat{\beta}^* - \beta^*\Big) &= \frac{\sqrt{n}}{\hat{\alpha}\alpha}\Big(\hat{\beta}\alpha - \beta\hat{\alpha}\Big)\\ &= \frac{\sqrt{n}}{\hat{\alpha}\alpha}\Big((\hat{\beta}-\beta)\alpha - \beta(\hat{\alpha}-\alpha)\Big)\\ &= \frac{\sqrt{n}}{\hat{\alpha}}\frac{1}{n}\sum_{i=1}^{n}\Big(r_\beta(Z_i) - r_\alpha(Z_i)\frac{\beta}{\alpha}\Big) + o_p(1) \equiv \frac{A_n}{\hat{\alpha}} + o_p(1).\end{aligned}$$

[5] This result has been shown in Powell and Stoker (1996).

The numerator is

$$A_n \equiv \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\Big(r_\beta(Z_i) - r_\alpha(Z_i)\beta^*\Big) \longrightarrow \mathcal{N}(0, V),$$

where V is the covariance matrix of the influence function $r_\beta(Z_i) - r_\alpha(Z_i)\beta^*$. Therefore, by Slutsky's theorem, I have the third part of Theorem 2.1.

B.2.5 Choice of nonparametric tuning parameters

Let the positive sequences $h \propto n^{-a}$, $h_1 \propto n^{-c}$, $h_0 \propto n^{-d}$, and $\delta \propto n^{-b}$, for some constants, $a, b, c, d > 0$. Choose a, b, c, d, ν, and ν_1 such that $\sqrt{n}R_n = O_p\Big(\Big(\delta^2 n^{1/2} h^q \sqrt{h_0}\Big)^{-1}\Big) = o(1)$, for the bias: $\sqrt{n}(h_0^2 + h^\nu) = o(1)$, $\sqrt{n}h_1^{\nu_1} = o(1)$, $\sqrt{n}(\text{III}) = \delta^{-1}(nh^q h_1^{q+2})^{-1/2} = o(1)$, for the U-statistics: $(nh^q)^{-1} = o(1)$, $(nh_1^q)^{-1} = o(1)$, and for the bias-reducing kernel: $(nh^q)^{1/2}(h^\nu + h_0^2) = o(1)$, $(nh_1^{q+2})^{1/2}h_1^{\nu_1}$, as $n \to \infty$.

Let's start with

$$bias(\sqrt{n}(I)) = O\big(\sqrt{n}(h_0^2 + h^\nu)\big) = o(1) \Longleftrightarrow d > \frac{1}{4}, a > \frac{1}{2\nu}, \tag{B.33}$$

$$bias(\sqrt{n}(\text{II})) = O\big(\sqrt{n}h_1^{\nu_1}\big) = o(1) \Longleftrightarrow c > \frac{1}{2\nu_1}, \tag{B.34}$$

$$\sqrt{n}R_n = o(1) \Longleftrightarrow b < \frac{1}{4}(1 - 2aq - d), \tag{B.35}$$

$$\sqrt{n}(\text{III}) = \delta^{-1}(nh^q h_1^{q+2})^{-1/2} = o(1) \Longleftrightarrow b < \frac{1}{2}(1 - aq - c(q+2)). \tag{B.36}$$

The upper bounds for b must be positive in (B.35) and (B.36), so

$$d < 1 - 2aq, \tag{B.37}$$

$$c < \frac{1 - aq}{q + 2}. \tag{B.38}$$

The upper bound must be larger than the lower bound, so for c, (B.34) and (B.38) give $\nu_1 > \frac{q+2}{2-2aq}$. For d, (B.33) and (B.37) give $a < \frac{3}{8q}$. Then, for a, together with (B.33), $\nu > \frac{4q}{3}$.

For the U-statistics, $(nh^q)^{-1} = o(1)$ and $(nh_1^q)^{-1} = o(1)$ are implied by (B.36). The condition for Proposition 2 are implied by controlling the bias in (B.35) and (B.37), and the remaining term in the Bahadur representation of $\hat{Q}_\tau(Y|X)$.

Therefore, choose $\nu > \frac{4q}{3}$, $a \in (\frac{1}{2\nu}, \frac{3}{8q})$, $\nu_1 > \frac{q+2}{2-2aq}$, $c \in (\frac{1}{2\nu_1}, \frac{1-aq}{q+2})$, $d \in (\frac{1}{4}, 1-2aq)$, and $b < min\{\frac{1}{4}(1-2aq-d), \frac{1}{2}(1-aq-c(q+2))\}$. Choosing $\nu_1 > \frac{4}{5}(q+2)$ is sufficient.

B.2.6 Trimming

Following Lavergne and Vuong (1996), choose c_n such that $c_n/\delta = o(1)$ and $c_n^{-1} \sup_| \hat{f}(X) - f(X)|\mathbf{1}_{\{f(X)\geq\delta\}} = o_p(1)$, which exists in view of (B.5). Since $c_n/\delta = o(1)$, I can work with the bound $\delta + c_n$, instead of δ.

$$\begin{aligned}\sqrt{n}(\tilde{\beta} - \hat{\beta}) &= \frac{-2}{\sqrt{n}} \sum_{i=1}^{n} \hat{Q}_i \nabla \hat{f}_i \Big(\mathbf{1}_{\{f(X_i)\geq\delta+c_n\}} - \mathbf{1}_{\{\hat{f}(X_i)\geq\delta\}} \Big) \\ &= \frac{-2}{\sqrt{n}} \sum_{i=1}^{n} \hat{Q}_i \nabla \hat{f}_i \Big(\mathbf{1}_{\{f(X_i)\geq\delta+c_n, \hat{f}(X_i)<\delta\}} - \mathbf{1}_{\{f(X_i)<\delta+c_n, \hat{f}(X_i)\geq\delta\}} \Big).\end{aligned}$$

For any $i \in \{1, 2, ..., n\}$, the event $\Big\{ f(X_i) \geq \delta + c_n, \hat{f}(X_i) < \delta \Big\} \subseteq \Big\{ |\hat{f}(X_i) - f(X_i)| > c_n, f(X_i) \geq \delta + c_n \Big\} \subseteq \Big\{$
$sup_i |\hat{f}(X_i) - f(X_i)| \mathbf{1}_{\{f(X_i)\geq\delta\}} > c_n \Big\}$ has asymptotic probability zero. Hence,
$sup_i \mathbf{1}_{\{f(X_i)\geq\delta+c_n, \hat{f}(X_i)<\delta\}} = 0$ with probability approaching one. So I need to consider the second term only. Define $I_i \equiv \mathbf{1}_{\{f(X_i)<\delta+c_n, \hat{f}(X_i)\geq\delta\}}$, for notational ease.

$$\begin{aligned}\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{Q}_i \nabla \hat{f}_i I_i &= n^{-1/2} \sum_{i=1}^{n} (\hat{Q}_i - Q_i) \nabla f_i I_i + n^{-1/2} \sum_{i=1}^{n} Q_i (\nabla \hat{f}_i - \nabla f_i) I_i \\ &\quad + n^{-1/2} \sum_{i=1}^{n} (\hat{Q}_i - Q_i)(\nabla \hat{f}_i - \nabla f_i) I_i + n^{-1/2} \sum_{i=1}^{n} Q_i \nabla f_i I_i. \qquad \text{(B.39)}\end{aligned}$$

For the last term in (B.39), $\Big\| n^{-1/2} \sum_{i=1}^{n} Q_i \nabla f_i I_i \Big\| \leq n^{-1/2} \sum_{i=1}^{n} \Big\| Q_i \nabla f_i \Big\| \mathbf{1}_{\{f_i<\delta+c_n\}}$. Therefore,

$$\begin{aligned}&E \Big\| n^{-1/2} \sum_{i=1}^{n} Q_i \nabla f_i I_i \Big\|^2 \leq n^{-1} E \left[\sum_{i=1}^{n} \Big\| Q_i \nabla f_i \Big\| \mathbf{1}_{\{f_i<\delta+c_n\}} \right]^2 \\ &= E \left[\Big\| Q_i \nabla f_i \Big\|^2 \mathbf{1}_{\{f_i<\delta+c_n\}} \right] + (n-1) \Big(E \Big[\Big\| Q_i \nabla f_i \Big\| \mathbf{1}_{\{f_i<\delta+c_n\}} \Big] \Big)^2 = o(1),\end{aligned}$$

where the first term is $o(1)$ by Lebesgue dominated convergence theorem with $E\|Q_i \nabla f_i\|^2 \leq \infty$ and $\delta + c_n \to 0$. In the second term, $\int_{B_{c_n}} \|Q_i \nabla f_i\| f_i dX_i = o(n^{-1/2})$ by Assumption A.T. So the last term in (B.39) vanishes in probability as $n \to \infty$.

Observe that $\{f(X) < \delta + c_n, \hat{f}(X) \geq \delta\}$ implies $\{|\hat{f}(X) - f(X)| < c_n, \hat{f}(X) \geq \delta\}$ which implies $\{f(X) \geq \delta - c_n\}$. So the first term of (B.39)

$$\begin{aligned}
&\|n^{-1/2} \sum_{i=1}^{n} \nabla f_i(\hat{Q}_i - Q_i) I_i\| \leq n^{-1/2} \sum_{i=1}^{n} \|\nabla f_i(\hat{Q}_i - Q_i)\| I_i \\
&\leq sup\Big\{ \Big| \hat{Q}_i - Q_i \Big| \mathbf{1}_{\{f(X_i) \geq \delta - c_n\}} \Big\} \cdot n^{-1/2} \sum_{i=1}^{n} \|\nabla f_i\| I_i \\
&= O_p(\delta^{-1}(nh^q)^{-1/2}) \cdot O_p(1) = o_p(1),
\end{aligned}$$

where $n^{-1/2} \sum_{i=1}^{n} \|\nabla f_i\| I_i = O_p(1)$ by the central limit theorem by $E\|\nabla f_i\|^2 < \infty$, and the uniform convergence of $\hat{Q}_i$ is implied by Assumption A.B. Similarly, the rest terms of (B.39) vanish in probability.

B.3 Proof of Theorem 2.2

For $r_{I\!I i}$, I follow Hardle and Stoker (1989) using the projection structure in its U-statistic. By (B.8), r_i is the limit of $r_n(Z_i) = E[\xi_{nij}|Z_i]$. So r_i can be estimated directly by its sample analogue of $r_n(Z_i)$, $\hat{r}_i \equiv (n-1)^{-1} \sum_{j \neq i} \hat{\xi}_{nij}$, where $\hat{\xi}_{nij}$ is obtained by a first-step nonparametric estimation of the unknown functions in ξ_{nij} (B.12).

By a similar argument for the trimming, it suffices to show consistency of $n^{-1} \sum \hat{r}_i \hat{r}_i' \mathbf{1}_{X_i}$ for $E(rr')$ and $n^{-1} \sum \hat{r}_i \mathbf{1}_{X_i}$ for $E(r)$. First, I need to show $sup|\hat{r}_i - r_i| \mathbf{1}_{X_i} = o_p(1)$. By the triangle inequality,

$$\begin{aligned}
&\sup |\hat{r}_i - r_i| \mathbf{1}_{X_i} \leq \sup |\hat{r}_{Ii} - r_{Ii}| \mathbf{1}_{X_i} + \sup |\hat{r}_{I\!I i} - r_{I\!I i}| \mathbf{1}_{X_i} \\
&\leq \sup |\hat{r}_{Ii} - r_{Ii}| \mathbf{1}_{X_i} + \sup \Big| \frac{1}{n-1} \sum_{j \neq i} (\hat{\xi}_{nij} - \xi_{nij}) \Big| \mathbf{1}_{X_i} \\
&+ \sup \Big| \frac{1}{n-1} \sum_{j \neq i} \xi_{nij} - E[\xi_{nij}|Z_i] \Big| \mathbf{1}_{X_i} + \sup \Big| r_{I\!I ni} - r_{I\!I i} \Big| \mathbf{1}_{X_i},
\end{aligned}$$

where $r_{I\!I ni} = E[\xi_{nij}|Z_i]$ and ξ_{nij} for $(I\!I)$ is defined in (B.12). The first and the second terms are $o_p(1)$ by the uniform convergence of the nonparametric estimation in Appendix A.1. The third term is $o_p(1)$ by the law of large number. The last term is $o_p(1)$ by the proof of Theorem B.2.

Since $\sup |\hat{r}_i - r_i| \mathbf{1}_{X_i} = o_p(1)$, the variance of r_i exists, and $Pr(f(X) \leq \delta) = o(1)$,

$$n^{-1} \sum_{i=1}^{n} \hat{r}_i \hat{r}_i' \mathbf{1}_{X_i} - E[r_i r_i'] = n^{-1} \sum_{i=1}^{n} (\hat{r}_i - r_i)(\hat{r}_i - r_i)' \mathbf{1}_{X_i} + n^{-1} \sum_{i=1}^{n} r_i (\hat{r}_i - r_i)' \mathbf{1}_{X_i}$$
$$+ n^{-1} \sum_{i=1}^{n} (\hat{r}_i - r_i) \hat{r}_i' \mathbf{1}_{X_i} - n^{-1} \sum_{i=1}^{n} \hat{r}_i \hat{r}_i' (1 - \mathbf{1}_{X_i}) + n^{-1} \sum_{i=1}^{n} r_i r_i' - E[r_i r_i'] = o_p(1).$$

Appendix C: Proofs of Theorems in Chapter 3

C.1 Proof of Theorem A.3

I implement the main results in Severini and Tripathi (2001). I start with the definitions and construct the Hilbert space. The unknown probability density or mass function of the random vector $(Y, X')' \in \Omega = S_Y \times S_X$ with respect to the measure P, the products of the Lebesgue measure μ_Y and μ_X, [1] is written as $f(Y,X) = f(Y|X)f(X) := \psi_0^2(Y|X)\phi_0^2(X)$. The functionals ψ_0 and ϕ_0 belong to the following sets defined by the regularity conditions:

$$\Psi_Y := \Big\{\psi \in S_Y \times S_X \to R, \psi^2(Y|X) \geq 0, \text{bounded and uniformly continuous in } y,$$
$$\text{uniformly in } x \text{ over the support of } X, \int_{S_Y} \psi^2(y|X)dy = 1\Big\},$$
$$\Phi := \Big\{\phi \in L^2(S_X;\mu_X), \phi^2(X) \geq 0, \int_{S_X} \phi^2(x)\mu_X(dx) = 1,$$
$$\int_{S_X} \|x\|^{2+\epsilon}\phi^2(x)\mu_X(dx) < \infty, \text{for some } \epsilon > 0\Big\}.$$

Let $\mathcal{A} := \Psi_Y \times \Phi$.

Definition A vector $\dot{\xi} = (\dot{\psi}, \dot{\phi})$ is said to be tangent to $\mathcal{A}$ at ξ_0 if it is the slope of $\xi_t := (\psi_t, \phi_t)$ at $t = 0$, i.e., $lim_{t\to 0}\|t^{-1}(\xi_t - \xi_0) - \dot{\xi}\| = 0$.

Definition The tangent space to $\mathcal{A}$ at the true value ξ_0, denoted as $\overline{lin\, T(\mathcal{A}, \xi_0)}$, is the smallest linear space which is closed under the L^2-norm and contains all $\dot{\xi} \in L^2(\Omega; \mu_Y \times \mu_X)$ tangent to $\mathcal{A}$ at ξ_0.

Severini and Tripathi (2001) show that the tangent space $\overline{lin\, T(\mathcal{A}, \xi_0)}$ is the product of $\overline{lin\, T(\Psi_Y, \psi_0)}$ and $\overline{lin\, T(\Phi, \phi_0)}$, where

$$\overline{lin\, T(\Psi_Y, \psi_0)} := \Big\{\dot{\psi} \in L^2(\Omega; \mu_Y \times X), \int_{S_Y} \dot{\psi}(y|X)\psi_0(y|X)\mu_Y(dy) = 0 a.s.\Big\}$$
$$\overline{lin\, T(\Phi, \phi_0)} := \Big\{\dot{\phi} \in L^2(S_X;\mu_X), \int_{S_X} \dot{\phi}(x)\phi_0(x)\mu_X(dx) = 0\Big\}.$$

[1] μ_X may not be a Lebesgue measure since I allow discrete components in the covariates X.

The pseudo-true model is the unconditional moment restriction $E_0\big[(\tau - \mathbf{1}_{\{Y \le X'\beta_0\}})X\big] = 0$ in (3.3). Here E_0 is the expectation with respect to the true density functions $\xi_0 = (\psi_0, \phi_0)$ and β_0 denotes the pseudo-true $\beta(\tau)$ for notational simplicity. The objective is to estimate the efficiency bound for estimating β_0. Equivalently, I can instead look at the efficiency bound for estimating the functional $\eta(\psi_0, \phi_0) := c'\beta_0$ for any arbitrary vector $c \in R^d$. Severini and Tripathi (2001) parameterize $\xi_0 = (\psi_0, \phi_0) \in \mathcal{A}$ and β_0 as a one-dimensional subproblem. For some $t_0 > 0$, let $t \mapsto (\xi_t, \beta_t)$ be a curve from $[0, t_0]$ into $\mathcal{A} \times R^d$ which passes through (ξ_0, β_0) at $t = 0$. That is, estimating $\eta(\xi_t) = c'\beta_t = t$ at the true parameter $t = 0$ is equivalent to estimating $t = 0$. The likelihood of estimating t using a single observation $(Y, X')'$ is given by $\psi_t^2(Y|X)\phi_t^2(X)$, so the score function for estimating $t = 0$ is

$$\begin{aligned} S_0(Y, X) &:= \frac{d}{dt}\, log\big[\psi_t^2(Y|X)\phi_t^2(X)\big]\mathbf{1}_{\{\psi_t(Y|X)>0\}}\mathbf{1}_{\{\phi_t(X)>0\}}\big|_{t=0} \\ &= 2\frac{\dot\psi(Y|X)}{\psi_0(Y|X)}\mathbf{1}_{\{\psi_0(Y|X)>0\}} + 2\frac{\dot\phi(X)}{\phi_0(X)}\mathbf{1}_{\{\phi_0(X)>0\}}. \end{aligned}$$

Then the Fisher information at $t = 0$ can be written as

$$\begin{aligned} i_F &= E[S_0(Y, X)S_0'(Y, X)] = \int_{S_X}\int_{S_Y} S_0(y, x)S_0'(y, x)\psi_0^2(y|x)\phi_0^2(x)\mu_Y(dy)\mu_X(dx) \\ &= 4E_X\Big[\int_{S_Y} \dot\psi(y|X)\dot\psi'(y|X)\mathbf{1}_{\{\psi_0(y|X)>0\}}\mu_Y(dy)\Big] + 4\int_{S_X} \dot\phi(x)\dot\phi'(x)\mathbf{1}_{\{\phi_0(x)>0\}}\mu_X(dx) \\ &:= < (\dot\psi, \dot\phi), (\dot\psi, \dot\phi) >_F, \end{aligned}$$

where the third equality is because $\dot\xi_0 = (\dot\psi_0, \dot\phi_0) \in \overline{lin\ T(\mathcal{A}, \xi_0)}$, and E_X denotes integrals with respect to the distribution of X. Therefore, the Fisher information inner product $< \cdot, \cdot >_F$ and the corresponding norm $\|\cdot\|_F$ are defined as

$$< \dot\xi_1, \dot\xi_2 >_F := 4E_X\Big[\int_{S_Y} \dot\psi_1(y|X)\dot\psi_2'(y|X)\mathbf{1}_{\{\psi_0(y|X)>0\}}\mu_Y(dy)\Big] + 4\int_{S_X} \dot\phi_1(x)\dot\phi_2'(x)\mathbf{1}_{\{\phi_0(x)>0\}}\mu_X(dx)$$

$$\|\dot\xi_1\|_F^2 = \|(\dot\psi_1, \dot\phi_1)\|_F^2 := < (\dot\psi_1, \dot\phi_1), (\dot\psi_1, \dot\phi_1) >_F$$

for any $\dot\xi_1.\dot\xi_2 \in \overline{lin\ T(\mathcal{A}, \xi_0)}$ which is a closed subset of $L^2(\Omega; P)$. Hence I have constructed the Hilbert space $(\overline{lin\ T(\mathcal{A}, \xi_0)}, < \cdot, \cdot >_F)$.

Now I are ready to derive the efficiency bounds. It is known that the information inequality holds for all regular estimators, i.e., the asymptotic covariance of the estimator

$\geq 1/i_F = \|\dot{\xi}_0\|_F^{-2}$. The semiparametric bound can be interpreted as the supremum of the asymptotic covariance over the parametric submodels. Severini and Tripathi (2001) shows that the lower bound is

$$\begin{aligned} l.b. &= \sup_{\{\dot{\xi}\in(\overline{lin\ T(\mathcal{A},\xi_0)}:\dot{\xi}\neq 0,\nabla\eta(\dot{\xi})=1\}} \|\dot{\xi}\|_F^{-2} = \sup_{\{\dot{\xi}\in(\overline{lin\ T(\mathcal{A},\xi_0)}:\|\dot{\xi}\|_F=1\}} |\nabla\eta(\dot{\xi})|^2 \\ &= \|\nabla\eta\|_*^2 = \|\xi^*\|_F^2. \end{aligned} \tag{C.1}$$

The third equality is the norm of linear functional $\nabla\eta$, the pathwise derivative of η (Luenberger, 1969, p.105). The forth equality is from the Riesz-Fréchet theorem: there exists a unique $\xi^* \in \overline{lin\ T(\mathcal{A},\xi_0)}$ for the continuous linear functional $\nabla\eta$ on the Hilbert space $(\overline{lin\ T(\mathcal{A},\xi_0)}, <\cdot,\cdot>_F)$ such that $\nabla\eta(\dot{\xi}) = <\xi^*, \dot{\xi}>_F$ for all $\dot{\xi} \in \overline{lin\ T(\mathcal{A},\xi_0)}$, i.e.,

$$\begin{aligned} \nabla\eta(\dot{\psi},\dot{\phi}) = c'\dot{\beta} &=< (\psi^*,\phi^*),(\dot{\psi},\dot{\phi}) >_F \\ &= 4E_X\Big[\int_{S_Y} \psi^*\dot{\psi}'\mathbf{1}_{\{\psi_0(y|X)>0\}}\mu_Y(dy)\Big] + 4\int_{S_X} \phi^*\dot{\phi}'\mathbf{1}_{\{\phi_0(x)>0\}}\mu_X(dx). \end{aligned} \tag{C.2}$$

So to find the lower bound by (C.1), I need to find ξ^* which is known as the representer of the continuous linear functionals $\nabla\eta$.

The submodel $(\psi_t, \phi_t, \beta_t)$ should also satisfy the unconditional moment restriction, $E_t[(\mathbf{1}_{\{Y\leq X'\beta_t\}} - \tau)X] = 0$. For any $\tau_1, \tau_2 \in (0,1)$, $\beta_t := (\beta_t'(\tau_1), \beta_t'(\tau_2))' := (\beta_{1t}', \beta_{2t}')'$. I simultaneously estimate $\beta_0 = (\beta_{10}', \beta_{20}')'$, a $2d$-dimensional vector, so the unconditional moment restriction is

$$\int_{S_X}\int_{S_Y} \begin{pmatrix} \big(\mathbf{1}_{\{y\leq x'\beta_{1t}\}} - \tau_1\big)x \\ \big(\mathbf{1}_{\{y\leq x'\beta_{2t}\}} - \tau_2\big)x \end{pmatrix} \psi_t^2(y|x)\phi_t^2(x)\mu_Y(dy)\mu_X(dx) = 0.$$

Taking the derivative with respect to t evaluated at $t = 0$, [2]

$$
\begin{aligned}
0 = &\int_{S_X} \begin{pmatrix} xx' f_Y(x'\beta_{10}|x)\dot{\beta}_1 \\ xx' f_Y(x'\beta_{20}|x)\dot{\beta}_2 \end{pmatrix} \phi_0^2(x)\mu_X(dx) \\
&+ 2\int_{S_X}\int_{S_Y} \begin{pmatrix} x\mathbf{1}_{\{y\le x'\beta_{10}\}} \\ x\mathbf{1}_{\{y\le x'\beta_{20}\}} \end{pmatrix} \psi_0(y|x)\dot{\psi}'(y|x)\mu_Y(dy)\phi_0^2(x)\mu_X(dx) \\
&+ 2\int_{S_X} \begin{pmatrix} x\big(F_Y(x'\beta_{10}|x) - \tau_1\big) \\ x\big(F_Y(x'\beta_{20}|x) - \tau_2\big) \end{pmatrix} \phi_0(x)\dot{\phi}'(x)\mu_X(dx).
\end{aligned}
$$

where the second term is because $\dot{\psi} \in \overline{lin\ T(\Psi_Y, \psi_0)}$ implies $\int_{S_Y} \psi_0 \dot{\psi}\mu_Y(dy) = 0$. Note that $\int_{S_X} xx' f_Y(x'\beta_0|x)$
$\phi_0^2(x)\mu_X(dx) = E_0[XX' f_Y(X'\beta_0|X)] = J(\tau)$ which is assumed to be positive definite by (R3), so $J(\tau)^{-1}$ exists. Define

$$
D := \begin{pmatrix} J(\tau_1) & 0 \\ 0 & J(\tau_2) \end{pmatrix},
$$

so D^{-1} exists. Then

$$
\begin{aligned}
\begin{pmatrix} \dot{\beta}_1 \\ \dot{\beta}_2 \end{pmatrix} = -2D^{-1}\Big[&\int_{S_X}\int_{S_Y} \begin{pmatrix} x\mathbf{1}_{\{y\le x'\beta_{10}\}} \\ x\mathbf{1}_{\{y\le x'\beta_{20}\}} \end{pmatrix} \psi_0(y|x)\dot{\psi}'(y|x)\mathbf{1}_{\{\psi_0(y|x)>0\}}\mu_Y(dy)\phi_0^2(x)\mu_X(dx) \\
&+2\int_{S_X} \begin{pmatrix} x\big(F_Y(x'\beta_{10}|x) - \tau_1\big) \\ x\big(F_Y(x'\beta_{20}|x) - \tau_2\big) \end{pmatrix} \phi_0(x)\dot{\phi}'(x)\mathbf{1}_{\{\phi_0(x)>0\}}\mu_X(dx)\Big].
\end{aligned}
\tag{C.3}
$$

I confirm that $\nabla\eta(\dot{\xi}) = c'\dot{\beta}$ is a continuous linear functional on $\overline{lin\ T(\mathcal{A}, \xi_0)}$, so η is indeed pathwise differentiable. From (C.2) and (C.3), I can find the representer for $\nabla\eta$ as

$$
\phi^*(x) = -\frac{1}{2}c'D^{-1} \begin{pmatrix} \big(F_Y(x'\beta_{10}|x) - \tau_1\big)x \\ \big(F_Y(x'\beta_{20}|x) - \tau_2\big)x \end{pmatrix} \phi_0(x),
$$

[2]The interchange of differentiation and integration is alloId, assumed throughout Severini and Tripathi (2001), by the smoothness of $\xi_t(Y, X)$ in $t \in [0, t_0]$ by the construction of regular parametric submodels; see Newey (1990) for details.

and

$$\psi^*(y|x) = -\frac{1}{2}c'D^{-1}\begin{pmatrix} \big(\mathbf{1}_{\{y\le x'\beta_{10}\}} - F_Y(x'\beta_{10}|x)\big)x \\ \big(\mathbf{1}_{\{y\le x'\beta_{20}\}} - F_Y(x'\beta_{20}|x)\big)x \end{pmatrix}\psi_0(y|x)$$

because $\dot{\psi} \in \overline{lin\ T(\Psi_Y, \psi_0)}$. It can be easily checked that $(\psi^*, \phi^*) \in \overline{lin\ T(\mathcal{A}, \xi_0)}$. For notational ease, denote $\mathbf{1}_i := \mathbf{1}_{\{y\le x'\beta_{i0}\}}$ and $F_i := F_Y(x'\beta_{i0}|x)$, $i = 1, 2$. Then the lower bound for regular $\sqrt{n}$-consistent estimators of $c'\beta_0$ is

$$\begin{aligned}
&\|(\psi^*, \phi^*)\|_F^2 \\
=&c'D^{-1}\Bigg\{\int_{S_X}\begin{pmatrix} \big(F_1-\tau_1\big)^2xx' & \big(F_1-\tau_1\big)\big(F_2-\tau_2\big)xx' \\ \big(F_1-\tau_1\big)\big(F_2-\tau_2\big)xx' & \big(F_2-\tau_2\big)^2xx' \end{pmatrix}\phi_0^2(x)\mathbf{1}_{\{\phi_0(x)>0\}}\mu_X(dx) \\
&+E\begin{pmatrix} E\Big[\big(\mathbf{1}_1-F_1\big)^2|X\Big]XX' & E\Big[\big(\mathbf{1}_1-F_1\big)\big(\mathbf{1}_2-F_2\big)|X\Big]XX' \\ E\Big[\big(\mathbf{1}_1-F_1\big)\big(\mathbf{1}_2-F_2\big)|X\Big]XX' & E\Big[\big(\mathbf{1}_2-F_2\big)^2|X\Big]XX' \end{pmatrix}\Bigg\}D^{-1}c \\
=&cD^{-1}\begin{pmatrix} \Gamma(\tau_1,\tau_1) & \Gamma(\tau_1,\tau_2) \\ \Gamma(\tau_1,\tau_2) & \Gamma(\tau_2,\tau_2) \end{pmatrix}D^{-1}c,
\end{aligned}$$

where

$$\begin{aligned}
\Gamma(\tau_1,\tau_2) &:= E\Big[E\big[(F_1-\tau_1)(F_2-\tau_2)+(\mathbf{1}_1-F_1)(\mathbf{1}_2-F_2)|X\big]XX'\Big] \\
&= E\Big[E\big[\tau_1\tau_2-\tau_1F_2-\tau_2F_1+\mathbf{1}_1\mathbf{1}_2|X\big]XX'\Big] \\
&= E\Big[E\big[\big(\tau_1-\mathbf{1}_{\{y\le X'\beta_{10}\}}\big)\big(\tau_2-\mathbf{1}_{\{y\le X'\beta_{20}\}}\big)|X\big]XX'\Big] \\
&= E\Big[\big(\tau_1-\mathbf{1}_{\{y\le X'\beta_{10}\}}\big)\big(\tau_2-\mathbf{1}_{\{y\le X'\beta_{20}\}}\big)XX'\Big]
\end{aligned}$$

by the law of iterated expectations, and so $\Gamma(\tau,\tau) = E\Big[\big(\tau-\mathbf{1}_{\{y\le X'\beta_0\}}\big)^2XX'\Big]$.

So the lower bound for estimating $\beta(\tau)$ is $J(\tau)^{-1}\Gamma(\tau,\tau)J(\tau)^{-1}$. The asymptotic covariance of the estimators for $\beta(\tau_1)$ and $\beta(\tau_2)$ cannot be smaller than $J(\tau_1)^{-1}\Gamma(\tau_1,\tau_2)J(\tau_2)^{-1}$.

Consider the efficiency bound for estimating one single quantile $\beta(\tau)$ by Newey's (1990) approach. Severini and Triphathi (2001) claim that Newey's efficient influence function for $c'\beta(\tau)$ is $2\psi^*/\psi_0 + 2\phi^*/\phi_0 = c'J(\tau)^{-1}X(\tau - F_Y(X'\beta|X)) + c'J(\tau)^{-1}X(F_Y(X'\beta|X) - \mathbf{1}_{\{Y\le X'\beta\}}) = c'J(\tau)^{-1}X(\tau - \mathbf{1}_{\{Y\le X'\beta\}})$. Then the efficient influence function for β_0 is

$(E[SS'])^{-1}S$, where the efficient score $S = J(\tau)\Gamma(\tau,\tau)^{-1}X(\tau - \mathbf{1}_{\{Y \leq X'\beta\}})$. Newey shows the semiparametric bound is $(E[SS'])^{-1}$. □

C.2 Proof of the Semiparametric Efficiency Bound for the Correct Linear Specified QR (3.4)

Under correct specification, $F_Y(X'\beta_0|X) = \tau$. The random vectors (Y, X) satisfy the conditional moment restriction $E\big[\mathbf{1}_{\{Y \leq X'\beta_0\}} - \tau | X\big] = 0$, i.e., $\int_{S_Y} \big(\mathbf{1}_{\{y \leq X'\beta_0\}} - \tau\big)\psi_0^2(y|X)\mu_Y(dy) = 0$, where the joint distribution of (Y, X) is $\psi_0^2(Y|X)\phi_0^2(X)$. The Hilbert space $(\overline{lin\ T(\mathcal{A}, \xi_0)}, < \cdot, \cdot >_F)$ and $\mathcal{A} = (\Psi_Y, \Phi)$ are defined in the proof of Theorem A.3. Consider any $\tau_1.\tau_2 \in (0,1)$, $\beta_t := (\beta_t'(\tau_1), \beta_t'(\tau_2))' := (\beta_{1t}', \beta_{2t}')'$. The parameterized submodel $(\psi_t, \phi_t, \beta_t)$ also have to satisfy the moment condition

$$\int_{S_Y} \begin{pmatrix} \big(\mathbf{1}_{\{y \leq X'\beta_{1t}\}} - \tau_1\big) \\ \big(\mathbf{1}_{\{y \leq X'\beta_{2t}\}} - \tau_2\big) \end{pmatrix} \psi_t^2(y|X)\mu_Y(dy) = 0,$$

Taking the derivative with respect to t evaluated at $t = 0$, I have

$$\begin{pmatrix} f_Y(X'\beta_1|X)X'\dot{\beta}_1 \\ f_Y(X'\beta_2|X)X'\dot{\beta}_2 \end{pmatrix} + 2\int_{S_Y} \begin{pmatrix} \big(\mathbf{1}_{\{y \leq X'\beta_{10}\}} - \tau_1\big) \\ \big(\mathbf{1}_{\{y \leq X'\beta_{20}\}} - \tau_2\big) \end{pmatrix} \psi_0(y|X)\dot{\psi}'(y|X)dy = 0, \qquad \text{(C.4)}$$

where $f_Y(y|X) = \psi_0^2(y|X)$. Define

$$D(X) := \begin{pmatrix} f_Y(X'\beta_1|X)X' & 0 \\ 0 & f_Y(X'\beta_2|X)X' \end{pmatrix}.$$

Note that (C.4) has overidentifying moment restriction that cannot uniquely solve $\dot{\beta}$. To locally identify $\dot{\beta}$, Severini and Tripathi (2001) give the sufficient condition by $W(X)$ which is some nonsingular (w.p.1) 2×2 matrix such that $E[D(X)'W(X)D(X)]$ is nonsingular. By assumption, $E[XX'f_Y^2(X'\beta(\tau)|X)] = E[XX'f_{\epsilon_\tau}^2(0|X)]$ exists and is nonsingular, so the same holds for $E[D'(X)D(X)]$. Hence, I can choose $W(X) = \mathbf{1}$, identity matrix. Multiply (C.4) by $D'(X)$:

$$D'(X)D(X)\begin{pmatrix} \dot{\beta}_1 \\ \dot{\beta}_2 \end{pmatrix} + D'(X)2\int_{S_Y} \begin{pmatrix} \big(\mathbf{1}_{\{y \leq X'\beta_{10}\}} - \tau_1\big) \\ \big(\mathbf{1}_{\{y \leq X'\beta_{20}\}} - \tau_2\big) \end{pmatrix} \psi_0(y|X)\dot{\psi}'(y|X)dy = 0.$$

Take expectations on both sides with respect to X and solve for $\dot{\beta}$: $(\beta_1', \beta_2')' =$

$$-2\Big(E[D'(X)D(X)]\Big)^{-1}E\Big[D'(X)2\int_{S_Y}\begin{pmatrix}\big(\mathbf{1}_{\{y\leq X'\beta_{10}\}}-\tau_1\big)\\ \big(\mathbf{1}_{\{y\leq X'\beta_{20}\}}-\tau_2\big)\end{pmatrix}\psi_0(y|X)\dot{\psi}'(y|X)\mathbf{1}_{\{\psi_0(y|X)>0\}}dy\Big].$$

Then for any arbitrary $c \in R^{2d}$, the representer for $\nabla\eta((\dot{\psi},\dot{\phi})) = c'\dot{\beta}$ is

$$\psi^*(y|X) = -\frac{1}{2}c'\Big(E[D'(X)D(X)]\Big)^{-1}D'(X)\begin{pmatrix}\big(\mathbf{1}_{\{y\leq X'\beta_{10}\}}-\tau_1\big)\\ \big(\mathbf{1}_{\{y\leq X'\beta_{20}\}}-\tau_2\big)\end{pmatrix}\psi_0(y|X)$$

$\in \overline{lin\ T(\Psi_Y,\psi_0)}$ by the conditional moment restriction. And $\phi^* = 0$ since ϕ_0 is just ancillary in this conditional moment case. Define $A := \Big(E\big[D'(X)D(X)\big]\Big)^{-1}$ and $\mathbf{1}_i := \mathbf{1}_{\{Y\leq X'\beta_{i0}\}}$ for $i = 1, 2$ for notational ease. Without loss of generality, assume $\tau_1 < \tau_2$. So the lower bound is $\|(\psi^*,\phi^*)\|_F^2 =$

$$c'E\left[AD'(X)\begin{pmatrix}E\Big[\big(\mathbf{1}_1-\tau_1\big)^2\Big|X\Big] & E\Big[\big(\mathbf{1}_1-\tau_1\big)\big(\mathbf{1}_2-\tau_2\big)\Big|X\Big]\\ E\Big[\big(\mathbf{1}_1-\tau_1\big)\big(\mathbf{1}_2-\tau_2\big)\Big|X\Big] & E\Big[\big(\mathbf{1}_2-\tau_2\big)^2\Big|X\Big]\end{pmatrix}D(X)A\right]c$$

$$= c'AE\left[D'(X)\begin{pmatrix}\tau_1(1-\tau_1) & \tau_1(1-\tau_2)\\ \tau_1(1-\tau_2) & \tau_2(1-\tau_2)\end{pmatrix}D(X)\right]Ac$$

$$= c'\begin{pmatrix}\tau_1(1-\tau_1)\Big\{E\big[XX'f^2_{\epsilon_{\tau_1}}(0|X)\big]\Big\}^{-1} & \tau_1(1-\tau_2)\Big\{E\big[XX'f_{\epsilon_{\tau_1}}(0|X)f_{\epsilon_{\tau_2}}(0|X)\big]\Big\}^{-1}\\ \tau_1(1-\tau_2)\Big\{E\big[XX'f_{\epsilon_{\tau_1}}(0|X)f_{\epsilon_{\tau_2}}(0|X)\big]\Big\}^{-1} & \tau_2(1-\tau_2)\Big\{E\big[XX'f^2_{\epsilon_{\tau_2}}(0|X)\big]\Big\}^{-1}\end{pmatrix}c$$

since $f_Y(X'\beta|X) = f_{\epsilon_\tau}(0|X)$ for correct specification.

Consider the efficiency bound for estimating one single quantile $\beta(\tau)$ by Newey's (1990) approach. Severini and Triphathi (2001) claim that Newey's efficient influence function for $c'\beta(\tau)$ is $2\psi^*/\psi_0 = c'\big(E\big[f^2_{\epsilon_\tau}(0|X)XX'\big]\big)^{-1}f_{\epsilon_\tau}(0|X)X(\tau - \mathbf{1}_{\{Y\leq X'\beta\}})$. Then the efficient influence function for β_0 is $(E[SS'])^{-1}S$, where the efficient score $S = (\tau-\tau^2)^{-1}f_{\epsilon_\tau}(0|X)X(\tau - \mathbf{1}_{\{Y\leq X'\beta\}})$. Newey shows the semiparametric bound is $(E[SS'])^{-1}$. □

Bibliography

Abadie, A. and G. W. Imbens (2006, 01). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Agüero, J., M. Carter, and I. Woolard (2010). The impact of unconditional cash transfers on nutrition: The south african child support grant. working paper.

Ai, C. and X. Chen (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics 170*(2), 442–457.

Altonji, J. G. and R. L. Matzkin (2005, 07). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Andrews, D. W. and X. Shi (2011, June). Inference based on conditional moment inequalities. Cowles Foundation Discussion Papers 1761, Cowles Foundation for Research in Economics, Yale University.

Angrist, J., V. Chernozhukov, and I. Fernndez-Val (2006, 03). Quantile regression under misspecification, with an application to the u.s. wage structure. *Econometrica 74*(2), 539–563.

Baez, J. E. and A. Camacho (2011). Assessing the long-term effect of conditional cash transfers on human capital: Evidence from colombia. World Bank Policy Research Working Paper 5681.

Behrman, J., S. Parker, and P. Todd (2011). Do conditional cash transfers for schooling generate lasting benefits? a five-year follow up of progesa/oportunidades. *Journal of Human Resources* (46), 93–122.

Behrman, J. R., Y. Cheng, and P. E. Todd (2004, February). Evaluating preschool programs when length of exposure to the program varies: A nonparametric approach. *The Review of Economics and Statistics 86*(1), 108–132.

Bhattacharya, D. (2007, April). Inference on inequality from household survey data. *Journal of Econometrics 137*(2), 674–707.

Bhattacharya, P. K. and A. K. Gangopadhyay (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics 18*(3), pp. 1400–1415.

Blundell, R. and J. L. Powell (2003). *Endogeneity in Nonparametric and Semiparametric Regression Models*, Volume II. Cambridge University Press, Cambridge, U.K.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–154.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2010). Robust data-driven inference for density-weighted average derivatives. *Journal of the American Statistical Association 105*(491), 1070–1083.

Chamberlain, G. (1987, March). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chaudhuri, P., K. Doksum, and A. Samarov (1997). On average derivative quantile regression. *The Annals of Statistics 25*(2), pp. 715–744.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics 36*(2), pp. 808–843.

Chen, X. and D. Pouzo (2009, September). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics 152*(1), 46–60.

Chernozhukov, V., A. Belloni, and I. Fernandez-Val (2011). Conditional quantile processes based on series and many regressors (with an application to gasoline demand). Technical report.

Chernozhukov, V., I. Fernandez-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica forthcoming*.

Chernozhukov, V., I. Fernndez-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Chernozhukov, V. and C. Hansen (2005, 01). An iv model of quantile treatment effects. *Econometrica 73*(1), 245–261.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica 71*(5), 1405–1441.

Dabrowska, D. M. (1992). Nonparametric quantile regression with censored data. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002) 54*(2), pp. 252–259.

Darolles, S., J.-P. Florens, and E. M. Renault (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996, September). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica 64*(5), 1001–44.

Donald, S. G., Y.-C. Hsu, and G. F. Barrett (2012). Incorporating covariates in the measurement of welfare and inequality: methods and applications. *The Econometrics Journal 15*(1), C1–C30.

Escanciano, J. C., D. Jacho-Chavez, and A. Lewbel (2012, May). Uniform convergence of weighted sums of non- and semi-parametric residuals for estimation and testing. Boston College Working Papers in Economics 756, Boston College Department of Economics.

Ferreira, Francisco H. G., F. D. and N. Schady. Own and sibling effects of conditional cash transfer programs : Theory and evidence from cambodia. Technical report.

Firpo, S. (2007, 01). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), 259–276.

Firpo, S. and C. Pinto (2011). Identification and estimation of distributional impacts of interventions using changes in inequality measures. Working paper.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008, 09). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica 76*(5), 1191–1206.

Flores, C. A. (2007, November). Estimation of dose-response functions and optimal doses with a continuous treatment. Working Papers 0707, University of Miami, Department of Economics.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012, February). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *The Review of Economics and Statistics 94*(1), 153–171.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica 79*(2), 437–452.

Haerdle, W., P. Janssen, and R. Serfling (1988). Strong uniform consistency rates for estimators of conditional functionals. *The Annals of Statistics 16*(4), 1428–1449.

Hahn, J. (1997, November). Bayesian bootstrap of the quantile regression estimator: A large sample study. *International Economic Review 38*(4), 795–808.

Hahn, J. (1998, March). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–332.

Hahn, J. and G. Ridder (2013). The asymptotic variance of semi-parametric estimators with generated regressors. *Econometrica 81*(1), 315340.

Hansen, B. E. (2008, June). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*(03), 726–748.

Hardle, W. and T. M. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association 84*(408), 986–995.

Heckman, J. J., H. Ichimura, and P. Todd (1998, April). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*(2), 261–94.

Hirano, K. and G. Imbens (2004). *The Propensity Score with Continuous Treatments*, Chapter 7. John Wiley and Sons.

Hirano, K., G. W. Imbens, and G. Ridder (2003, 07). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hoderlein, S. and E. Mammen (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica 75*(5), 1513–1518.

Hoderlein, S. and E. Mammen (2009). Identification and estimation of local average derivatives in non-separable models without monotonicity. *Econometrics Journal 12*(1), 1–25.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Ibarraran, P. and J. Villa (2010). Labor insertion assessment of conditional cash transfer programs: A dose-response estimate for mexicos oportunidades. Mimeo, Inter-American Development Bank.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics 159*, 252–266.

Imai, K. and D. van Dyk (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association 99*(467), 854–866.

Imbens, G. and J. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 586.

Imbens, G. W. and W. K. Newey (2009, 09). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Jacho-Chávez, D. T. (2009, June). Efficiency bounds for semiparametric estimation of inverse conditional-density-weighted functions. *Econometric Theory 25*(03), 847–855.

Kasy, M. (2013). Identification in general triangular systems. ¡p¿Original version: January 2012¡/p¿.

Khan, S. (2001, February). Two-stage rank estimation of quantile index models. *Journal of Econometrics 100*(2), 319–355.

Khan, S. and E. Tamer (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica 78*(6), 2021–2042.

Kim, T.-H. and H. White (2003). Estimation, inference, and specification testing for possibly misspecified quantile regression. *Advances in Econometrics 17*, 107–132.

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao (2012). Evaluating continuous training programs using the generalized propensity score. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(2), 587–617.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression quantile. *Econometrica 46*, 33–50.

Komunjer, I. (2005). Quasi-maximum likelihood estimation for conditional quantiles. *Journal of Econometrics 128*(1), 137–164.

Kong, E. and Y. Xia (2012, 7). A single-index quantile regression model and its estimation. *Econometric Theory 28*, 730–768.

Kosorok, M. R. (2008). Springer: New York.

Lavergne, P. and Q. H. Vuong (1996). Nonparametric selection of regressors: The nonnested case. *Econometrica 64*(1), 207–19.

Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory 19*(01), 1–31.

Lee, Y.-Y. (2011). Nonparametric density-weighted average quantile derivative. Working paper.

Lewbel, A. (1998, January). Semiparametric latent variable model estimation with endogenous or mismeasured regressors. *Econometrica 66*(1), 105–122.

Ma, L. and R. Koenker (2006). Quantile regression methods for recursive structural equation models. *Journal of Econometrics 134*(2), 471 – 506.

Magnac, T. and E. Maurin (2007). Identification and information in monotone binary models. *Journal of Econometrics 139*, 76–104.

Mammen, E., C. Rothe, and M. Schienle (2012a). Nonparametric regression with nonparametrically generated covariates. *Annals of Statistics 40*(SFB649DP2010-059), 1132–1170.

Mammen, E., C. Rothe, and M. Schienle (2012b). Semiparametric estimation with generated covariates. working paper.

Matzkin, R. L. (2007). Nonparametric identification. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, Chapter 73. Elsevier.

Newey, W. (1994a). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382.

Newey, W. and T. M. Stoker (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica 61*(5), 1199.

Newey, W. K. (1990, April-Jun). Semiparametric efficiency bounds. *Journal of Applied Econometrics 5*(2), 99–135.

Newey, W. K. (1994b, June). Kernel estimation of partial means and a general variance estimator. *Econometric Theory 10*(02), 1–21.

Newey, W. K. and J. L. Powell (1990, September). Efficient estimation of linear and type 1 censored regression models under conditional quantile restrictions. *Econometric Theory 6*(03), 295–317.

Newey, W. K., J. L. Powell, and F. Vella (1999, May). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–604.

Otsu, T. (2008, January). Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics 142*(1), 508–538.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Conference Board of the Mathematical Science: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics 32*(1), 143 – 155.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), pp. 1403–1430.

Powell, J. L. and T. M. Stoker (1996, December). Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics 75*(2), 291–316.

Rodrguez-Oreggia, E. and S. Freije (2012). Long term impact of a cash-transfers program on labor outcomes of the rural youth. Cid working paper.

Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics 155*, 56–70.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data in the fisher randomization test" by d. basu. *Journal of the American Statistical Association 75*, 591–593.

Severini, T. A. and G. Tripathi (2001, May). A simplified approach to computing efficiency bounds in semiparametric models. *Journal of Econometrics 102*(1), 23–66.

Sherman, R. (1994). Maximal inequalities for degenerate u-processes with applications to optimization estimators. *The Annals of Statistics 22*(1), 439–459.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall.

Song, K. (2008). Uniform convergence of series estimators over function spaces. *Econometric Theory 24*, 1463–1499.

Song, K. (2012a). On the smoothness of conditional expectation functionals. *Statistics & Probability Letters 82*(5), 1028–1034.

Song, K. (2012b). Semiparametric models with single-index nuisance parameters. working paper.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica 54*(6), pp. 1461–1481.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: with Application to Statistics*. New York: Springer-Verlag.

Whang, Y.-J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory 22*, 173–205.

White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review 21*, 149–170.

White, H. and K. Chalak (2013). Identification and identification failure for treatment effects using structural systems. *Econometric Reviews 32*(3), 273–317.

Wu, T. Z., K. Yu, and Y. Yu (2010). Single-index quantile regression. *Journal of Multivariate Analysis 101*(7), 1607 – 1621.