Sample Selection Models with Monotone Control Functions

Ruixuan Liu and Zhengfei Yu
Emory University and University of Tsukuba

Abstract. The celebrated Heckman selection model yields a selection correction function (control function) proportional to the inverse Mills ratio, which is monotone. This paper studies a sample selection model which does not impose parametric distributional assumptions on the latent error terms, while maintaining the monotonicity of the control function. We show that a positive (negative) dependence condition on the latent error terms is sufficient for the monotonicity of the control function. The condition is equivalent to a restriction on the copula function of latent error terms. Utilizing the monotonicity, we propose a tuning-parameter-free semiparametric estimation method and establish root n-consistency and asymptotic normality for the estimates of finite-dimensional parameters. A new test for selectivity is also developed exploring the shape-restricted estimation. Simulations and an empirical application are conducted to illustrate the usefulness of the proposed methods.

Key words: Copula, Sample Selection Models, Isotonic Regression, Semiparametric Estimation, Shape Restriction

JEL Classification: C14, C21, C24, C25

1. Introduction

The sample selection problem arises frequently in economics when observations are not taken from a random sample of the population. Understanding the self-selection process and correcting selection bias is a central task in empirical studies of the determinants of occupational wages (Roy, 1951; Heckman and Honore, 1990), the labor supply behavior of females (Heckman, 1974; Gronau, 1974; Arellano and Bonhomme, 2017), schooling

The first draft: October 18, 2018. This version: March 22, 2019.
Corresponding Address: Ruixuan Liu, Department of Economics, Emory University, 201 Dowman Drive, Atlanta, GA, USA, 30322, E-mail: *ruixuan.liu@emory.edu*.
We would like to thank Stephane Bonhomme, Yanqin Fan, Marc Henry, Essie Maasoumi, and Peter Robinson for helpful comments.

choices (Willis and Rosen, 1979; Cameron and Heckman, 1998), unionism status (Lee, 1978; Lemieux, 1998), and migration decisions (Borjas, 1987; Chiquiar and Hanson, 2005), among others. A prototypical sample selection model consists of the following outcome and selection equations:

$$
\begin{aligned}
Y_i^* &= X_i'\beta_0 + \varepsilon_i, \\
D_i &= \mathbb{I}\{W_i'\gamma_0 + \nu_i > 0\}, \\
Y_i &= Y_i^* D_i, \text{ for } i = 1, \cdots, n,
\end{aligned} \tag{1.1}
$$

where (Y_i, D_i, X_i', Z_i') are observed variables and (ε_i, ν_i) are latent error terms. The conditional mean function of the observed dependent variable Y_i is equal to

$$
\mathbb{E}[Y_i | X_i, W_i, D_i = 1] = X_i'\beta_0 + \lambda_0(W_i'\gamma_0), \tag{1.2}
$$

where $\lambda_0(W_i'\gamma_0) = \mathbb{E}[\varepsilon_i | \nu_i > -W_i'\gamma_0, W]$ corrects for the sample selection bias and is known as the control function[1] (Heckman and Robb, 1985, 1986).

Since the seminal work of Heckman (1979), Heckman's two-step method has been the default choice for estimating the sample selection model (1.1). The approach assumes the joint normality on the error terms (ε, ν). As a result, the control function has a known parametric form: $\lambda_0(W_i'\gamma_0)$ is proportional to the inverse Mills ratio $\phi(W_i'\gamma_0)/\Phi(W_i'\gamma_0)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution functions of the standard normal distribution, respectively. An interesting, yet somewhat neglected, property of the inverse Mills Ratio is its monotonicity.

In this paper, we consider a semiparametric sample selection model where the control function is monotone. We prove that a positive (or negative) dependence condition on (ε, ν), formally known as the *right tail increasing (decreasing)* (Esary and Proschan, 1972) is sufficient for the monotonicity of the control function. Intuitively, the right tail increasing (RTI) means that whenever ν is large, it is more likely that ε is large. This condition only depends on the copula function without imposing any distributional assumption on the latent errors either in the outcome or selection equation. In particular, a positive (negative) correlation coefficient of the Gaussian copula (as in the generalized selection model of Lee (1983)) leads to a monotonically decreasing (increasing) control function, regardless of marginal distribution specifications. We also show that the condition is easily verified for

[1] In some alternative formulation (Heckman and Vytlacil, 2007a,b), the control function $\tilde{\lambda}_0(\cdot)$ is defined as a function of the propensity score $P_i = \Pr\{D_i = 1 | W_i\}$. Therefore, for the model (1.1) one has $\lambda_0(v) = \tilde{\lambda}_0(\bar{F}_\nu(-v))$ where $\bar{F}_\nu$ is the survivor function of ν. However, this does not affect our discussion regarding the monotonicity of the control function, as it is straightforward to see that λ_0 and $\tilde{\lambda}_0$ are equivalent up to a monotone transformation.

many parametric families including the Archimedean copula models, Generalized Farlie-Gumbel-Morgenstern copula models, and normal mixture models. In practice, the choice between a positive and negative dependence is up to the researcher because it is often possible to postulate whether one gets positive or negative sorting for empirical questions.

Maintaining the monotonicity assumption of the control function, we propose a new semiparametric estimation method and a new test for selectivity that explore this shape restriction. Our method is fully automatic and free of any tuning parameter. The resulting estimators of the regression coefficients β_0 and γ_0 are root-n consistent and asymptotically normal. Compared with existing semiparametric procedures that make use of kernel or sieve estimation for certain nonparametric components, the main advantage of our approach is its tuning-parameter-free nature. The implementation of our method circumvents the need to pick bandwidths in kernel smoothing, penalization parameters in cubic splines or the order of polynomials in series estimation which are required by the majority of existing semiparametric approaches and are often chosen in ad-hoc ways. One exception is Cosslett (1991), who studies a tuning-parameter-free method different from our approach.[2] However, Cosslett (1991) only presents a consistency proof based on sample-splitting, whereas the rate of convergence and the asymptotic distribution remain unknown.

Our estimation method consists of two stages. In the first stage, we use the likelihood function for the binary choice data (D_i, W_i) in terms of the regression coefficient and latent error distribution in the selection equation:

$$\mathbb{L}_{1n}(\gamma, F) = \Pi_{i=1}^{n} \left\{ F(-W_i'\gamma)^{1-D_i} \left[1 - F(-W_i'\gamma)\right]^{D_i} \right\}, \tag{1.3}$$

to get our estimates $(\hat{\gamma}_n, \hat{F}_{n\nu}(\cdot; \hat{\gamma}_n))$, following Groeneboom and Hendrickx (2018). In the second stage, we obtain the estimator $\hat{\beta}_n$ and $\hat{\lambda}_n$ by estimating a partial linear model with a monotone nonparametric component (Huang, 2002) and generated regressor $W'\hat{\gamma}_n$:

$$(\hat{\beta}_n, \hat{\lambda}_n) = \arg\min_{\beta,\lambda} \sum_{i=1}^{n} D_i \left[Y_i - X_i'\beta - \lambda(W_i'\hat{\gamma}_n)\right]^2, \tag{1.4}$$

where λ is restricted to be either a decreasing or increasing function. Note that our estimation method utilizes two monotonicity restrictions, i.e., one on the marginal distribution function of latent error ν and the other on the control function λ. Both nonparametric estimates are piece-wise constant functions with implicit window widths automatically determined by the data. Another useful feature resides in the computational simplicity as efficient algorithms are available (Groeneboom and Hendrickx, 2018; Meyer, 2013)[3], so no delicate optimization problems arise in the calculation.

[2]See Remark 3.2 for a detailed comparison between our approach and Cosslett (1991).
[3]The computation algorithms are available in R packages "isotone" and "coneproj".

Within our framework, the presence of the sample selection bias can be formally tested by testing the constancy of the control function λ against a non-constant monotone function. For this purpose, we adapt the likelihood ratio type test[4] of Robertson, Wright, and Dykstra (1988) and Sen and Meyer (2017) to our setting. It is well-known that the null asymptotic distribution of the test statistic is complicated and tabulating the null critical value based on the asymptotic distribution is impractical. We prove that a residual bootstrap procedure approximates the null distribution of the test statistic and our test is consistent against general alternatives. The substantial advantage of our test over the kernel type test (Fan and Li, 1996) developed by Christofides, Li, Liu, and Min (2003) is that our test sidesteps any bandwidth selection, which could involve sophisticated higher-order expansions if the optimal version is desired (Gao and Gijbels, 2008).

Our main contributions are three-fold. First of all, we find a simple sufficient condition for the monotone control function, which is related to an intuitive dependence concept of two latent error terms. This demonstrates the monotonicity of the inverse Mills ratio in the original Heckman model (Heckman, 1974, 1979) is shared by a much larger family without requiring any parametric assumption. Not surprisingly, our framework nests some existing parametric generalizations (Lee, 1983; Marchenko and Genton, 2012) as special cases. Second, our methodology complements the existing semiparametric approaches (Ahn and Powell, 1993; Das, Newey, and Vella, 2003; Newey, 2009; Li and Wooldridge, 2002) in the sense that we develop fully data-driven estimation and inference methods that free applied researchers from choosing any tuning parameter, as long as the monotonicity of the control function is assumed. The aforementioned existing semiparametric approaches do not impose any shape restriction on the control function, but one has to specify bandwidths, orders of polynomials, or trimming sequences in the estimation or testing. We argue that the imposed shape restriction is reasonable in the setting where researchers do have certain prior knowledge regarding the dependence between latent errors. The intuitive dependence relationship is formulated precisely in terms of the right tail increasing or decreasing. Both the Monte Carlo simulation and real data application demonstrate the robust performance of our procedures, whereas the kernel-based approaches are sensitive to the bandwidths selection. An important application of our methodology regards estimating various treatment effects for evaluating policy changes based on the generalized Roy model (Heckman, Tobias, and Vytlacil, 2003; Heckman and Vytlacil, 2005; Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018; Kline and Walters, 2019). Our semiparametric estimates directly deliver tuning-parameter-free estimates of the average treatment effect (ATE), the average treatment effect on the treated (TTE), and the local average

[4]See the test statistic $\bar{E}_{01}^2$ in Chapter 2 of Robertson, Wright, and Dykstra (1988).

treatment effect (LATE) without any parametric assumption on the error terms. Last but not least, from a theoretical perspective, our work also contributes to the literature of two-stage estimation and testing that involves shape restricted nonparametric components. A distinction from Huang (2002) and Cheng (2009) is that the regressor $W'\hat{\gamma}_n$ depends on the estimated coefficient $\hat{\gamma}_n$ from the selection equation so that we have to characterize its asymptotic effect on the estimation of the outcome equation. Unlike the sieve or kernel approach adopted by Newey (2009) or Li and Wooldridge (2002), our estimator for the control function is only a piece-wise constant function with random jump locations determined by the data. As a consequence, the estimated control function not only converges at a slower rate than the kernel or sieve estimator, but also cannot be simply differentiated to determine the asymptotic influence of $\hat{\gamma}_n$ from the first stage estimation. Aiming at those challenges, our proofs that make novel use of the empirical process theory and the characterization of isotonic regression are also of independent interest.

1.1. Related Literature

The joint normality assumption on (ε, ν) in Heckman (1974, 1979) is more for convenience than necessity for the sample selection model. Indeed, the imposition of false distributional assumptions leads to inconsistent estimates and invalid inferences (Arabmazar and Schmidt, 1982), motivating the development of non-normal parametric selection model (Lee, 1983; Marchenko and Genton, 2012) and more flexible semi/non-parametric estimation methods.

Substantial theoretical advances have been made where either a kernel or sieve type of estimator is used to estimate nonparametric components of the selection model (Gallant and Nychka, 1987; Newey, 2009; Robinson, 1988; Andrews, 1991; Ahn and Powell, 1993; Andrews and Schafgans, 1998; Chen and Lee, 1998; Das, Newey, and Vella, 2003). We refer readers to Vella (1998) and Chapter 10 of Li and Racine (2007) for a comprehensive review. The implication is that one must choose a tuning parameter such as the bandwidth in kernel smoothing or the number of sieve base functions, but the optimal choice is not clear and could be quite delicate (Cattanoe, Farrell, and Jansson, 2018) in this context. Inevitably, these methods require a considerable amount of intervention and judgment on the part of the practitioner. Indeed, as noted by Heckman and Vytlacil (2007a)[p.4783], "progress in implementing these procedures in practical empirical problems has been slow and empirical applications of semi-parametric methods have been plagued by issues of sensitivity of estimates to choices of smoothing parameters, trimming parameters, and the like."

One attempt that gives rise to a tuning-parameter-free estimation of the sample selection model is made by Cosslett (1991). In this approach, the profile maximum likelihood

estimation of Cosslett (1983) is used to estimate the selection equation in the first stage. The estimated marginal distribution $\hat{F}_\nu$ is used in the second stage when the partial linear model of the outcome equation is fitted to the nonparametric component approximated by a piece-wise constant function, where the jump locations are taken as those of $\hat{F}_\nu$. Cosslett (1991) proves the consistency of his estimator; however, the corresponding asymptotic distribution and the rate of convergence remain unknown. Our work is inspired by Cosslett (1991), but the key distinction between our approach and his is that we also impose a shape restriction on the control function in the second stage. Building on the recent breakthrough on semiparametric shape-restricted estimation and inference (Groeneboom and Jongbloed, 2014; Groeneboom and Hendrickx, 2018; Baladbaoui, Groeneboom, and Hendrickx, 2017; Sen and Meyer, 2017), we establish the root-n consistency and asymptotic normality of our estimators for finite dimensional parameters.

Starting with Ayer, Brunk, Ewing, Reid, and Silverman (1955) and Grenander (1956), there is a voluminous literature dealing with shape-restricted estimation, and we shall content ourselves to mention only a few references related to our semiparametric models, such as Cosslett (1983), Groeneboom and Wellner (1992), and Groeneboom and Hendrickx (2018), while referring to Groeneboom and Jongbloed (2014) for a comprehensive account. There has also been continued interest in shape-restricted estimation and inference in the econometrics literature, as demonstrated by Matzkin (1991, 1993), Banerjee, Mukherjee, and Mishra (2009), Lee, Tu, and Ullah (2014), and Chernozhukov, Newey, and Santos (2015). Also, see the excellent review provided by Chetverikov, Santos, and Shaikh (2018) for an extensive list of references in econometrics. The synthesis of these works is a shape constraint on the nonparametric component that is suggested by theoretical models or background knowledge. Within the context of sample selection models, Chen and Zhou (2010) and Chen, Zhou, and Ji (2018) make use of another type of shape restriction; namely, a symmetry condition on the control functions, thus eliminating selection bias through the proper matching of propensity scores. However, this matching approach resorts to kernel smoothing, which again depends on a properly chosen kernel bandwidth.

1.2. Organization and Notation

The rest of our paper is organized as follows. Section 2 characterizes a sufficient condition for the monotonicity of the control function. Section 3 proposes an automatic semiparametric estimation method and a new test for the presence of sample selection bias. Section 4 establishes the asymptotic results. Section 5 extends our methodology to the Type-3 Tobit model, the Generalized Roy model, and the panel selection model. Section 6 conducts Monte Carlo simulations. Section 7 applies our method to a real data-set. The last section

concludes. Proofs of main theorems are presented in Appendix A; whereas, proofs of more technical lemmas are delegated to Appendix B. We end this section by introducing some basic notations.

Throughout the paper, we work with the i.i.d. data (Y_i, D_i, X_i, W_i) for $i = 1, ..., n$. It is convenient to introduce the indicator $\bar{D}_i$, defined by $\bar{D}_i = 1 - D_i$ for $i = 1, \cdots, n$. Let p denote the dimensionality of covariates X and write $\beta_0 \equiv (\beta_{01}, \beta_{02}, ..., \beta_{0p})'$. The covariates X do not contain the constant term as the intercept term is absorbed into the control function for identification purposes; see Andrews and Schafgans (1998) and Das, Newey, and Vella (2003). Similarly, we let q denote the dimensionality of covariates W and we write $\gamma_0 \equiv (\gamma_{01}, \gamma_{02}, ..., \gamma_{0q})'$. A normalization by taking $\gamma_{01} = 1$ is adopted, following Ichimura (1993) and Klein and Spady (1993).

We use the standard empirical process notations as follows. For a function $f(\cdot)$ of a random vector $Z = (Y, D, X', W')'$ that follows distribution P, we let $Pf = \int f(z)dP(z), \mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(Z_i)$, and $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n - P)f$. Function f can be replaced by a random function $z \mapsto \hat{f}_n(z; Z_1, \cdots, Z_n)$. Therefore, $P\hat{f}_n = \int f(z; Z_1, \cdots, Z_n)dP(z)$, $\mathbb{P}_n\hat{f}_n = n^{-1}\sum_{i=1}^n f(Z_i; Z_1, \cdots, Z_n)$ and $\mathbb{G}_n\hat{f}_n = n^{1/2}(\mathbb{P}_n - P)\hat{f}_n$.

Regarding the joint distribution of two latent error terms, we denote the copula function of (ε, ν) by $C(\cdot, \cdot)$ and let F_ε and F_ν represent their marginal distribution functions. We write $\bar{F}_\varepsilon$ and $\bar{F}_\nu$ as the corresponding survivor functions. Also, let the survivor copula function be $\hat{C}(u, v) \equiv u + v - 1 + C(1 - u, 1 - v)$. Moreover, the conditional distribution $F_{\varepsilon|\nu>t}(s)$ stands for $\Pr\{\varepsilon \leq s|\nu > t\}$.

2. Monotonicity of the Control Function

The sample selection bias arises when the latent error terms ε and ν in the selection and outcome equations are dependent, leading to a non-constancy control function $\lambda(W'\gamma_0) = \mathbb{E}[\varepsilon|\nu > -W'\gamma_0, W]$. In Heckman's original set up, (ε, ν) has a bivariate normal distribution which gives rise to a monotone control function λ proportional to the inverse Mills ratio. It is interesting to explore whether this is a special shape induced by the joint normality assumption, or it is an omnipresent feature shared by a much larger family without parameterizing the joint distribution of the error terms (ε, ν). In this section, we show that if the latent error terms ν and ε exhibit certain positive (negative) dependence, then the control function is monotonically decreasing (increasing).

Beyond the standard correlation coefficient, there exists a wealth of notions characterizing the positive (negative) dependence between two random variables, as exemplified in the pioneering work of Lehmann (1966). Two popular measures in Lehmann (1966) are the

positive quadrant dependence (PQD)[5] and stochastic increasing (SI)[6]. Complementing the work of Lehmann (1966), Esary and Proschan (1972) proposed the notation of right tail increasing, which is weaker than SI, yet stronger than PQD, and is defined as the following.

Definition 2.1. *A random variable ε is said to be right tail increasing (decreasing) in ν, which we denote as $RTI(\varepsilon|\nu)$ ($RTD(\varepsilon|\nu)$), if $P\{\varepsilon > s|\nu > t\}$ is an increasing (decreasing) function of t for all s.*

The choice between RTI and RTD can be determined in empirical studies since applied researchers do have certain prior knowledge about the sign or direction of the selection bias. To avoid repetition, we focus on $RTI(\varepsilon|\nu)$ since the conditions related to $RTD(\varepsilon|\nu)$ can be stated analogously. RTI is an intuitive positive dependence condition in the sense that ε is more likely to take large values when ν increases. Considering the wage equation where Y^* is the wage offer of an individual and D is the labor supply decision, RTI simply means that those with a higher willingness to work are more likely to earn a higher wage conditional on observed characteristics.

Referring to the benchmark selection model, i.e., the Roy model (Heckman and Honore, 1990; Heckman and Vytlacil, 2007a), the outcomes Y_1 and Y_0 are wages attached to different sectors (or different education levels) that have the following specifications:

$$\begin{aligned} Y_1 &= X'\beta_1 + u_1, \\ Y_0 &= X'\beta_0 + u_0, \end{aligned} \tag{2.1}$$

with an observable switching cost (or price) $C = \tilde{W}'\beta_C$. The decision rule states that the individual self selects into the sector with a higher wage modulo the switching cost:

$$D = \mathbb{I}\{X'(\beta_1 - \beta_0) - \tilde{W}'\beta_C + (u_1 - u_0) > 0\}. \tag{2.2}$$

Suppose one only observes the wage corresponding to sector 1; i.e., $Y = D \times Y_1$. In terms of our notation, we use $W = (X', \tilde{W}')'$, $\gamma_0 = ((\beta_1 - \beta_0)', \beta_C')'$, and $\nu = u_1 - u_0$ in the selection equation. The latent error in the outcome equation for sector 1 is $\varepsilon = u_1$. For this simple Roy model, $RTI(\varepsilon|\nu)$ is an appealing concept since it means that when $u_1 - u_0$ is larger, it is more likely that u_1 is large as well.

An equivalent characterization of RTI is given by the copula function,[7] which we record as the following lemma (see Theorem 5.2.2. in Nelsen (2006)).

[5]Random variables ε and ν are positive quadrant dependent if $P\{\varepsilon > s, \nu > t\} \geq P\{\varepsilon > s\}P\{\nu > t\}$ for any s and t.

[6]The random variable ε is stochastic increasing in ν if $P\{\varepsilon > s|\nu = t\}$ is an increasing function of t for all s.

[7]For sample selection models and generalized Roy models, the copula has been successfully employed to obtain bounds on distributional treatment effects (Abbring and Heckman, 2007; Fan and Wu, 2010; Fan, Guerre, and Zhu, 2017) and aid in identification for non-separable models (Arellano and Bonhomme, 2017).

Lemma 2.1. *We get $RTI(\varepsilon|\nu)$ if and only if*

$$\frac{1-u-v+C(u,v)}{1-v} \quad \textit{is increasing inv;} \tag{2.3}$$

or equivalently, if and only if

$$\frac{u-C(u,v)}{1-v} \quad \textit{is decreasing inv.} \tag{2.4}$$

The main theorem in this section shows that RTI implies that the control function is monotonically decreasing.

Theorem 2.1. *If ε is right tail increasing in ν, then the control function $\lambda(\cdot)$ is monotonically decreasing.*

Proof. The following formula is more convenient for our purpose:

$$\begin{aligned}\mathbb{E}[\varepsilon|\nu>t] &= \int_{-\infty}^{+\infty} s dF_{\varepsilon|\nu>t}(s) \\ &= \int_0^{+\infty} \bar{F}_{\varepsilon|\nu>t}(s)ds - \int_{-\infty}^{0} F_{\varepsilon|\nu>t}(s)ds \\ &= \int_0^{+\infty} \frac{\hat{C}(\bar{F}_\varepsilon(s),\bar{F}_\nu(t))}{\bar{F}_\nu(t)}ds - \int_{-\infty}^{0} \frac{F_\varepsilon(s)-C(F_\varepsilon(s),F_\nu(t))}{1-F_\nu(t)}ds.\end{aligned} \tag{2.5}$$

See Proposition 4.2 in Shorack (2000). We examine the two terms on the right-hand side of (2.5) separately. On one hand, we get

$$\begin{aligned}&\int_0^{+\infty} \frac{\hat{C}(\bar{F}_\varepsilon(s),\bar{F}_\nu(t))}{\bar{F}_\nu(t)}ds \\ &= \int_0^{+\infty} \frac{1-F_\varepsilon(s)-F_\nu(t)+C(F_\varepsilon(s),F_\nu(t))}{1-F_\nu(t)}ds.\end{aligned}$$

Hence, $\int_0^{+\infty} \bar{F}_{\varepsilon|\nu>t}(s)ds$ is an increasing function of t by (2.3).
On the other hand, it is straightforward to see that

$$-\int_{-\infty}^{0} \frac{F_\varepsilon(s)-C(F_\varepsilon(s),F_\nu(t))}{1-F_\nu(t)}ds$$

is again monotonically increasing with respect to v given (2.1). Finally, the control function $\lambda(t)=\mathbb{E}[\varepsilon|\nu>-t,W]$ is monotonically decreasing with respect to t. □

We now provide some parameterized joint distributions of (ε,ν) that yield monotone control functions. For simplicity, we focus on examples with positively dependent pairs ε and ν, generating decreasing control functions $\lambda(\cdot)$.

Example 2.1 (Joint Gaussian Distribution/Gaussian Copula)**.** The original Heckman's model (Heckman, 1974, 1979) under the joint normality assumption on (ε,ν) serves as our

starting point. In this case, the control function has the well-known form depending on the inverse Mill's ratio:

$$\lambda(W'\gamma) = \frac{\rho\sigma_\varepsilon}{\sigma_\nu}\left\{\frac{\phi(W'\gamma)}{\Phi(W'\gamma)}\right\}, \tag{2.6}$$

where ρ is the correlation coefficient and σ_ε, σ_ν stand for the individual standard deviation. If $\rho > 0$, then the control function is decreasing because the inverse Mill's ratio $\frac{\phi(\cdot)}{\Phi(\cdot)}$ is a decreasing function following after the log-concavity of normal distribution (Heckman and Honore, 1990). In fact, the monotonicity property here only depends on the Gaussian copula $C(u, v; \rho) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$ and the sign of its correlation coefficient denoted by ρ . Without restricting the marginal distribution to be Gaussian, Lee (1983) first proposes a generalized selection model with arbitrary (but known) marginal distributions coupled with the Gaussian copula. A straightforward calculation shows that the partial derivative of any Gaussian copula is

$$\frac{\partial}{\partial v}C(u, v; \rho) = \Phi\left(\frac{\Phi^{-1}(u) - \rho\Phi^{-1}(v)}{\sqrt{1-\rho^2}}\right), \tag{2.7}$$

which is an increasing function of v if and only if the correlation coefficient $\rho \geq 0$. Hence, by Theorem 5.2.10 of Nelsen (2006), the non-negative correlation implies the stochastic increasing property, which further implies $RTI(\varepsilon|\nu)$. A complete analog shows that a non-positive correlation leads to $RTD(\varepsilon|\nu)$. In sum, $RTI(\varepsilon|\nu)$ is equivalent to $\rho \geq 0$ in case of the Gaussian copula model.

Example 2.2 (Archimedean Copula)**.** When the copula function is Archimedean, i.e., $C(u, v) = \psi^{[-1]}(\psi(u) + \psi(v))$ with ψ as the generator function. Consider the following cross-ratio function proposed by Oakes (1989):

$$CR(u) = -u\frac{\psi^{(2)}(u)}{\psi^{(1)}(u)} \tag{2.8}$$

for $u \in [0, 1]$, where $\psi^{(j)}$ denotes the j-th order derivative of the generator ψ, for $j = 1, 2$. As shown by Spreeuw (2014), $RTI(\varepsilon|\nu)$ is equivalent to Oakes' cross-ratio function being greater or equal to 1; i.e., $CR(u) \geq 1$ for any u. One popular Archimedean copula is the Clayton copula:

$$C(u, v; \alpha) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad 0 \leq u, v \leq 1, \tag{2.9}$$

where the parameter $\alpha \geq 0$. Its generator function $\psi(u; \alpha) = u^{-\alpha} - 1$. One could easily verify that the cross-ratio function $CR(u) = \alpha + 1$ for any $u \in [0, 1]$, which is always greater or equal to 1. Hence, within the whole Clayton copula family, we have $RTI(\varepsilon|\nu)$.

Example 2.3 (Generalized FGM Copula)**.** A copula function belongs to the generalized Farlie-Gumbel-Morgenstern (FGM) family if $C(u,v;\theta) = uv + \theta\varphi(u)\varphi(v)$ (Amblard and Girard, 2002) with φ as the generator function and θ as the parameter. According to Amblard and Girard (2002), $RTI(\varepsilon|\nu)$ is equivalent to the condition that $\varphi(u)/(u-1)$ is monotone. The original FGM copula specifies $\varphi(u) = u(1-u)$ so that $C(u,v;\theta) = uv + \theta uv(1-u)(1-v)$. Note that $\varphi(u)/(u-1) = u$ in the original FGM, which is indeed monotone; therefore giving rise to $RTI(\varepsilon|\nu)$ in our context.

On some occasions, it is easier to directly verify the monotonicity of the control function than its sufficient condition $RTI(\varepsilon|\nu)$, as shown in the following normal mixture model.

Example 2.4 (A Normal Mixture)**.** Let $g(\cdot,\cdot;\sigma_1,\sigma_2,\rho)$ be the joint density function of the bivariate normal distribution $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}\right)$. Suppose that (ε,ν) is a mixture of two bivariate normals with "small" and " large" variances. The joint distribution is

$$f_{\varepsilon,\nu}(s,t) = \pi g(s,t;\sigma_1,\sigma_2,\rho) + (1-\pi)g(s,t;k\sigma_1,k\sigma_2,\rho), \tag{2.10}$$

where the second normal component has a covariance matrix amplified by $k > 1$. The conditional density of $\varepsilon|\nu = t$ can then be written as

$$\begin{aligned} f_{\varepsilon|\nu}(s|t) &= \frac{\pi g(s,t;\sigma_1,\sigma_2,\rho) + (1-\pi)g(s,t;k\sigma_1,k\sigma_2,\rho)}{\pi\phi(t/\sigma_2)/\sigma_2 + (1-\pi)\phi(t/k\sigma_2)/k\sigma_2} \\ &= \Pi(t;\sigma)\frac{g(s,t;\sigma_1,\sigma_2,\rho)}{\phi(t/\sigma_2)/\sigma_2} + (1-\Pi(t;\sigma))\frac{g(s,t;k\sigma_1,k\sigma_2,\rho)}{\phi(t/k\sigma_2)/k\sigma_2} \\ &= \Pi(t;\sigma)\phi\left(\frac{s-\rho t\sigma_1/\sigma_2}{\sqrt{1-\rho^2}\sigma_1}\right)\frac{1}{\sqrt{1-\rho^2}\sigma_1} + (1-\Pi(t;\sigma))\phi\left(\frac{s-\rho t\sigma_1/\sigma_2}{\sqrt{1-\rho^2}k\sigma_1}\right)\frac{1}{\sqrt{1-\rho^2}k\sigma_1}, \end{aligned}$$

where $\phi(\cdot)$ is the density of the standard normal and $\Pi(t;\sigma)$ is given by

$$\Pi(t;\sigma) \equiv \frac{\pi\phi(t/\sigma_2)/\sigma_2}{\pi\phi(t/\sigma_2)/\sigma_2 + (1-\pi)\phi(t/k\sigma_2)/k\sigma_2}.$$

In other words, $\varepsilon|\nu = t$ is a normal mixture with components $N\left(\rho t\sigma_1/\sigma_2, (1-\rho^2)\sigma_1^2\right)$, $N\left(\rho t\sigma_1/\sigma_2, (1-\rho^2)k^2\sigma_1^2\right)$, and the mixing coefficient $\Pi(t;\sigma)$. As a result, the conditional expectation $E[\varepsilon|\nu = t] = \rho t\sigma_1/\sigma_2$, which is increasing in t as long as $\rho > 0$. Consider any $t < t'$, the following inequality holds for the weighted averages where the conditional expectation is weighted by the density of ν:

$$\frac{\int_t^\infty \mathbb{E}[\varepsilon|\nu = u]f_\nu(u)du}{1-F_\nu(t)} < \frac{\int_{t'}^\infty \mathbb{E}[\varepsilon|\nu = u]f_\nu(u)du}{1-F_\nu(t')},$$

which is the same as $\mathbb{E}[\varepsilon|\nu \geq t] < \mathbb{E}[\varepsilon|\nu \geq t']$. This is equivalent to the control function $\lambda(t) = \mathbb{E}[\varepsilon|\nu \geq -t]$ being monotonically decreasing with respect to t.

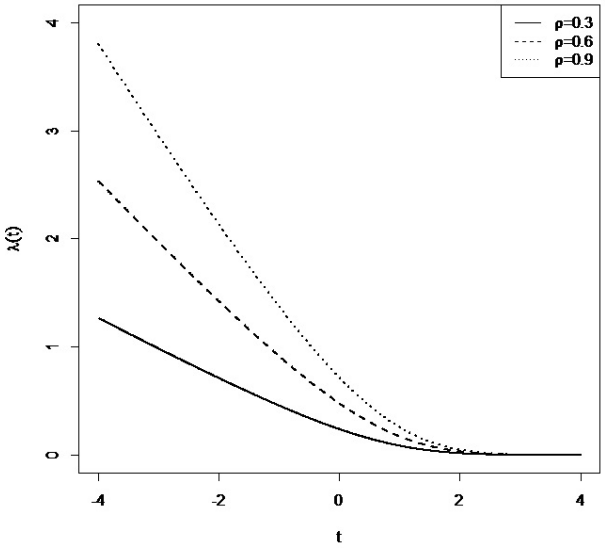
Figure 1 plots the control function $\lambda(t) = \mathbb{E}[\varepsilon|\nu \geq -t]$ based on the previous examples using specific joint distributions of the error terms (ε, ν). Panels (a) and (b) have the same Gaussian copula, but with different marginal distributions: $N(0,1)$ and $t(5)$, respectively, where $t(5)$ denotes a t distribution with the degree of freedom equal to 5. In each panel, three lines represent the control functions coming from Gaussian copulas with different correlations: $\rho = 0.3$, 0.6, and 0.9. The control functions in Panel (a) have the form $\lambda(t) = \rho\phi(t)/\Phi(t)$ and in Panel (b) are $\lambda(t) = \rho\phi(\Phi^{-1} \circ F_\nu(t))/F_\nu(t)$, where $F_\nu(t)$ is the CDF of $t(5)$.

Panels (c) and (d) depict $\lambda(t)$ for joint distributions that have the same $t(5)$ marginal distribution, but with different copulas: the Clayton copula (see Example 2.2, with $\alpha = 1, 5, 15$) and the FGM copula (see Example 2.3, with $\theta = 0.5, 0.75, 1$). Because a FGM copula can only model a relatively weak dependence, the resulting $\lambda(t)$ has limited variations. Panels (e) and (f) show $\lambda(t)$ for the joint distribution described in Example 2.4: a mixture of bivariate normal components with correlation ρ, mixing coefficient π, and standard deviations $\sigma_1 = \sigma_2 = 5$. Panel (e) fixes the mixing coefficient at $\pi = 0.9$ and presents $\lambda(t)$ for the correlation $\rho = 0.3, 0.6$, and 0.9. Panel (f) fixes the correlation $\rho = 0.9$ and varies the value of the mixing coefficients among $\pi = 0.3, 0.6$, and 0.9.

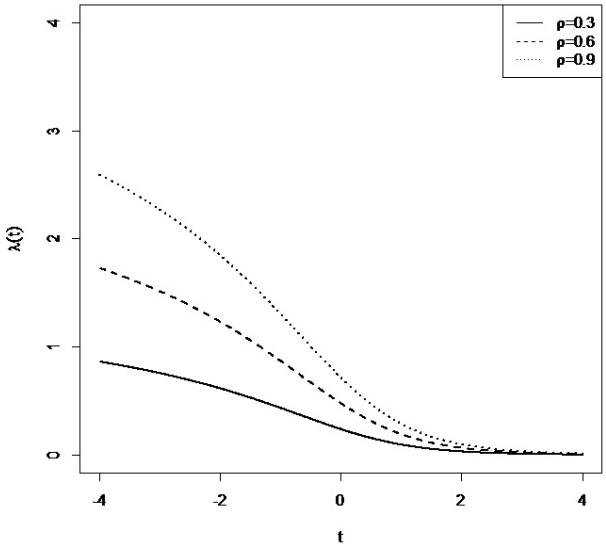
Several interesting observations follow from the exhibited control functions. First, all the control functions depicted in Figure 1 are decreasing by design, yet their shapes substantially differ depending on the marginal distribution or the copula function. For the joint normal case [Panel (a)], the dependence measure (correlation coefficient ρ) only changes the control function proportionally; whereas for other cases, the dependence measure can also affect the shape and curvature of λ. Furthermore, the overall range of dispersion of a control function is related to the range of the dependence measure or parameter in the copula function [compare Panels (c) and (d)]. Namely, the FGM copula is only suitable for modeling moderate dependence, whereas the Clayton copula allows for a much wider range of dependence relationship. Last, but not least, the more curved portion of the control function can be either on the left or right as shown in Panels (e) and (f).

FIGURE 1. Plots of the control function $\lambda(t) = \mathbb{E}[\varepsilon|\nu \geq -t]$ for different joint distributions of (ε, ν).

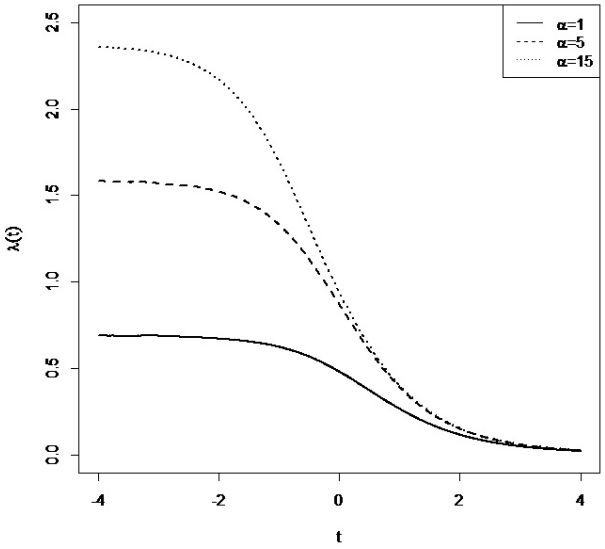
(a) Gaussian copula with correlation ρ. Marginal distributions: $N(0,1)$.



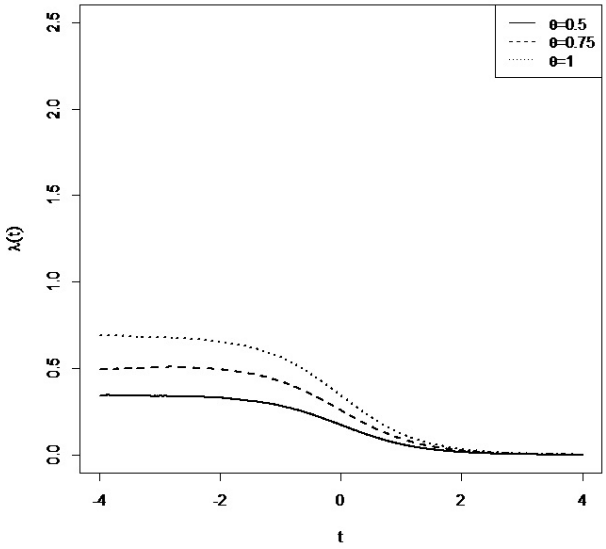
(b) Gaussian copula with correlation ρ. Marginal distributions: $t(5)$.



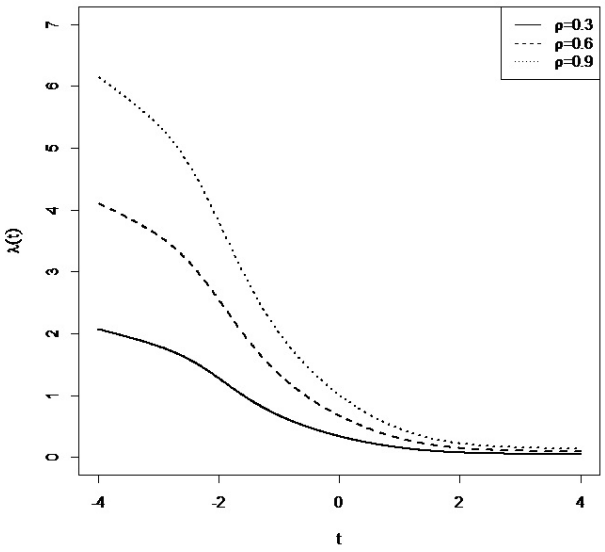
(c) Clayton copula with parameter α. Marginal distributions: $t(5)$.



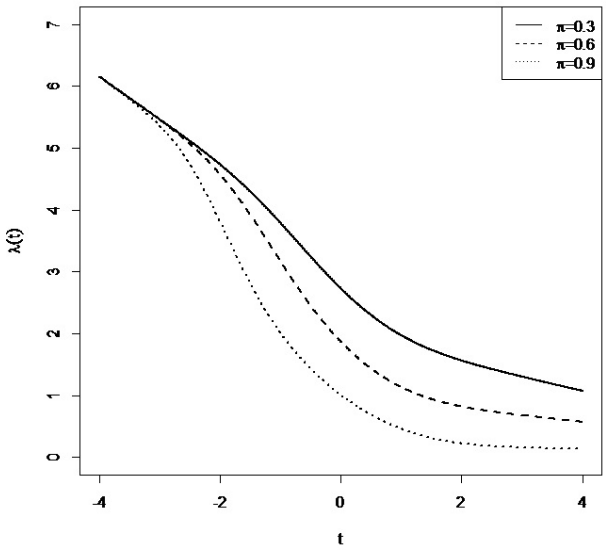
(d) FGM Copula with parameter α. Marginal distributions: $t(5)$.



(e) Normal mixture with correlation ρ. $\sigma_1 = \sigma_2 = 5$; $\pi = 0.9$.



(f) Normal mixture with mixing coefficient π. $\sigma_1 = \sigma_2 = 5$; $\rho = 0.9$.



3. Shape-restricted Estimation and Testing

In this section, we propose a simple two-stage semiparametric estimation method of $(\beta, \lambda(\cdot))$ that does not require any user-specified tuning parameter. We also develop a new sensitivity test for the presence of sample selection bias exploring the shape restricted estimation.

3.1. A Semiparametric Estimator without Tuning Parameters

Our estimation method is inspired by Cosslett (1991) in the sense that we obtain a two-stage semiparametric estimation making use of shape restricted estimation of non-parametric components in the model. The differences are mainly two-fold. First, we adapt the important breakthrough by Groeneboom and Hendrickx (2018) to estimate the linear index in the selection equation, which delivers root-n consistent and asymptotic normal estimators $\hat{\gamma}_n$, unlike the profile maximum likelihood estimator (Cosslett, 1983), which is only known to be consistent. More importantly, we also impose the shape restriction on the control function in the second stage and utilize the isotonic regression technique (Huang, 2002). The detailed procedure is described as follows.

Stage 1(i). For any γ, we compute the NPMLE for $F_\nu(\cdot)$ in the selection equation:

$$\hat{F}_{n\nu}(\cdot;\gamma) = \arg\max_{F} \sum_{i=1}^{n} \left[\bar{D}_i \log F(-W_i'\gamma) + (1-\bar{D}_i)\log(1-F(-W_i'\gamma))\right], \tag{3.1}$$

where $\bar{D}_i \equiv 1 - D_i$. The above optimization problem is well-defined and it generates a piece-wise constant function $\hat{F}_{n\nu}(\cdot;\gamma)$ that can be characterized as follows. Fixing the parameter γ, we consider the values of $\bar{V}_1^{(\gamma)} = -W_1'\gamma, \cdots, \bar{V}_n^{(\gamma)} = -W_n'\gamma$. Let $\bar{V}_{(1)}^{(\gamma)} \leq \cdots \leq \bar{V}_{(n)}^{(\gamma)}$ be the order statistics with corresponding indicators $\bar{D}_i^{(\gamma)}$ for $i = 1, \cdots, n$. Thereafter, $\hat{F}_{n\nu}(\cdot;\gamma)$ is equal to the left derivative of the convex minorant of a cumulative sum diagram consisting of the points $(0,0)$ and

$$\left(i, \sum_{j=1}^{i} \bar{D}_{(j)}^{(\gamma)}\right) \quad \text{for} \quad i = 1, \cdots, n,$$

as in Groeneboom and Hendrickx (2018).

Stage 1(ii). Given $\hat{F}_{n\nu}(\cdot;\gamma)$ at hand, our estimator $\hat{\gamma}_n$ for the regression coefficient is the zero-crossing point of the estimation equation[8]

$$\frac{1}{n}\sum_{i=1}^{n} W_i \left[\bar{D}_i - \hat{F}_{n\nu}(-W_i'\hat{\gamma}_n;\hat{\gamma}_n)\right] = 0. \tag{3.2}$$

Stage 2. Given $\hat{\gamma}_n$, we estimate β and $\lambda(\cdot)$ by the least squares estimator under the monotonicity restriction for λ:

$$(\hat{\beta}_n, \hat{\lambda}_n) = \arg\min_{\beta\in\mathbf{B},\lambda\in\mathcal{D}} \sum_{i=1}^{n} D_i \left[Y_i - X_i'\beta - \lambda(W_i'\hat{\gamma}_n)\right]^2. \tag{3.3}$$

This optimization problem involves minimizing a convex function over a convex set; therefore, $(\hat{\beta}_n, \hat{\lambda}_n)$ exist and are well-defined (Huang, 2002; Meyer, 2013). The efficient single-cone-projection algorithm[9] in Meyer (2013) can be directly applied to obtain $(\hat{\beta}_n, \hat{\lambda}_n)$, which give rises to a monotone piece-wise constant function $\hat{\lambda}_n$ with jump sizes and locations determined by the data.

Now we provide a heuristic discussion of each step. The first stage NPMLE $\hat{F}_{n\nu}(\cdot;\gamma)$ and its characterization date back to Ayer, Brunk, Ewing, Reid, and Silverman (1955) in analyzing current status data (Groeneboom and Wellner, 1992). Within the context of binary choices models, the NPMLE is utilized by Cosslett (1983) to define the profile maximum likelihood estimator. However, only consistency results are available for Cosslett's estimator given the challenge that the estimated error distribution is neither linear nor smooth. The key to developing a root-n consistent and asymptotic normal estimator for

[8]As $\hat{F}_{n\nu}(\cdot;\hat{\gamma}_n)$ is a step function, the estimating equation here may not hold exactly. Therefore, one needs to search for the zero-crossing point as outlined in Groeneboom and Hendrickx (2018).

[9]This algorithm is available in the R package "coneproj" (Liao and Meyer, 2014).

β_0 while also maintaining the tuning-parameter-free feature is in Stage 1 (ii); we adapt the Z-estimator from Groeneboom and Hendrickx (2018). Modulo the estimated latent distribution function, one makes use of the population level moment condition

$$\mathbb{E}[W(\bar{D} - F_\nu(-W'\gamma_0))] = 0, \tag{3.4}$$

and plug in the first-step estimator $\hat{F}_{n\nu}(\cdot;\gamma)$ in the sample analog[10]. Referring to the second stage assuming a monotone control function, it becomes straightforward to run the isotonic regression after the inclusion of $W'\hat{\gamma}_n$ to control for the endogeneity.

Remark 3.1. *We highlight a connection of our method with the sieve/series type estimator in Das, Newey, and Vella (2003) and Newey (2009). When the control function is within a nice functional class that can be approximated by sieves, it is natural to consider the approximation* $\lambda_n(\cdot) = \sum_{j=1}^{K_n} b_j P_j(\cdot)$*, where* $P_1(\cdot), \cdots, P_{K_n}(\cdot)$ *are basis functions in the sieve space. Given a user-specified* K_n*, the coefficients* $b_1, \cdots, b_{K_n}$ *can be obtained from the least squares estimation, so the resulting sieve estimator is* $\tilde{\lambda}_n(W'\hat{\gamma}_n) = \sum_{j=1}^{K_n} \tilde{b}_j P_j(W'\hat{\gamma}_n)$ *with estimated* $\tilde{b}_1, \cdots, \tilde{b}_{K_n}$*. It turns out that our monotonic estimator* $\hat{\lambda}_n$ *can also be expanded in terms of certain basis as noted by Meyer (2013). First of all,* $\hat{\lambda}_n$ *is a piece-wise constant function with possible jumps at observed* $W_i'\hat{\gamma}_n$ *for* $i = 1, \cdots, n$*. Denote the vector* $\boldsymbol{\lambda}_n = (\hat{\lambda}_n(W_1'\hat{\gamma}_n), \cdots, \hat{\lambda}_n(W_n'\hat{\gamma}_n))'$*. This vector belongs to a convex cone; i.e.,* $\boldsymbol{\lambda}_n \in \boldsymbol{\Lambda}$*. Proposition 2.2 in Meyer (2013) shows that* $\boldsymbol{\lambda}_n = \sum_{j=1}^{K_0} \hat{b}_j \mathbf{e}_j$ *where* $K_0 + 1$ *is the number of distinct values of* $\hat{\lambda}_n$ *and* $\mathbf{e}_j$ *are edges of the cone* $\boldsymbol{\Lambda}$*. Hence, there are two main differences between our approach and the sieve method. First, the number of terms* K_0 *is determined by the data itself and is not chosen by practitioners. Second, the basis terms are formed by edges of a cone associated with the shape restriction rather than smooth functions.*

Remark 3.2. *Cosslett (1991) has proposed an ingenious two-step procedure in which no tuning parameter is needed. He first estimates* γ_0 *and* $F_{\nu 0}(\cdot)$ *by the profile maximum likelihood estimator defined in Cosslett (1983). Note that the resulting estimators* $\tilde{\gamma}_n$ *and* $\tilde{F}_{n\nu}(\cdot)$ *are different from the ones in Groeneboom and Hendrickx (2018) that we adopt in our first stage. The estimated marginal distribution function* $\tilde{F}_{n\nu}(\cdot)$ *is a step-wise function that is constant on a finite number* K_n *of intervals* $I_j = [c_{i-1}, c_j)$*, for* $j = 1, ..., K_n$ *and* $c_0 = -\infty, c_{K_n} = +\infty$*. In the second stage, Cosslett (1991) estimates the outcome equation while approximating the control function* $\lambda(\cdot)$ *by* K_n *indicator variables* $\{\mathbb{I}(W'\tilde{\gamma}_n \in I_j)\}_{j=1}^{K_n}$*. Only consistency results are derived for all estimates in Cosslett (1991) based on a sample-splitting argument. The most important distinction of our method is that we impose the*

[10]The main improvement made by Groeneboom and Hendrickx (2018) over Cosslett (1983) to restore standard distributional theory for γ is that one does not need the error's density function in the moment condition (3.4). In contrast, one has to handle the error density in the likelihood based estimation appearing in the score function, whereas the NPMLE $\hat{F}_{n\nu}(\cdot;\gamma)$ itself is not differentiable.

monotonicity restriction on the control function $\lambda(\cdot)$. Although our estimated $\hat{\lambda}_n(\cdot)$ is also a piece-wise function, it is monotone and the jump locations are determined by the second stage estimation. In contrast, the estimated control function in Cosslett (1991) is not necessarily monotone and its jump locations are determined by the first stage estimation. The major theoretical improvement of our approach over Cosslett (1991) is that we obtain root-n consistent and asymptotically normal estimators for the finite dimensional parameters.

3.2. A Shape-restricted Test for Selectivity

Under the null hypothesis of no selectivity bias, Heckman (1979) proposes a t-test on the regression coefficient associated with the inverse Mill's ratio, assuming joint normality of the latent error terms. Melino (1982) shows that the t-test in Heckman (1979) is the Lagrange multiplier test statistic, which inherits all the optimal properties in this context; also see Vella (1998).

Within our framework, one does not face selection bias if the control function λ_0 is constant, whereas it becomes a non-constant decreasing (increasing) function in the presence of selection bias. Building on this idea, we develop a new test to detect the sample selection. To focus on the main idea, we consider the case where one has a decreasing control function λ_0. The cases with increasing control functions can be dealt with analogously. Let $\mathcal{D}$ be the space of decreasing functions and $\mathcal{C}$ be the space of constant functions for λ_0. The null hypothesis is H_0: $\lambda_0 \in \mathcal{C}$ and the alternative is H_1: $\lambda_0 \in \mathcal{D} \setminus \mathcal{C}$.

The following notations facilitate our presentation. Denote $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ and $\mathbf{X}$ as the $n \times p$ matrix of covariates in the outcome equation. Let $\mathcal{X}$ be the linear space spanned by the column vectors of $\mathbf{X}$. The testing for selectivity regards the conditional mean function $\mathbb{E}[Y|D = 1, X, W]$. We write the null space as $\mathcal{S}_0 = \mathcal{X} + \mathcal{C}$ and the alternative space as $\mathcal{S}_1 = \mathcal{X} + \mathcal{D}$. For any vector $\mathbf{Y} = (Y_1, \cdots, Y_n)'$, define the following norm $\| \mathbf{Y} \|_{n,D}$ as $\sqrt{\sum_{i=1}^n D_i (Y_i)^2}$. Given the norm $\| \cdot \|_{n,D}$, we write $\Pi(\mathbf{Y}|\mathcal{S}_j)$ as the projection of $\mathbf{Y}$ on the null and alternative spaces for $j = 0, 1$, respectively.[11]

Our test statistic is inspired by the likelihood ratio type test in Robertson, Wright, and Dykstra (1988) and it compares the sum of squared residuals under the null and alternative hypotheses:

$$T_n = \frac{\| \Pi(\mathbf{Y}|\mathcal{S}_0) - \Pi(\mathbf{Y}|\mathcal{S}_{1,\hat{\gamma}_n}) \|^2_{n,D}}{\| \mathbf{Y} - \Pi(\mathbf{Y}|\mathcal{S}_0) \|^2_{n,D}}, \tag{3.5}$$

[11]Considering the norm $\| \cdot \|_{n,D}$, only those observations in the selection subsample matter; i.e., the values of Y_i where its corresponding $D_i = 1$. Therefore, the projection $\Pi(\mathbf{Y}|\mathcal{S}_j)$ only depends on the observed dependent variables Y_i for which $D_i = 1$ and the coordinate values for which $D_i = 0$ can be defined arbitrarily. Similar remarks apply to $\Pi(\boldsymbol{\epsilon}|\mathcal{S}_j)$ for $j = 0, 1$ in Section 4.2.

where the additional subscript $\hat{\gamma}_n$ on the space $\mathcal{S}_1$ signifies the fact that the linear index $v = w'\gamma_0$ has to be estimated by $w'\hat{\gamma}_n$. Note that under the null hypothesis, the residual term $\mathbf{Y} - \Pi(\mathbf{Y}|\mathcal{S}_0)$ is simply the residual term from the ordinary leasts square (OLS) estimation over the subsample with $D = 1$.

The asymptotic distribution of T_n under the null hypothesis is very complicated (see Section 2.3 of Robertson, Wright, and Dykstra (1988)). The recent breakthrough by Sen and Meyer (2017) shows that the null critical value for this type of test statistic can be approximated by the bootstrap method. Considering the sample selection model, because the control function boils down to a constant term under H_0, a centered residual bootstrap suffices. Let $\mathcal{A}_n \equiv \{i = 1, 2, ...n : D_i = 1\}$ and $n_1 \equiv \sum_{i \in \mathcal{A}_n} D_i$. Let $\hat{\epsilon}_i, i \in \mathcal{A}_n$ be the OLS residual obtained from regressing Y_i on the constant term and covariates X_i for the subsample with $D_i = 1$, and $\bar{\epsilon}_n = \sum_{i \in \mathcal{A}_n} \hat{\epsilon}_i / n_1$. In each bootstrap sample ($b = 1, 2, ...B$), one obtains $\epsilon^*_{i,b}$ for $i \in \mathcal{A}$ by re-sampling the centered residuals $\hat{\epsilon}_i - \bar{\epsilon}_n$. One then generates $Y^*_{i,b} = \tilde{\alpha}_n + X_i'\tilde{\beta}_n + \epsilon^*_{i,b}$ for $i \in \mathcal{A}_n$, where $\tilde{\alpha}_n$ and $\tilde{\beta}_n$ denote the OLS estimate for the intercept and slope coefficient, respectively. Finally, by letting $\mathbf{Y}^*_b = (Y^*_{1,b}, \cdots, Y^*_{n,b})'$, the bootstrap version of our test statistic is

$$T^*_{n,b} = \frac{\| \Pi(\mathbf{Y}^*_b|\mathcal{S}_0) - \Pi(\mathbf{Y}^*_b|\mathcal{S}_{1,\hat{\gamma}_n}) \|^2_{n,D}}{\| \mathbf{Y}^*_b - \Pi(\mathbf{Y}^*_b|\mathcal{S}_0) \|^2_{n,D}}. \tag{3.6}$$

One can easily repeat the above process B times and obtain the desired critical value by tabulating $(T^*_{n1}, \cdots, T^*_{nB})$.

4. Main Results

In this section, we establish root-n consistency and the asymptotic normality of our estimator of $\hat{\gamma}_n$ and $\hat{\beta}_n$. The nonparametric estimates for λ_0 and $F_{\nu 0}$ converge at the cubic root rate (modulo some $\log n$ term). We also justify the bootstrap procedure in Section 3 to approximate the null sampling distribution and show the consistency of our test.

4.1. Asymptotic Properties of the Semiparametric Estimation

We start with some preliminary notations borrowed from Newey (2009). Denote $V_i = W_i'\gamma_0$ and

$$U_i = D_i(X_i - \mathbb{E}[X_i|D_i = 1, V_i]). \tag{4.1}$$

We assume $H_\beta \equiv \mathbb{E}[U_i U_i']$ is non-singular. Moreover, we define the centered error term as

$$\epsilon_i = D_i(Y_i - X_i'\beta_0 - \lambda_0(V_i)) \tag{4.2}$$

with $\Sigma \equiv \mathbb{E}[\epsilon_i^2 U_i U_i']$ and $H_\gamma \equiv \mathbb{E}[U_i \frac{\partial \lambda_0(v_i)}{\partial v_i} W_i]$. Regarding the first-stage estimation, the NPMLE $\hat{F}_{n\nu}$ in Cosslett (1983) provides an estimate of

$$F_\nu(u;\gamma) \equiv P\left\{\bar{D}^{(\gamma)} | -V^{(\gamma)} = u\right\} = \int F_{\nu 0}(u - w'(\gamma_0 - \gamma)) f_{W|W'\gamma}(w| - W'\gamma = u) dw, \tag{4.3}$$

for any fixed γ; see Groeneboom and Hendrickx (2018). In the sequel, we also denote its density by $f_\nu(u;\gamma)$¿ Short-hand notations such as $F_{\nu 0}(u)$ and $f_{\nu 0}(u)$ are used for $F_\nu(u;\gamma_0)$ and $f_\nu(u;\gamma_0)$ in case where one plugs in the true γ_0.

The following regularity conditions will be assumed throughout the paper.

Condition 1. We assume both Y and X have sub-exponential tails, i.e., there exists some finite constant terms M and σ_0 such that

$$2M^2\left(\mathbb{E}[e^{|Y|/M}] - 1 - \mathbb{E}|Y|/M\right) \leq \sigma_0^2 \tag{4.4}$$

and

$$2M^2\left(\mathbb{E}[e^{|X|/M}] - 1 - \mathbb{E}|X|/M\right) \leq \sigma_0^2. \tag{4.5}$$

Condition 2. The latent error terms (ε, ν) are independent of (X, W).

Condition 3. There exists a local neighborhood $\mathcal{N}_0$ around γ_0 such that for any $\gamma \in \mathcal{N}_0$, $W'\gamma$ is a non-degenerate random variable conditional on X.

Condition 4. The true regression parameters β_0 and γ_0 belong to the interior of some compact sets in $\mathcal{R}^p$ and $\mathcal{R}^q$, respectively.

Condition 5. The true monotone control function λ_0 is continuously differentiable with its derivative denoted by $\dot{\lambda}(\cdot)$. Moreover, its inverse denoted by $\lambda_0^{-1}(\cdot)$ is globally Lipschitz continuous.

Condition 6. The function $F_\nu(\cdot;\gamma)$ has a strictly positive continuous derivative which stays away from zero for all γ in the parameter space. Moreover, the function $F_\nu(u;\gamma)$ is twice continuously differentiable with respect to u on the interior of its support for all γ in the parameter space.

Condition 7. The probability $\Pr\{D = 1\}$ is bounded away from zero.

Condition 8. The density $f_\nu(u;\gamma)$ and conditional expectations $\mathbb{E}[W|W'\gamma = u]$ and $E[WW'|W'\gamma = u]$ are twice continuously differentiable w.r.t. u. The functions $\gamma \mapsto$

$f_\nu(u;\gamma)$, $\gamma \mapsto \mathbb{E}[W|W'\gamma = u]$ and $\gamma \mapsto \mathbb{E}[WW'|W'\gamma = u]$ are continuous functions for u in the definition domain and all γ in the parameter space. The support of W is compact.

Condition 9. The conditional mean function $\chi(u) \equiv \mathbb{E}[X|D = 1, W'\gamma_0 = u]$ is globally Lipschitz continuous, i.e., for any u_1, u_2, one has

$$|\chi(u_1) - \chi(u_2)| \leq L|u_1 - u_2|, \tag{4.6}$$

for some positive finite constant L. The matrix $E[XX'|D = 1]$ is of full rank.

The assumptions are standard and adapted from Ichimura (1993), Klein and Spady (1993), Huang (2002), Heckman and Vytlacil (2007b), Groeneboom and Hendrickx (2018), and Newey (2009). The only condition that we want to emphasize concerns the exclusion restriction of W in Condition (3). Namely, we strengthen the identification condition (A-2) in Heckman and Vytlacil (2007b) to ensure that any linear combination $W'\gamma$ is a non-degenerate random variable conditional on X for γ in a local neighborhood $\mathcal{N}_0$ around γ_0, not just for the true linear index $W'\gamma_0$. Recall the estimated $\hat{\lambda}_n$ is not differentiable, so this technical requirement is needed to obtain the consistency and convergence rates for the parameters in the outcome equation given the first stage estimate $\hat{\gamma}_n$; see the details in our proof of Lemma (10.9). For empirical applications, the distinction is rather minor, because X_i variables are typically a strict subset of W_i and there are additional independent variables in W_i altering the selection equation without affecting the outcome equation.

We define two matrices appearing in the asymptotic covariance matrix of the estimator in Groeneboom and Hendrickx (2018) as follows:

$$A = \mathbb{E}\left[f_{\nu 0}(-W'\gamma_0)\left\{W - E[W|W'\gamma_0]\right\}^{\otimes 2}\right] \tag{4.7}$$

and

$$B = \mathbb{E}\left[\left\{(F_{\nu 0}(-W'\gamma_0) - \bar{D})(W - E[W|W'\gamma_0])\right\}^{\otimes 2}\right]. \tag{4.8}$$

The following lemma regarding the asymptotic analysis of $\hat{\gamma}_n$ and $\hat{F}_{n\nu}(\cdot;\gamma)$ is directly form Groeneboom and Hendrickx (2018).

Lemma 4.1. *Under Conditions 1 to 9, $\hat{\gamma}_n$ is root-n consistent and asymptotically normal.*

$$n^{1/2}\left(\hat{\gamma}_n - \gamma_0\right) \Rightarrow \mathbb{N}(0, V_\gamma), \tag{4.9}$$

where V_γ is equal to $A^{-1}BA^{-1}$. Regarding the latent error distribution, one gets the following cubic rate uniform convergence (modulo the logarithm factor):

$$\sup_u \left|\hat{F}_{n\nu}(u;\hat{\gamma}_n) - F_{\nu 0}(u)\right| = O_p(\log n \times n^{-1/3}). \tag{4.10}$$

Our first main theorem in this section shows the consistency of $(\hat{\beta}_n, \hat{\lambda}_n)$ and gives a crude yet fast enough rate to establish the asymptotic normality in Theorem (4.2). For the nonparametric component, we use the following L_2 norm to metrize its convergence:

$$\| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \|^2 \equiv \int \left(\hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0)\right)^2 f_{W|D=1}(w)dw, \tag{4.11}$$

where $f_{W|D=1}(\cdot)$ is the conditional density of W given $D = 1$.

Theorem 4.1. *Suppose Conditions 1 to 9 hold, then one has*

$$|\hat{\beta}_n - \beta_0| + \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| = O_p(n^{-1/3} \log n). \tag{4.12}$$

The preceding result regarding the convergence of the control function is stated depending on the estimated $\hat{\gamma}_n$. The next statement decouples $\hat{\lambda}_n$ and $\hat{\gamma}_n$ and it implies the uniform convergence of $\hat{\lambda}_n$ to λ_0 over any compact set within the interior of the support.

Lemma 4.2. *Assume the conditional density function of W given $D = 1$ is uniformly bounded from below by a positive constant $\underline{q}$ in its support. Let $[\underline{v}, \bar{v}]$ denote the support of $V = W'\gamma_0$, then*

$$\left(\int_{\underline{v}+\omega_n}^{\bar{v}-\omega_n} \left(\hat{\lambda}_n(v) - \lambda_0(v)\right)^2 dv \right)^{1/2} = O_p(n^{-1/3} \log n) \tag{4.13}$$

for all sequence ω_n such that $n^{1/2}\omega_n \to \infty$ and $\underline{v} + \omega_n \leq \bar{v} - \omega_n$.

Remark 4.1. *There are general results on establishing consistency and rate of convergence for two-step semiparametric estimation methods [Chen, Linton, and Van Keilegom (2003); Chen, Lee, and Sung (2014)], however, these results are not directly applicable to our scenario mainly because the estimated control function is not smooth. Specifically, Theorem 2 in Chen, Linton, and Van Keilegom (2003) focuses on the case where the second stage estimates converge at the root-n rate. Furthermore, since our estimated control function is not differentiable and cannot be directly separated from the first stage estimation, the Condition (B.4) in Lemma B.1 of Chen, Lee, and Sung (2014) is hard to verify in our context. To exemplify the challenge from a different perspective, the consistency proof in Cosslett (1991) relies on the sample-splitting trick in which the selection equation and outcome equation are estimated using separate subsamples. A rigorous proof based on the full sample is absent in Cosslett (1991).*

The large sample property of $\hat{\beta}_n$ is more complicated and is our main focus. Unlike the setup in Newey (2009) or Li and Wooldridge (2002), where the nonparametric control function is subject to certain smoothness restriction, the control function is estimated utilizing the monotonicity restriction in the outcome equation for our model. As a consequence, the

estimated control function $\hat{\lambda}_n(\cdot)$ is piece-wise constant with random jump locations and it is not differentiable. The crux of our proof is to determine the asymptotic contribution of the estimated $\hat{\gamma}_n$ to $\hat{\beta}_n$ based on the characterization of the isotonic regression for partial linear models (Huang, 2002; Mammen and Yu, 2007; Cheng, 2009)] and the empirical process theory (Groeneboom and Hendrickx, 2018; Baladbaoui, Durot, and Jankowski, 2016; Baladbaoui, Groeneboom, and Hendrickx, 2017)].

Theorem 4.2 (Asymptotic Normality)**.** *Suppose Conditions 1 to 9 hold, then we get*

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \Rightarrow \mathbb{N}(0, V_\beta), \tag{4.14}$$

where

$$V_\beta \equiv H_\beta^{-1}\left(\Sigma + H_\gamma V_\gamma H_\gamma'\right) H_\beta^{-1}$$

and V_γ is the asymptotic covariance matrix for $\hat{\gamma}_n$ in Lemma 4.1.

Remark 4.2. *The asymptotic variance matrix for $\hat{\beta}_n$ takes the generic form of two-step estimator in Newey (2009). The first part, $H_\beta^{-1}\Sigma H_\beta^{-1}$, is the asymptotic covariance of an oracle estimator assuming that γ_0 is known; whereas $H_\beta^{-1}H_\gamma V_\gamma H_\gamma' H_\beta^{-1}$ captures the effect from estimating γ_0 in the first stage. Given the additive structure of V_β, a more efficient estimator for γ_0 in the selection equation would improve the performance of $\hat{\beta}_n$. In our approach, the Groeneboom and Hendrickx (2018) estimator is not as efficient as the one in Klein and Spady (1993). However, the advantage is that one avoids picking any tuning parameter by an ad-hoc method.*

Remark 4.3. *A close examination of our proof reveals that only root-n consistency of $\hat{\gamma}_n$ is needed in deriving the asymptotic properties of $\hat{\beta}_n$ and $\hat{\lambda}_n$. In the first stage estimation of the selection equation, the maximum rank correlation estimator of Han (1987) can be used for the coefficient γ_0, which is also tuning-parameter-free. Our preference is mainly driven by two concerns. First, the computational cost associated with the maximum rank correlation estimator is quite non-trivial, because one has to maximize over the indicator functions. Second, Han (1987) sidesteps the estimation of the marginal distribution $F_\nu(\cdot)$, which is needed in estimating the treatment effect of treated when applied to the generalized Roy model (Heckman, 1990); see the discussion in Section 5.2 of this paper.*

4.2. Validity of the Semiparametric Test

We show the validity of the bootstrap inference procedure described in Section 3. Let H_n be the distribution function of T_n and H_n^* be the (conditional) distribution function

of $T^*_{n,b}$ given the observations $(Y_i, D_i, X_i', Z_i')_{i=1}^n$. Furthermore, we define the vector $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)'$.

Theorem 4.3. *Assume Conditions 1 to 9 hold. Let d_L denote the Levy distance between two distribution functions. Also, suppose the sequence*

$$\mathbb{E}[n_1 \parallel \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \parallel_{n,D}^{-2}] < +\infty, \tag{4.15}$$

then we have

$$d_L(H_n, H_n^*) \to 0 \quad a.s. \tag{4.16}$$

Remark 4.4. *The bound in equation (4.15) is from Theorem 1 in Sen and Meyer (2017). They state it as a high-level assumption. Note that the equation (4.15) is imposed to ensure the existence of $\mathbb{E}[(\boldsymbol{\epsilon}'Q\boldsymbol{\epsilon})^{-1}]$ for some idempotent matrix Q with rank equal to $n_1 - (p+1)$. When the error terms $\boldsymbol{\epsilon}$ follow a normal distribution, one can resort to Lemma 2 in Chapter 2 of Ullah (2004), which requires $n_1 - (p+1) > 4$. For general cases where the distribution of $\boldsymbol{\epsilon}$ belongs to the exponential family, analogous conditions can be found in Section 2.3 or 2.4 of Ullah (2004).*

A direct consequence of the above theorem is the validity of using bootstrap critical value (Lemma 23.3 in Van Der Vaart (1998)). The lower p-th quantile of bootstrap distribution is denoted by the quantity c_{np}.

Corollary 4.1. *Under the null hypothesis, for any $\alpha \in (0,1)$, we have*

$$\mathbb{P}_{\lambda_0}\{T_n > c_{n,1-\alpha}\} \to \alpha \tag{4.17}$$

as $n \to \infty$.

We analyze the power property of our test against the alternative hypothesis $H_1 : \lambda_0 \in \mathcal{D} \setminus \mathcal{C}$. To facilitate the presentation, we denote $\boldsymbol{\xi} \equiv (\xi_1, \cdots, \xi_n)' \equiv (X_1'\beta + \lambda(W_1'\gamma), \cdots, X_n'\beta + \lambda(W_n'\gamma))'$. Let the projections to the null and alternative spaces be $\boldsymbol{\xi}_{\mathcal{S}_0}$ and $\boldsymbol{\xi}_{\mathcal{S}_1}$, respectively. To highlight the asymptotic framework, we explicitly denote the dependence on the sample size n of the quantities involved so that we write $\lambda_{0,n}$.

Theorem 4.4. *For any sequence $\{\lambda_{0,n}\} \in \mathcal{D} \setminus \mathcal{C}$, if the following conditions hold:*

$$\lim_{n\to\infty} \frac{\parallel \boldsymbol{\xi}_{\mathcal{S}_0} - \boldsymbol{\xi}_{\mathcal{S}_1} \parallel_{n,D}^2}{n} = c \tag{4.18}$$

and

$$\lim_{n\to\infty} \frac{\parallel \mathbf{Y} - \boldsymbol{\xi}_{\mathcal{S}_0} \parallel_{n,D}^2}{n} = \sigma^2 \tag{4.19}$$

for some positive constant c and σ^2, then

$$\mathbb{P}_{\lambda_{0,n}}\{T_n > c_{n,1-\alpha}\} \to 1 \tag{4.20}$$

as $n \to \infty$.

Remark 4.5. *When the control function is constant, the isotonic estimator is still consistent. In fact, the rate of convergence is almost close to the parametric root-n rate as shown by Zhang (2002) when the underlying function is (piece-wise) constant, leading to $T_n = o_p(1)$ under the null hypothesis. On the other hand, the T_n is bounded away from zero under the alternative hypothesis for functions deviating from constant in a non-trivial way. The latter condition is formalized by equation (4.18), which is also needed in studying the power properties of related tests in Sen and Meyer (2017).*

5. Extensions

In this section, we discuss three different extensions of our proposed methodology.

5.1. A Type-3 Tobit Model

Our framework can be easily extended to the Type-3 Tobit model (Amemiya, 1984) where the selection equation involves a censored dependent variable rather than a binary choice. The model consists of the following two equations for the latent dependent variables:

$$\begin{aligned} Y_i^* &= X_i'\beta_0 + \varepsilon_i; \\ T_i^* &= W_i'\gamma_0 + \nu_i. \end{aligned} \tag{5.1}$$

One observes the censored dependent variable $T_i = \max\{T_i^*, 0\}$ and the indicator $D_i \equiv \mathbb{I}\{T_i^* > 0\}$ from the selection equation. Furthermore, the dependent variable from the outcome equation is only observed when the censored variable is positive; i.e., $Y_i = Y_i^* D_i$, from the outcome equation for $i = 1, \cdots, n$. In a typical labor economics application, $\max\{T_i^*, 0\}$ represents the working hours for the $i-$th worker, whereas Y_i denotes the (log-)wage if he/she is indeed working. In contrast to the standard sample selection model (1.1), one observes working hours when it is positive, whereas in the model (1.1) one only knows whether working hours are positive or zero. Many ingenious semiparametric methods have been proposed for estimating the Type-3 Tobit model, including Powell (1987), (Ahn and Powell, 1993; Lee, 1994; Chen, 1997; Honoré, Kyriazidou, and Udry, 1997; Li and Wooldridge, 2002), among others. It is worthwhile to note that under certain symmetry conditions, the methods by Chen (1997) and Honoré, Kyriazidou, and Udry (1997) are

tuning-parameter-free and do not require the exclusion restriction in the selection equation; i.e., one can take $X = W$.

Our approach complements the aforementioned works in the case where a shape-restricted control function is incorporated into the model (5.1). Since the conditional mean function of the observed dependent variable Y has the following form:

$$\mathbb{E}[Y|X, W, D = 1] = X'\beta_0 + \lambda_0(W'\gamma_0), \tag{5.2}$$

our estimation and testing procedure is directly applicable if one only utilizes the binary choice data (D_i, W_i) in the first stage. However, one could also modify our first step as any other Tobit type estimator can be used to deliver a tuning-parameter-free and root-n consistent estimator $\hat{\gamma}_n$, like the censored quantile regression estimator in Powell (1987). Given $\hat{\gamma}_n$, we estimate β_0 and $\lambda_0(\cdot)$ by

$$(\hat{\beta}_n, \hat{\lambda}_n) = \arg \min_{\beta \in \mathbf{B}, \lambda \in \mathcal{D}} \sum_{i=1}^{n} D_i \left[Y_i - X_i'\beta - \lambda(W_i'\hat{\gamma}_n) \right]^2, \tag{5.3}$$

under the monotonicity restriction such that λ belongs to the space of decreasing functions $\mathcal{D}$.

5.2. A Generalized Roy Model

An important feature of a sample selection model is its use for evaluating potential outcomes and various treatment effects with the corresponding policy implications (Heckman and Vytlacil, 2007a). We consider the generalized Roy model (or the Type-5 Tobit model in Amemiya (1984)) where the treatment outcome $Y(1)$, control outcome $Y(0)$, and the treatment status D are specified by

$$\begin{aligned} Y_i(1) &= X'\beta_{0,1} + \varepsilon_{1i}, \\ Y_i(0) &= X'\beta_{0,0} + \varepsilon_{0i}, \\ D_i &= \mathbb{I}\{W_i'\gamma_0 + \nu_i > 0\}. \end{aligned} \tag{5.4}$$

However, one only observes (Y_i, D_i, X_i, W_i) with $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ for $i = 1, \cdots, n$. Since the conditional mean functions of the observed dependent variables are

$$\mathbb{E}[Y(1)|X, W, D = 1] = X'\beta_{0,1} + \lambda_{0,1}(W'\gamma_0), \tag{5.5}$$

$$\mathbb{E}[Y(0)|X, W, D = 0] = X'\beta_{0,0} + \lambda_{0,0}(W'\gamma_0), \tag{5.6}$$

with control functions $\lambda_{0,1}$ and $\lambda_{0,0}$, it is straightforward to apply the two-step estimation separately for the treatment and control groups (Amemiya, 1984).

In the program evaluation, researchers are mainly interested in the average treatment effect (given $X = x$):

$$ATE(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = x'(\beta_{0,1} - \beta_{0,0}), \tag{5.7}$$

and the average treatment effect on the treated (given $X = x$ and $W = w$):

$$\begin{aligned} TTE(x,w) =& \mathbb{E}[Y(1) - Y(0)|D = 1, X = x, W = w] \\ &= x'(\beta_{0,1} - \beta_{0,0}) + \mathbb{E}[\varepsilon_1 - \varepsilon_0|D = 1, X = x, W = w]; \end{aligned} \tag{5.8}$$

see Heckman and Vytlacil (2007a). According to Heckman (1990), one has

$$-\mathbb{E}[\varepsilon_0|D = 1, X = x, W = w] = \frac{\Pr\{D = 0|W = w\}}{\Pr\{D = 1|W = w\}}\mathbb{E}[\varepsilon_0|D = 0, X = x, W = w]. \tag{5.9}$$

Therefore, the treatment effect on the treated is also identifiable as

$$\begin{aligned} TTE(x,w) =& x'(\beta_{0,1} - \beta_{0,0}) + \lambda_{0,1}(w'\gamma_0) + \frac{\Pr\{D = 0|W = w\}}{\Pr\{D = 1|W = w\}}\lambda_{0,0}(w'\gamma_0) \\ =& x'(\beta_{0,1} - \beta_{0,0}) + \lambda_{0,1}(w'\gamma_0) + \frac{F_{\nu 0}(-w'\gamma_0)}{\bar{F}_{\nu 0}(-w'\gamma_0)}\lambda_{0,0}(w'\gamma_0), \end{aligned} \tag{5.10}$$

where $\bar{F}_{\nu 0}(\cdot) \equiv 1 - F_{\nu 0}(\cdot)$ denotes the marginal survival function of ν. Assuming monotone control functions $\lambda_{0,1}$ and $\lambda_{0,0}$, one can apply our semiparametric method to two sets of sample selection data, $(Y_i(1), D_i, X_i, W_i)$ and $(Y_i(0), D_i, X_i, W_i)$, separately to obtain $(\hat{\beta}_{n,1}, \hat{\lambda}_{n,1})$, $(\hat{\beta}_{n,0}, \hat{\lambda}_{n,0})$, and $(\hat{\gamma}_n, \hat{F}_{n\nu})$, which deliver consistent estimates for $ATE(x)$ and $TT(x,w)$ without any tuning parameter.

Considering the equivalence result in Vytlacil (2002), one can also make use of the generalized Roy model to uncover the Local Average Treatment Effect (LATE), which is the average effect for a subpopulation of compliers compelled to alter their treatment status by an external instrument (Imbens and Angrist, 1994). In particular, one has

$$LATE(X = x, D(w) = 1, D(\bar{w}) = 0) = x'(\beta_{0,1} - \beta_{0,0}) + \mathbb{E}[\varepsilon_1 - \varepsilon_0| - w'\gamma_0 \leq \nu \leq -\bar{w}'\gamma_0].$$

Under the marginal mean restriction that $\mathbb{E}[\varepsilon_1] = \mathbb{E}[\varepsilon_0] = 0$, one can derive the selection correction term $\mathbb{E}[\varepsilon_1 - \varepsilon_0| - w'\gamma_0 \leq \nu \leq -\bar{w}'\gamma_0]$ explicitly, which leads to

$$LATE(X = x, D(w) = 1, D(\bar{w}) = 0) = x'(\beta_{0,1} - \beta_{0,0}) + \Gamma_1(w,\bar{w}) - \Gamma_0(w,\bar{w}), \tag{5.11}$$

where

$$\Gamma_1(w,\bar{w}) = -\frac{\bar{F}_{\nu 0}(-\bar{w}'\gamma_0)\lambda_{0,1}(\bar{w}) + F_{\nu 0}(-w'\gamma_0)\bar{\lambda}_{0,1}(w)}{F_{\nu 0}(-\bar{w}'\gamma_0) - F_{\nu 0}(-w'\gamma_0)}, \tag{5.12}$$

$$\Gamma_0(w,\bar{w}) = -\frac{\bar{F}_{\nu 0}(-\bar{w}'\gamma_0)\bar{\lambda}_{0,0}(\bar{w}) + F_{\nu 0}(-w'\gamma_0)\lambda_{0,0}(w)}{F_{\nu 0}(-\bar{w}'\gamma_0) - F_{\nu 0}(-w'\gamma_0)}, \tag{5.13}$$

with

$$\bar{\lambda}_{0,j}(w) \equiv -\frac{\bar{F}_{\nu 0}(-w)\lambda_{0,j}(w)}{F_{\nu 0}(-w)} \quad \text{for} \quad j = 0, 1. \tag{5.14}$$

Now it becomes clear that the semiparametric estimates produced by our procedure deliver a semiparametric estimation of LATE, which generalizes Heckman, Tobias, and Vytlacil (2003) where these structural control function estimators are based on multivariate normal or t distributions of the error terms.

5.3. A Panel Selection Model

We consider a two-period panel data model in Kyriazidou (1997):

$$\begin{aligned} Y_{it}^* &= X_{it}'\beta_0 + \alpha_i + \varepsilon_{it}; \\ D_{it} &= \mathbb{I}\{W_{it}'\gamma_0 + \eta_i + \nu_{it} > 0\}. \end{aligned} \tag{5.15}$$

We only observe the dependent variable for the selected sample with $D_{it} = 1$,; i.e., $Y_{it} = Y_{it}^* D_{it}$ for $i = 1, \cdots, n$ and $t = 1, 2$. In order to utilize the control function approach, we make the following assumptions regarding the latent errors $(\varepsilon_{it}, \nu_{it})$ and unobserved heterogeneity terms (α_i, η_i).

Condition 10. The heterogeneity η_i in the selection equation is independent of W_i and ν_{it}. ε_{it} is independent of $\nu_{it'}$ given ν_{it} for $t \neq t'$.

Thereafter, we have the following identity:

$$\mathbb{E}[Y_{i1} - Y_{i2}|D_{i1} = 1, D_{i2} = 1, W_i] = (X_{i1} - X_{i2})'\beta_0 + \lambda_{01}(W_{i1}'\gamma_0) - \lambda_{02}(W_{i2}'\gamma_0), \tag{5.16}$$

where

$$\lambda_{0t}(W_{it}'\gamma_0) = \int \mathbb{E}[\varepsilon_{it}|\nu_{it} > -W_{it}'\gamma_0 - \eta_i] dF_\eta(\eta_i) \quad \text{for} \quad t = 1, 2, \tag{5.17}$$

where $F_\eta(\cdot)$ stands for the distribution of η_i. Apparently, the condition we present in Section 2 leads to the monotonicity of $\mathbb{E}[\varepsilon_{it}|\nu_{it} > -W_{it}'\gamma_0 - \eta_i]$ with respect to $W_{it}'\gamma_0$ for any η_i. Integrating out η_i does not alter the monotonicity, so the exact same monotone restriction is inherited by the control functions $\lambda_{0t}(W_{it}'\gamma_0)$ for $t = 1, 2$.

Our assumptions regarding the heterogeneity terms are stronger than Kyriazidou (1997), yet are weaker than Wooldridge (1995), in the sense that α_i in the outcome equation is a fixed effect that can depend on covariates and η_i in the selection equation is a random effect that is independent of covariates and other error terms. Nevertheless, there is no parametric assumption on any unobserved error term in our model. In comparison, the model considered by Kyriazidou (1997) imposes no restriction on the dependence structure

of latent error terms (nor on the relationship between error and covariates), whereas the heterogeneity η_i is excluded from the selection equation in the model of Wooldridge (1995).

We describe below the simple semiparametric estimation for the panel data setting.

Stage 1 (i) For any γ, we compute the NPMLE for $F(\cdot)$ for two selection equations in both time periods:

$$\hat{F}_{n\nu_t}(\cdot;\gamma) = \arg\max_{F} \sum_{i=1}^{n} \left[\bar{D}_{it} \log F(-W_{it}'\gamma) + (1-\bar{D}_{it}) \log(1 - F(-W_{it}'\gamma))\right], \tag{5.18}$$

where $\bar{D}_{it} = 1 - D_{it}$ for $i = 1, \cdots, n$ and $t = 1, 2$.

Stage 1 (ii) Given $\hat{F}_{n\nu_t}(\cdot;\gamma)$ at hand, our estimator $\hat{\gamma}_{nt}$ for the regression coefficient is the zero-crossing point of the estimation equation:

$$\frac{1}{n}\sum_{i=1}^{n} W_{it}\left[\bar{D}_{it} - \hat{F}_{n\nu_t}(-W_{it}'\hat{\gamma}_{nt};\hat{\gamma}_{nt})\right] = 0. \tag{5.19}$$

Stage 2 Given $\hat{\gamma}_{nt}$, we estimate β_0 and $\lambda_{0t}(\cdot)$ by the least squares estimator under the shape restriction for λ_t:

$$(\hat{\beta}_n, \hat{\lambda}_{n1}, \hat{\lambda}_{n2}) = \arg \min_{\beta\in\mathbf{B},\lambda_1,\lambda_2\in\mathcal{D}} \sum_{D_{i1}=D_{i2}=1} \left[\Delta Y_i - \Delta X_i'\beta - \lambda_2(W_{i2}'\hat{\gamma}_{n2}) + \lambda_1(W_{i1}'\hat{\gamma}_{n1})\right]^2, \tag{5.20}$$

where $\Delta Y_i \equiv (Y_{i2} - Y_{i1})$ and $\Delta X_i \equiv (X_{i2} - X_{i1})$. This optimization problem boils down to minimizes a convex function over a convex set; therefore, the estimator $(\hat{\beta}_n, \hat{\lambda}_{n1}, \hat{\lambda}_{n2})$ exists and is well-defined (Meyer, 2013).

One can adapt our theorems and the proofs in Mammen and Yu (2007) and Cheng (2009) to show the root-n consistency and asymptotic normality for $\hat{\beta}_n$. In contrast, the estimator for β_0 proposed by Kyriazidou (1997) relies on kernel smoothing to control for the endogeneity associated with η_i for the more general model and the convergence rate is slower than $\sqrt{n}$.

6. Monte Carlo Simulations

This section conducts Monte Carlo simulations to evaluate the finite sample performance of our estimator based on a monotone control function. In the following, we refer to it as "monotone CF" estimator. Two alternative procedures are considered for comparison. The first one is Heckman's two-step estimator requiring joint normality on the latent errors. The second one is a kernel-based estimator, which treats the control function $\lambda(\cdot)$ as completely unknown and does not impose any monotonicity restriction. Referring to the latter one, empiircal researchers often combine Ichimura (1993) or Klein and Spady (1993)'s estimator

for the single index model and Robinson (1988) for the partial linear model (Schafgans, 1998, 2000; Martins, 2001).

We consider a simulation design with the following outcome and selection equations:

$$\begin{aligned} Y_i^* &= \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \\ D_i &= \mathbb{I}\{-1 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 \tilde{W}_i + \nu_i > 0\}, \end{aligned} \tag{6.1}$$

with $\beta_1 = -1$, $\beta_2 = 1$, $\gamma_1 = 1$,$\gamma_2 = 0.5$, and $\gamma_3 = -2$. The observed random vectors consist of $\{(Y_i, D_i, X_{1i}, X_{2i}, \tilde{W}_i)\}_{i=1}^n$, where $Y_i = Y_i^* D_i$. Let X_{1i} follow the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$, X_{2i} follow the standard normal distribution, $\tilde{W}_i$ follow the exponential distribution with a unit variance.[12] The joint distribution of (ε, ν) is a bivariate normal mixture as follows.[13]

$$\begin{bmatrix} \varepsilon \\ \nu \end{bmatrix} \sim \pi N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right) + (1-\pi) N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \times 15^2 \right),$$

with $\pi = 0.9$, $\sigma = 0.25$, and $\rho = 0.9$. The error terms ε and ν have standard deviations around 1.21. The proportion of $D_i = 1$ is about 0.37. Simulation results are based on $1,000$ replications.

Table 1 presents the median bias and the mean absolute value (MAE) of our shape restricted estimator (shape rest.), Heckman's two-step estimator (Heckit), and a kernel-based semiparametric estimator (Klein-Spady-Robinson). The kernel-based estimator uses the Klein-Spady estimator for γ_2 and γ_3 in the first stage and the Robinson estimator (for the partial linear model) as the second stage. Note that two bandwidths are required in this kernel-based estimator. Our simulations set bandwidth choices $c_1 \times h_{cv,1}$ for the Klein-Spady estimator and $c_2 \times h_{cv,1}$ for the Robinson estimator, where both bandwidths $(h_{cv,1}, h_{cv,2})$ are selected by cross-validation. We further vary the constant terms $c_1 = \{1/2, 1, 2\}$ and $c_2 = \{1, 2, 3, 4\}$ to investigate the sensitivity of kernel-based estimators related to the bandwidth choice. Table 2 presents the median bias and the MAE for the first stage estimation (binary choice model) regarding parameters γ_2 and γ_3, with γ_1 normalized to 1.

Table 1 shows that Heckman's two-step estimator is not consistent for the normal mixture error terms as the median bias and the MAE are not only large in magnitude and but also do not diminish when the sample size increases. This is not surprising, since the Heckman's two-step method does not account for any deviation from the joint normality assumption. The monotone CF estimator yields a much smaller median bias and MAE for β_1 and β_2 than Heckman's two-step approach. Moreover, both the median bias and MAE of the monotone CF estimator decrease substantially when the sample size increases. When it

[12] The density function of $\tilde{W}_i$ is $g(w) = \mathbb{I}\{w > -1\} \exp(-w - 1)$.

[13] It corresponds to $\sigma_1 = \sigma_2 = \sigma$ and $k = 15$ in Example 2.4.

comes to the kernel-based estimator, its performance critically depends on the choices of bandwidths. Using the cross-validated bandwidth in both stages ($c_1 = c_2 = 1$), the kernel-based estimator performs slightly better than the monotone CF estimator in terms of MAE and yields a notably smaller bias. When the bandwidth coefficients (c_1, c_2) become $(1, 2)$ or $(2, 2)$, the MAE of the kernel-based estimator is similar to that of the monotone CF estimator. Moreover, the monotone CF estimator has a smaller MAE than the kernel-based estimator if the latter uses the bandwidth of $h_{cv,1}/2$ in the first stage. Similar patterns can be found in Table 2. When considering the first stage estimation of the selection equation, the Heckman's two-step approach is again not consistent. Comparing two semi-parametric estimators, one can see that the MAE of the kernel-based estimator is smaller when the "good" bandwidth is used ($c_1 = 1$) but it can be larger than the monotone CF estimator when other bandwidths are used, say $c_1 = 1/2$ and 2.

In sum, our simulation experiments show that the monotone CF estimator have a robust finite sample performance when the error terms are non-normal. Free from any user-specified tuning parameter, our estimator does not suffer from sensitivity with respect to bandwidths in the kernel based estimation. In terms of MAE, our estimator is comparable to the kernel-based estimator using "good" bandwidths and outperforms the latter when "bad" bandwidths are chosen.

TABLE 1. Finite sample performances of estimators for the outcome equation. The bandwidth for the Klein-Spady estimation is $c_1 \times h_{cv,1}$ and the bandwidth for the Robinson estimation is $c_2 \times h_{cv,2}$, where $h_{cv,1}$ and $h_{cv,2}$ are the bandwidth from cross-validation, respectively. A Gaussian kernel is used.

n	Method	Bandwidths	β_1		β_2	
		(c_1, c_2)	Med.bias	MAE	Med.bias	MAE
1,000	Monotone CF		.0952	.1172	.0421	.0668
	Heckit		.1767	.1881	.0681	.0856
	Klein-Spady-Robinson	$(1,1)$	-.0224	.0945	-.0167	.0628
		$(1,2)$	-.0693	.1156	-.0331	.0714
		$(1,3)$	-.1208	.1506	-.0569	.0837
		$(1,4)$	-.1720	.1881	-.0783	.0956
		$(1/2,1)$	.0393	.1635	.0002	.0897
		$(1/2,2)$	-.0074	.1607	.0154	.0886
		$(1/2,3)$	-.0685	.1666	-.0431	.0925
		$(1/2,4)$	-.1264	.1819	-.0678	.0992
		$(2,1)$	-.0442	.0985	-.0175	.0618
		$(2,2)$	-.0919	.1247	-.0362	.0712
		$(2,3)$	-.1472	.1634	-.0597	.0843
		$(2,4)$	-.1906	.1997	-.0791	.0958
2,000	Monotone CF		.0690	.0842	.0329	.0493
	Heckit		.1862	.1870	.0747	.0796
	Klein-Spady-Robinson	$(1,1)$	-.0072	.0676	-.0043	.0431
		$(1,2)$	-.0483	.0817	-.0229	.0492
		$(1,3)$	-.0969	.1163	-.0422	.0608
		$(1,4)$	-.1465	.1552	-.0596	.0730
		$(1/2,1)$	.0418	.1180	.0117	.0628
		$(1/2,2)$	.0109	.1144	-.0021	.0623
		$(1/2,3)$	-.0413	.1216	-.0233	.0655
		$(1/2,4)$	-.0982	.1411	-.0457	.0719
		$(2,1)$	-.0224	.0705	-.0059	.0433
		$(2,2)$	-.0605	.0878	-.0221	.0495
		$(2,3)$	-.1073	.1237	-.0400	.0615
		$(2,4)$	-.1562	.1617	-.0596	.0741

TABLE 2. Finite sample performances of estimators for the selection equation. The bandwidth for the Klein-Spady estimation is $c_1 \times h_{cv,1}$, where $h_{cv,1}$ is the cross-validated bandwidth. A Gaussian kernel is used.

n	Method	Bandwidths	γ_2		γ_3	
		c_1	Med.bias	MAE	Med.bias	MAE
1,000	Monotone CF		-.0119	.0334	.0514	.0856
	Probit		-.0185	.0442	.2151	.2194
	Klein-Spady	1	-.0034	.0310	.0205	.0834
		1/2	-.0203	.0764	-.1796	.2844
		2	.0178	.0392	-.0692	.1203
2,000	Monotone CF		-.0063	-.0227	.0415	.0511
	Probit		-.0194	.0331	.2105	.2116
	Klein-Spady	1	.0007	.0218	.0079	.0579
		1/2	-.0154	.0543	.1390	.2245
		2	.0124	.0252	-.0380	.0785

7. An Empirical Application

In this section, we apply our estimation and testing methods to re-examine the wage equations of the Malaysian Chinese, using the monotone control function to correct for the sample selection bias. The data is drawn from the Second Malaysian Family Life Survey and is provided by Schafgans (1998). The choice of dependent and independent variables follows Schafgans (1998). The latent dependent variable Y_i^* represents the ith individual's latent hourly wage offer (in logarithms) and D_i is a dummy variable indicating whether this individual is a paid worker. One has $D_i = 1$ if the offered wage exceeds the reservation wage (Gronau, 1974; Heckman, 1974). Exogenous variables, W_i, entering the selection equation are: age, age squared (divided by 100), years of primary schooling, years of secondary schooling and above, dummy variable "Fail"(whether the individual failed the certificate at the education level he/she completed), dummy variable "Urban"(the location of the individual's residence), and non-employment variables including unearned income (average annual property income of the household), house ownership (a house ownership indicator times the cost of the housing), and land ownership (the amount of land-holding). We impose the standard exclusion restriction such that the non-employment variables does not appear in the wage offer equation; that is, they alter the reservation wage without affecting the offered wage. Exogenous variables X_i entering the wage offer equation consist of potential experience, potential experience squared (divided by 100), years of primary

schooling, years of secondary schooling and above, and two dummy variables "Fail" and "Urban".

Tables 3 and 4 present the estimates of the coefficients in the wage equation using three approaches: Heckman's two-step estimator, our monotone CF semiparametric estimator, and Schafgans (1998)'s kernel-based semi-parametric approach.[14] When implementing the monotone CF estimator, the selection correction function (control function) λ is assumed to be decreasing for working men (Table 3) and increasing for working women (Table 4). This choice is made by combining several pieces of evidence together. First of all, Heckman's two-step estimates of the coefficient attached to the inverse Mill's ratio are 0.3891 for men and -0.2787 for women. Second, the monotonicity assumption of the control function $\lambda(\cdot)$ is also supported by Figure 2, which compares the plots of monotone CF estimate of λ (solid line) versus the unrestricted kernel estimate (dash line).[15] Both estimates are decreasing for male workers and both show an increasing trend for female workers, despite some small fluctuations in the kernel estimate. Last but not least, the choice is also consistent with the reported p-values in the selectivity tests. One plausible explanation for the increasing control function for Chinese women may be due to an assortative matching in marriage, so a married women with higher productivity may have less incentive to work.

Tables 3 and 4 show that for most slope parameters, the monotone CF estimates are comparable to the other two estimates. For parameters where the Heckman's two-step estimate and Schafgans (1998)'s kernel estimate noticeably differ, such as with the coefficients on the variables "Secondary schooling" and "Fail" in Table 4, our estimates are closer to the ones from the kernel-based approach. We further present the Oaxaca (1973) decomposition using different estimates. The actual difference in the means of the log-wages for men and women workers is 0.3662. In the OLS case where no selection correction is made, 17.09% of this gender differential is explained by the term $(\bar{X}_m - \bar{X}_f)'\beta_f$, which describes the difference in wage-related characteristics.[16] Heckmen's two-step approach attributes 21.16% of the wage differential to the difference in characteristics and this percentage is 25.12% in Schafgans (1998)'s kernel-based approach. The monotone CF test suggests that a percentage as large as 28.78% owes to the difference in wage-related characteristics.

We also conduct a formal test for the presence of labor market selection. Table 5 reports the p-values of the t-test based on Heckman's selection model, our selectivity test

[14]Schafgans (1998)'s semi-parametric approach estimates the selection equation using Ichimura (1993)'s technique and then estimates the outcome equation, which is a partial linear model using Robinson (1988). The numbers in the last column of Tables 3 and 4 are drawn from Table III of Schafgans (1998).

[15]The kernel estimate uses the slope estimates of Schafgans (1998) in the wage offer equation and bandwidths are chosen by cross-validation.

[16]Here $\bar{X}_m$ and $\bar{X}_f$ denote the mean of X for men and women, respectively, and β_f denotes the coefficients on X for women.

based on a monotone control function under the general alternative in Section 3.2,[17] and a kernel-based test in the spirit of Christofides, Li, Liu, and Min (2003).[18] For our tests, both increasing and decreasing cases are considered. The cases consistent with the control functions depicted in Figure 2 (so that the control function is decreasing for men and increasing for women) are in bold font. For female workers, the test assuming an increasing control function detects stronger evidence against the no selection null hypothesis (p-value is .134) than the one with a decreasing control function (p-value .592). This is also consistent with the kernel-based estimate of the control function plotted in Figure 2 (the right panel). Compared with the t-test based on Heckman's selection model, our test based on an increasing control function produces a p-value (.134) closer to the kernel-based test (p-value is .060). For male workers, the test based on a decreasing control function reveals stronger evidence against the null hypothesis. Once again, this finding is in line with the kernel-based estimate in Figure 2 (the left panel). However, in the left panel, the piece-wise constant $\hat{\lambda}$ (monotone CF estimate) is much steeper than the smoothed $\hat{\lambda}$ (kernel estimate), which is also reflected in the p-values: the p-value is .080 for the test based on a decreasing control function, while the value is .866 for the kernel-based test.

[17]The critical values are calculated from 500 bootstrap samples.

[18]The kernel-based test rejects the null hypothesis of no sample selection if $n_1 h^{1/2} I_n/\hat{\sigma} > z_{1-\alpha}$ where $I_n = 1/(n_1^2 h)\sum_{i=1}^{n_1}\sum_{j=1,j\neq i}^{n_1}\hat{\varepsilon}_i\hat{\varepsilon}_j K((W_i'\hat{\gamma} - W_j'\hat{\gamma})/h)$, $\hat{\sigma}^2 = 2/(n_1^2 h)\sum_{i=1}^{n_1}\sum_{j=1,j\neq i}^{n_1}\hat{\varepsilon}_i^2\hat{\varepsilon}_j^2 K^2((W_i'\hat{\gamma} - W_j'\hat{\gamma})/h)$, $n_1 = \sum_{i=1}^{n} D_i$, $\hat{\varepsilon}_i$ is the OLS residual $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}_{ols}$, $\hat{\gamma}$ is the semiparameteric estimates of the selection equation in Schafgans (1998), $K(\cdot)$ is the Gaussian kernel function, and the bandwidth h is computed by cross-validation for estimating $\mathbb{E}(\varepsilon_i|W_i'\hat{\gamma})$.

TABLE 3. Wage equation for Chinese males using different corrections for sample selection. Number of total obs =1,190; number of working obs =559.

	Heckit	Monotone CF	Schafgans (1998)
Experience	.1109	.1237	.1051
	[.0887, .1331]	[-.0937, .1396]	[.0837, .1265]
Experience squared	-.1840	-.2130	-.1750
	[-.2316, -.1364]	[-.2452, -.1411]	[-.2213, -.1287]
Primary schooling	.0235	.0260	.0232
	[-.0205, .0674]	[-.0345, .0760]	[-.0184 .0648]
Secondary schooling	.1638	.1693	.1565
	[.1404, .1872]	[.1381, .1924]	[.1341, .1789]
Fail	-.1142	-.1148	-.1298
	[-.2416, .0132]	[-.2496, .0095]	[-.2455 -.0141]
Urban	.0751	.0543	.1047
	[-.0376, .1878]	[-.0311, .1837]	[-.0025, .2119]

Note: The confidence interval for the monotone CF estimate is calculated from 500 bootstrap samples.

TABLE 4. Wage equation for Chinese females using different corrections for sample selection. Number of total obs =1,298; number of working obs =371.

	Heckit	Monotone CF	Schafgans (1998)
Experience	.0551	.0394	.0564
	[.0298, .0804]	[.0171, .0870]	[.0295 .0833]
Experience squared	-.0511	-.0142	-.0635
	[-.1138, .0116]	[-.1274, .0411]	[-.1289, .0019]
Primary schooling	.1094	.1299	.0965
	[.0460, .1728]	[.0002, .1917]	[.0383, .1547]
Secondary schooling	.1451	.0859	.0821
	[.0892, .2010]	[.0426, .1885]	[-.0016, .1658]
Fail	-.2145	-.2999	.4214
	[-.3754, -.0536]	[-.4360, -.0914]	[-.6872, -.1556]
Urban	.0275	.0091	.0163
	[-.0974, .1520]	[-.1082, .1390]	[-.1038, .1364]

Note: The confidence interval for the Monotone CF estimate is calculated from 500 bootstrap samples.

FIGURE 2. The estimated control function $\hat{\lambda}(\hat{\gamma}'W)$

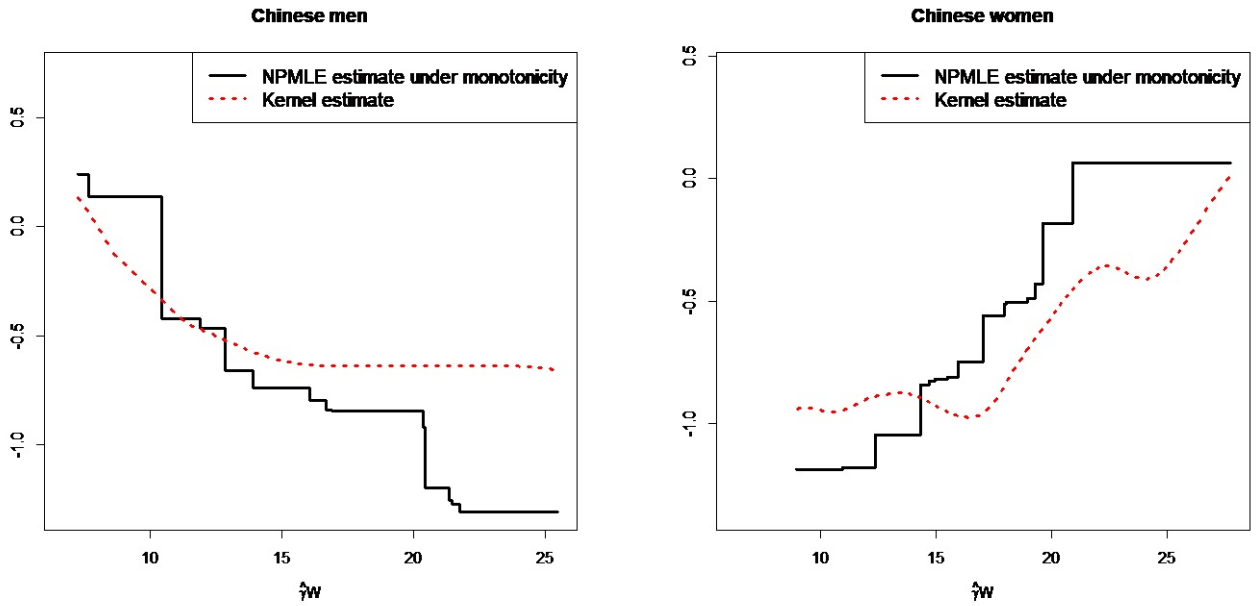


TABLE 5. The p-values for testing the presence of sample selection bias in the wage equation of Malaysian Chinese. H_0 : the control function λ is a constant.

	Heckit t-test	Test with monotone CF		Kernel-based test
		Decreasing	Increasing	
Men	.174	**.080**	.528	.866
Women	.357	.592	**.134**	.060

Note: For the selectivity test based on a monotone control function, both increasing and decreasing cases are considered. The cases consistent with control functions depicted in Figure 2 (so that the control function is decreasing for men and increasing for women) are in bold font.

8. Conclusion

This paper proposes a semiparametric sample selection model with a monotonicity constraint on the selection correction function. Nonrandom selection is both a source of bias in empirical research and a fundamental aspect of many social processes. The popularity of Heckman's two-step procedure to correct selectivity bias is witnessed by its profound impact on all of these fields; Heckman (1979) has received more than 28,000 Google Scholar citations. Lying between the original Heckman selection model and the semi-parametric selection model (Robinson, 1988; Newey, 2009; Das, Newey, and Vella, 2003; Ahn and Powell, 1993) where the control function is completely unknown, our new sample selection model imposes no parametric distributional assumptions and delivers automatic semiparametric estimation and testing. Therefore, the proposed method shares the generality of semiparametric approaches while keeps the main convenience of parametric approaches as its implementation is free from any tuning parameter.

This research will also add to the rich and evolving literature exploring various shape restrictions in estimation and inference. Shape restrictions, such as convexity, homogeneity, and monotonicity, frequently arise either as important assumptions or consequences of assumptions of economic models. Embedding shape restrictions into the estimation has a key advantage in that the underlying criterion function can be meaningfully maximized (or minimized) without additional penalization or smoothing. Meanwhile, the estimated component automatically satisfies the imposed shape restriction, which makes it more attractive to practitioners. We take the first step of introducing shape restrictions to the sample selection model by converting an intuitive concept regarding the dependence between latent errors into a precise condition known as RTI (RTD). The technical contributions we present

are also of independent interest for the general two-stage semiparametric estimation and testing involving shape-restricted components.

9. Appendix A: Proofs of Main Results

In the Appendix, we denote a large positive constant by M, whose value might change line by line. We introduce additional subscripts when there are multiple constant terms in the same display. For two sequences a_n, b_n, we write $a_n \lesssim b_n$, if $a_n \leq b_n$ for some large M independent of n.

Proof of Theorem 4.1. Given Groeneboom and Hendrickx (2018), we know that $\hat{\gamma}_n$ defined in the second step exists with probability tending to 1, $\hat{\gamma}_n \to_p \gamma_0$, and $\hat{F}_{n\nu}(u; \hat{\gamma}_n)$ converges uniformly to $F_{\nu 0}(u)$. For the (finite and infinite dimensional) parameters and estimators in the outcome equation, we use the short-hand notations $\theta_0 = (\beta_0, \lambda_0(\cdot))$ and $\hat{\theta}_n = (\hat{\beta}_n, \hat{\lambda}_n(\cdot))$. Moreover, we define the following metric:

$$d(\theta, \theta_0; \gamma) = |\hat{\beta}_n - \beta_0| + \parallel \hat{\lambda}_n(w'\gamma) - \lambda_0(w'\gamma_0) \parallel . \tag{9.1}$$

Because $\hat{\beta}_n$ and $\hat{\lambda}_n$ are solutions of the least squares problem, we get

$$\mathbb{P}_n[Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n)]^2 \leq \mathbb{P}_n[Y - X'\beta_0 - \lambda_0(W'\hat{\gamma}_n)]^2. \tag{9.2}$$

Hence, this leads to

$$\begin{aligned} &P\left[D_i\left\{(Y_i - X_i'\beta_0 - \lambda_0(W_i'\hat{\gamma}_n))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right\}\right] \\ \leq &(\mathbb{P}_n - P)\left[D_i\left\{(Y_i - X_i'\beta_0 - \lambda_0(W_i'\gamma_0))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right\}\right]. \end{aligned} \tag{9.3}$$

Thereafter, we prove in Lemma 10.9 that the left-hand side (l.h.s.) of inequality (9.3) can be bounded below by

$$\begin{aligned} &|\hat{\beta}_n - \beta_0|^2 + \parallel \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \parallel^2 - O_p(n^{-1/2}) \\ \lesssim &P\left[D_i\left\{(Y_i - X_i'\beta_0 - \lambda_0(W_i'\hat{\gamma}_n))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right\}\right]. \end{aligned} \tag{9.4}$$

The right-hand side (r.h.s.) of inequality (9.3) can be bounded up by

$$\begin{aligned} &(\mathbb{P}_n - P)\left[D_i\left\{(Y_i - X_i'\beta_0 - \lambda_0(W_i'\gamma_0))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right\}\right] \\ \leq &\sup_{\theta, |\gamma - \gamma_0| \leq M_1 n^{-1/2}, \sup|\lambda| \leq M_2 \log n} (\mathbb{P}_n - P) f^1_{\theta,\gamma} + o_p(1) \end{aligned}$$

for the function $f^1_{\theta,\gamma}$ defined in Lemma (10.6) because $|\hat{\gamma}_n - \gamma_0| = O_p(n^{-1/2})$ and $\sup_w |\hat{\lambda}_n(w'\hat{\gamma}_n)| = O_p(\log n)$. Therefore, by the Glivenko-Cantelli property of the functional class $f^1_{\theta,\gamma}$, the r.h.s. of inequality (9.3) is $o_p(1)$, which concludes the proof of consistency.

In order to obtain the rate of convergence, we get

$$\begin{aligned}
&\Pr\left\{d(\hat{\theta}_n, \theta_0; \hat{\gamma}_n) \geq \eta\right\} \\
\leq &\Pr\left\{\sup_{d(\theta,\theta_0;\gamma)\geq\eta, |\gamma-\gamma_0|\leq M_1 n^{-1/2}, \sup|\lambda|\leq M_2 \log n} (\mathbb{P}_n - P)[f^1_{\theta,\gamma}] - d^2(\theta, \theta_0; \gamma) \geq 0\right\} \\
+ &\Pr\{|\hat{\gamma}_n - \gamma_0| \geq M_1 n^{-1/2}\} + \Pr\left\{\sup_w |\hat{\lambda}_n(w'\hat{\gamma}_n)| \geq M_2 \log n\right\} \\
\equiv &P_{1n} + P_{2n} + P_{3n}
\end{aligned}$$

for any small positive η. It is clear that the last two terms converge to zero. Therefore, by the peeling argument and Theorem 3.4.2 in Van Der Vaart and Wellner (1996), we have

$$\begin{aligned}
P_{1n} &\leq \sum_{s=0}^{\infty} \mathbb{P}\left\{\sup_{2^s\eta\leq d(\theta,\theta_0;\gamma)\leq 2^{s+1}\eta, |\gamma-\gamma_0|\leq M_1 n^{-1/2}, \|\lambda\|\leq M_2 \log n} \mathbb{G}_n f^1_{\theta,\gamma} \geq n^{1/2} 2^{2s} \eta^2\right\} \\
&\leq \sum_{s=0}^{\infty} \mathbb{P}\left\{\sup_{f\in\mathcal{F}_{M2^{s+1}\eta}} \mathbb{G}_n f^1_{\theta,\gamma} \geq n^{1/2} 2^{2s} \eta^2\right\}.
\end{aligned}$$

Then, we apply the maximal inequality in equation (10.7) and the entropy bounds in equation (10.5) to get

$$\mathbb{E}\left\{\sup_{f\in\mathcal{F}_{M2^{s+1}\eta}} \mathbb{G}_n f^1_{\theta,\gamma}\right\} \lesssim M^{1/2} (\log n)^2 n^{-1/6} 2^{(s+1)/2},$$

where we take $\eta = M \log n \times n^{-1/3}$.

We can now bound P_{1n} by

$$\begin{aligned}
P_{1n} &\leq \sum_{s=0}^{\infty} \frac{M^{1/2} (\log n)^2 n^{-1/6} 2^{(s+1)/2}}{n^{1/2} 2^{2s} \eta^2} \\
&= \sum_{s=0}^{\infty} \frac{(\log n)^2 n^{-1/6} 2^{(s+1)/2}}{M^{3/2} n^{1/2} 2^{2s} (\log n)^2 n^{-2/3}} \\
&= M^{-3/2} \sum_{s=0}^{\infty} 2^{-3s/2},
\end{aligned}$$

which can be made arbitrarily small for a large enough M. Therefore, the stated convergence result holds. □

Proof of Theorem 4.2. The solution $(\hat{\beta}_n, \hat{\lambda}_n)$ of the shape-restricted optimization is characterized by a set of equality and inequality restrictions; see Robertson, Wright, and Dykstra (1988), Groeneboom and Wellner (1992), or Groeneboom and Jongbloed (2014). For our purpose, we only need the equality restriction expressed via the following score functions:

$$\mathbb{P}_n\left[D(Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n))X\right] = 0,$$
$$\mathbb{P}_n\left[D(Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n))g_n(W'\hat{\gamma}_n)\right] = 0,$$

where $g_n(\cdot)$ is any piece-wise constant function that has the same jump locations with $\hat{\lambda}_n(\cdot)$. Therefore, we start with the following characterization condition for our estimator $(\hat{\beta}_n, \hat{\lambda}_n)$:

$$\mathbb{P}_n\left[D(Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])\right] = 0, \tag{9.5}$$

as in Equations (3.3) and (3.4) of Huang (2002). Hence, one obtains

$$\begin{aligned}\sqrt{n}\mathbb{E}&\left[D(X'(\hat{\beta}_n - \beta_0) + \hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])\right] \\ &= \mathbb{G}_n\left[D(Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])\right],\end{aligned} \tag{9.6}$$

given the fact that $\mathbb{E}[\varepsilon|W, D=1] = 0$. Regarding the r.h.s. of Equation (9.6), we utilize the P-Donsker property in Lemma 10.6 to show that

$$\begin{aligned}&\mathbb{G}_n\left[D(Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])\right] \\ &\mathbb{G}_n\left[D(Y - X'\beta_0 - \lambda_0(W'\gamma_0))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \lambda_0(W'\gamma_0)])\right] + o_p(1) \\ &= \mathbb{G}_n\left[\epsilon(X - \mathbb{E}[X|D=1, W'\gamma_0])\right] + o_p(1).\end{aligned} \tag{9.7}$$

Furthermore, we decompose the l.h.s. of Equation (9.6) into two terms, J_{1n} and J_{2n}, defined as follows:

$$J_{1n} = \sqrt{n}\mathbb{E}\left[D(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])X'\right](\hat{\beta}_n - \beta_0), \tag{9.8}$$

$$J_{2n} = \sqrt{n}\left[D(\hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])\right]. \tag{9.9}$$

In our Lemmas 10.7 and 10.8, we prove that

$$J_{1n} = \mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])X'\right]\sqrt{n}(\hat{\beta}_n - \beta_0) + o_p(1 + \sqrt{n}|\hat{\beta}_n - \beta_0|), \tag{9.10}$$

and

$$J_{2n} = \mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])\dot{\lambda}_0(W'\gamma_0)W'\right]\sqrt{n}(\hat{\gamma}_n - \gamma_0) + o_p(1). \tag{9.11}$$

In sum, the desired linear representation for $\hat{\beta}_n$ follows after collecting the leading terms in J_{1n}, J_{2n}:

(9.12)

$$\begin{aligned}
&\mathbb{E}\left[D(X-\mathbb{E}[X|D=1,W'\gamma_0])X'\right]\sqrt{n}(\hat{\beta}_n-\beta_0)\\
=&\mathbb{G}_n\left[\epsilon(X-\mathbb{E}[X|D=1,W'\gamma_0])\right]-\mathbb{E}\left[D(X-\mathbb{E}[X|D=1,W'\gamma_0])\dot{\lambda}_0(W'\gamma_0)W'\right]\sqrt{n}(\hat{\gamma}_n-\gamma_0)\\
&+o_p(1+\sqrt{n}|\hat{\beta}_n-\beta_0|).
\end{aligned}$$

Finally, referring to the linear representation of $\hat{\gamma}_n$ and the fact that $\mathbb{E}[\epsilon|D=1,W]=0$, the two leading terms on the r.h.s. of Equation (9.12) are uncorrelated, which gives rise to the particular form of the asymptotic covariance matrix in Theorem 4.2. □

Before investigating the asymptotic properties of our test, we need to introduce additional definitions that characterize the weak convergence. These results are standard and we refer readers to Shorack (2000). For two distribution functions, F_1 and F_2, the Levy distance d_L is defined as

$$d_L(F_1,F_2)\equiv\inf\{\eta>0:F_1(x-\eta)-\eta\leq F_2(x)\leq F_1(x+\eta)+\eta,\quad\forall x\in\mathbb{R}\}. \tag{9.13}$$

The Levy distance metrizes weak convergence in the sense that $G_n\Rightarrow G$ if and only if $d_L(G_n,G)\to 0$ as $n\to\infty$. For two distribution functions, F_1 and F_2, the p-Wasserstein distance d_p is defined via

$$d_p(F_1,F_2)\equiv\inf_J\left\{[\mathbb{E}|S-T|^p]^{1/p}:S\sim F_1,T\sim F_2\right\}, \tag{9.14}$$

where the infimum is taken over all joint distributions J with two marginals equal to F_1 and F_2. In the sequel, we make use of the fact that $d_L(F_1,F_2)\leq\sqrt{d_1(F_1,F_2)}$.

Proof of Theorem 4.3. The proof essentially follows the route in Sen and Meyer (2017). First of all, let $\hat{G}_n$ be a sequence of random distribution functions of the bootstrap residuals ϵ^*. In the residual bootstrap, ϵ^* is obtained by re-sampling the centered residual $\hat{\epsilon}$. By Lemma 2.6 of Freedman (1981), $\hat{G}_n$ converges to G the distribution of ϵ, almost surely by the 2-Wasserstein distance; i.e.,

$$d_L(\hat{G}_n,G)\to 0\quad\text{and}\quad\int x^2d\hat{G}_n(x)\to\int x^2dG(x)\quad\text{a.s.} \tag{9.15}$$

By the projection nature of the operation, we have

$$\Pi(\mathbf{Y}|\mathcal{S}_0)-\Pi(\mathbf{Y}|\mathcal{S}_{1,\hat{\gamma}_n})=\Pi(\boldsymbol{\epsilon}|\mathcal{S}_0)-\Pi(\boldsymbol{\epsilon}|\mathcal{S}_{1,\hat{\gamma}_n}) \tag{9.16}$$

under the null hypothesis. Thereafter, one proceeds as

(9.17)

$$
\begin{aligned}
&\| \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} \\
\leq &\| \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) \|_{n,D} + \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} + \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} \\
&\leq 2 \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_{n,D} + \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D},
\end{aligned}
$$

which gives us

$$
\| \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} - \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} \leq 2 \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_{n,D} . \tag{9.18}
$$

To emphasize the dependence on the residual terms, we write our test statistics as $T_n(\epsilon)$ and $T_n(\epsilon^*)$. Thus, the following bound holds:

$$
\begin{aligned}
&|T_n^{1/2}(\epsilon) - T_n^{1/2}(\epsilon^*)| \\
&\leq \frac{\| \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} - \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D}}{\| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}} \\
&+ \frac{\| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D}}{\| \boldsymbol{\epsilon}^* - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) \|_{n,D}} \frac{\| \boldsymbol{\epsilon}^* - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \boldsymbol{\epsilon} + \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}}{\| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}} \\
&\leq \frac{2 \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_{n,D}}{\| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}} + \frac{\| \boldsymbol{\epsilon}^* - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \boldsymbol{\epsilon} + \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}}{\| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}} \\
&\leq 4 \frac{\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_{n,D}}{\| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}},
\end{aligned} \tag{9.19}
$$

where we have used Inequality (9.18) for the first term on the r.h.s. of the first inequality. For the second term,

$$
\frac{\| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D}}{\| \boldsymbol{\epsilon}^* - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) \|_{n,D}} \leq 1. \tag{9.20}
$$

Therefore, we have

(9.21)

$$
\begin{aligned}
d_1(H_n, H_n^*) \leq &\mathbb{E}|T_n(\epsilon) - T_n(\epsilon^*)| \leq 2\mathbb{E}|T_n^{1/2}(\epsilon) - T_n^{1/2}(\epsilon^*)| \\
\leq &8\mathbb{E}\left[\frac{\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_n}{\| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_{n,D}} \right] \leq 8\sqrt{\mathbb{E}[n_1 \| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|_n^{-2}] \mathbb{E}[n_1^{-1} \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon} \|_{n,D}^2]} \\
\to &0, \quad \text{a.s.},
\end{aligned} \tag{9.22}
$$

which leads to the conclusion that the bootstrap can approximate the null distribution of the test statistic. □

Proof of Theorem 4.4. The intuition behind the proof is as follows. Under the null where the control function is constant, the shape-restricted estimator is still consistent (with an

even faster rate), so that $T_n = o_p(1)$; whereas under the alternative specified in Theorem (4.4), the test statistic converges to a positive constant in probability.

Under the null hypothesis that the control function is a constant term, one can combine the proof of our Theorem (4.1) and Theorem 2.2 in Zhang (2002) to get $T_n = O_p(\log n/n)$, which leads to $c_{n,\alpha} = o(1)$. We then show that under the alternative hypothesis T_n converges to a positive constant in probability, giving the desired claim that $\mathbb{P}_{\lambda_{0,n}}\{T_n > c_{n,\alpha}\} \to 1$ for $\mathcal{H}_1 : \lambda_{0,n} \in \mathcal{D}$.

Regarding the power property, we show that

$$\frac{\| \boldsymbol{\xi}_{\mathcal{S}_1} - \boldsymbol{\xi}_{\mathcal{S}_{1,\hat{\gamma}_n}} \|_{n,D}^2}{n} \to_p 0,$$

by the Glivenko-Cantelli property of the corresponding functional, which leads to

$$T_n \to_p c/\sigma^2, \tag{9.23}$$

as $n \to +\infty$, combining with the two conditions stated in Theorem 4.4. □

10. Appendix B: Proofs of Technical Lemmas

First, we record Lemma 25.86 in Van Der Vaart (1998) here, which is needed in the proof of Theorem 4.1.

Lemma 10.1. *For any random variable Z, if $(\mathbb{E}[g_1(Z)g_2(Z)])^2 \le c\mathbb{E}[g_1^2(Z)]\mathbb{E}[g_2^2(Z)]$ for some $c \le 1$, then*

$$\mathbb{E}[(g_1(Z) + g_2(Z))^2] \ge (1 - \sqrt{c})(\mathbb{E}[g_1^2(Z)] + \mathbb{E}[g_2^2(Z)]). \tag{10.1}$$

The following Lemma adapts Lemma 10.1 in Baladbaoui, Groeneboom, and Hendrickx (2017), incorporating an estimated $\hat{\gamma}_n$ from the first stage estimation.

Lemma 10.2. *Under our Conditions, we have*

$$\sup_v |\hat{\lambda}_n(v)| = O_p(\log n). \tag{10.2}$$

Proof. Based on the max-min characterization of the additive isotonic regression (see equation 11 in Mammen and Yu (2007)), we have

$$\min_{1\le k\le n} \frac{\sum_{i=1}^k (Y_i - X_i'\hat{\beta}_n)}{k} \le \hat{\lambda}_n(W_i'\hat{\gamma}_n) \le \max_{1\le k\le n} \frac{\sum_{i=k}^n (Y_i - X_i'\hat{\beta}_n)}{n-k+1}, \tag{10.3}$$

which leads to

$$\min_i (Y_i - X_i'\hat{\beta}_n) \le \hat{\lambda}_n(W_i'\hat{\gamma}_n) \le \max_i (Y_i - X_i'\hat{\beta}_n). \tag{10.4}$$

Hence, one gets

$$\sup|\hat{\lambda}_n(v)| \leq \max_i |Y_i - X_i'\hat{\beta}_n| \lesssim \max_i |Y_i| + \max_i \parallel X_i \parallel . \tag{10.5}$$

Given the exponential tails of both Y and X, we obtain

$$\sup_v |\hat{\lambda}_n(v)| = O_p(\log n), \tag{10.6}$$

by Lemma 2.2.2 in Van Der Vaart and Wellner (1996). □

Here we restate some necessary definitions and Theorem 2.4.1 in Van Der Vaart and Wellner (1996) that will be used repeatedly in the sequel. $\|\cdot\|_\infty$ is the usual L_∞-norm for a function f with $\|f\|_\infty < \infty$ and $\|\cdot\|_2$ stands for the L_2-norm. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_2)$ for subclass $\mathcal{F}$ is defined to be the minimum of m such that $\exists$ $f_1^L, f_1^U, ..., f_m^L, f_m^U$ for $\forall f \in \mathcal{F}$, $f_j^L \leq f \leq f_j^U$ for some j, and $\left\|f_j^U - f_j^L\right\|_2 \leq \epsilon$. Denote $H_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_2) \equiv \log N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_2)$. Furthermore, the corresponding bracketing entropy integral is $\mathcal{J}_{[]}(\eta, \mathcal{F}, \|\cdot\|_2) = \int_0^\eta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_2)}d\epsilon$. The following lemma which is a restatement of Lemma 3.4.2 in van der Vaart and Wellner (1996) based on the L_2-norm is useful to bound the normalized empirical process $\mathbb{G}_n = \sqrt{n}(P_n - P)$ and $\|\mathbb{G}_n\|_\mathcal{F} = \sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|$.

Lemma 10.3. *Let $\mathcal{F}$ be a uniformly bounded class of measurable functions such that $\| f \|_2 \leq \delta$ and $\|f\|_\infty \leq M_0$, then*

$$\mathbb{E}\|\mathbb{G}_n\|_\mathcal{F} \lesssim \mathcal{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|_2)\left[1 + \frac{\mathcal{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|_2) M_0}{\delta^2\sqrt{n}}\right]. \tag{10.7}$$

Let $\mathcal{D}_M$ be the class of monotone decreasing functions with values in $[-M, M]$, then for all $\epsilon > 0$, one has

$$H_{[]}(\epsilon, \mathcal{D}_M, \|\cdot\|_2) \lesssim \frac{M}{\epsilon}; \tag{10.8}$$

see Van Der Vaart and Wellner (1996). We emphasize here only the range of the function is required to be bounded, not its domain.

The next lemma provides the entropy bound for an important functional class in our remaining proofs. Our proof makes use of a similar construction as in Lemma 4.9 of Baladbaoui, Durot, and Jankowski (2016).

Lemma 10.4. *Consider the following function class:*

$$\mathcal{F}^2_{K\delta,0} = \{\lambda(w'\gamma) : d(\theta, \theta_0) \leq \delta, \sup|\lambda| \leq K_1, |\gamma - \gamma_0| \leq K_2\}, \tag{10.9}$$

where the function $\lambda(\cdot)$ belongs to the class of monotone functions, then the following entropy bound holds:

$$H_{[]}\left(\epsilon, \mathcal{F}^2_{K\delta,0}, \|\cdot\|_2\right) \leq \frac{MK_1}{\epsilon}, \tag{10.10}$$

for some finite constant M.

Proof. For any small ϵ_γ, the compact neighborhood of γ_0 can be covered by N_γ neighborhoods with diameters no larger than ϵ_γ, where $N_\gamma \leq M\epsilon_\gamma^{-q}$. Thus, for any γ, we can find $i \in \{1, \cdots, N_\gamma\}$ such that $|\gamma - \gamma_i| \leq \epsilon$. For the monotone function λ, we can find brackets $[\lambda_j^L, \lambda_j^U]$ with size ϵ covering the class of monotone functions with range restricted to $[-K_1, K_1]$. Moreover, the number of brackets N_λ is bounded by $\exp(K_1\epsilon^{-1})$ up to some finite constant.

Consider any function $f(w)$ in $\mathcal{F}^2_{K\delta,0}$, one has

$$f(w) \equiv \lambda(w'\gamma) = \lambda(w'\gamma_i + w'(\gamma - \gamma_i)), \tag{10.11}$$

which leads to

$$\lambda(w'\gamma_i - M\epsilon_\gamma) \leq f(w) \leq \lambda(w'\gamma_i + M\epsilon_\gamma), \tag{10.12}$$

given that the covariates W have compact support. Therefore, we can cover the element in $\mathcal{F}^2_{K\delta,0}$ by

$$\lambda_j^L(w'\gamma_i - M\epsilon_\gamma) \leq f \leq \lambda_j^U(w'\gamma_i + M\epsilon_\gamma), \tag{10.13}$$

for a pair $[\lambda_j^L, \lambda_j^U]$ that covers λ.

Now we verify the size of new bracket $[\lambda_j^L(w'\gamma_i - M\epsilon_\gamma), \lambda_j^U(w'\gamma_i + M\epsilon_\gamma)]$ is less than ϵ up to some finite constant with a proper choice of ϵ_γ. We start with the following decomposition:

$$\begin{aligned}
&\| \lambda_j^U(w'\gamma_i + M\epsilon_\gamma) - \lambda_j^L(w'\gamma_i - M\epsilon_\gamma) \|_2 \\
\leq &\| \lambda_j^U(w'\gamma_i + M\epsilon_\gamma) - \lambda(w'\gamma_i + M\epsilon_\gamma) \|_2 \\
+ &\| \lambda(w'\gamma_i + M\epsilon_\gamma) - \lambda(w'\gamma_i - M\epsilon_\gamma) \|_2 \\
+ &\| \lambda(w'\gamma_i - M\epsilon_\gamma) - \lambda_j^L(w'\gamma_i - M\epsilon_\gamma) \|_2 .
\end{aligned}$$

Apparently, the first and third terms are bounded up by ϵ by the construction of $[\lambda_j^L, \lambda_j^U]$. Considering the second term, one get

$$\| \lambda(w'\gamma_i + M\epsilon_\gamma) - \lambda(w'\gamma_i - M\epsilon_\gamma) \|_2^2 \leq M \int_{-2M}^{2M} \left(\lambda(t) - \lambda(t - 2M\epsilon_\gamma)\right)^2 dt$$

by the change of variable. Now given the monotonicity of λ and the fact that it is bounded in absolute value by K, we have

$$\begin{aligned}\int_{-2M}^{2M}(\lambda(t)-\lambda(t-2M\epsilon_\gamma))^2\,dt \leq & M\int_{-2M}^{2M}(\lambda(t-2M\epsilon_\gamma)-\lambda(t))\,dt \\ = & M\left[\int_{-2M-2\epsilon_\gamma M}^{-2M}\lambda(t-2M\epsilon)dt-\int_{2M-2\epsilon_\gamma M}^{2M}\lambda(t)dt\right] \\ \lesssim & \epsilon_\gamma.\end{aligned}$$

Then we take $\epsilon_\gamma=\epsilon^2$, we get $\parallel \lambda(w'\gamma_i+M\epsilon_\gamma)-\lambda(w'\gamma_i-M\epsilon_\gamma)\parallel_2\lesssim \epsilon$. Thus, the overall bracketing entropy number is bounded by:

$$\begin{aligned}H_{[]}\left(\epsilon,\mathcal{F}^2_{K\delta,0},\|\cdot\|_2\right) \leq & \log N_\gamma+\log N_\lambda \\ \leq & 2q\log(\epsilon^{-1})+\frac{MK_1}{\epsilon}\lesssim\frac{MK_1}{\epsilon}.\end{aligned}$$

□

Now we obtain the entropy bounds for two key functional classes in the proofs of Theorem (4.1) and Theorem (4.2).

Lemma 10.5. *Consider the following functional classes for $j=1,2$*

$$\mathcal{F}^j_{K\delta}=\{f^j_{\theta,\gamma}(z):d(\theta,\theta_0)\leq\delta,\sup|\lambda|\leq K,|\gamma-\gamma_0|\leq Mn^{-1/2}\}, \tag{10.14}$$

where

$$f^1_{\theta,\gamma}(z)=d\left[(y-x'\beta_0-\lambda_0(w'\gamma_0))^2-(y-x'\beta-\lambda(w'\gamma))^2\right] \tag{10.15}$$

and

(10.16)

$$f^2_{\theta,\gamma}(z)=d\left[(y-x'\beta-\lambda(w'\gamma))(x-\chi\circ\lambda_0^{-1}(\lambda(w'\gamma)))-(y-x'\beta_0-\lambda_0(w'\gamma_0))(x-\chi(w'\gamma_0))\right].$$

Recall here $\chi(u)=\mathbb{E}[X|D=1,W'\gamma_0=u]$. For both clases, we have the following bounds hold

$$H_{[]}\left(\epsilon,\mathcal{F}^j_{K\delta},\|\cdot\|_2\right)\lesssim\frac{\delta}{\epsilon} \tag{10.17}$$

for $j=1$ and 2.

Proof. We only prove the results related to the functional class $\mathcal{F}^2_{K\delta}$ which is the more difficult one to handle. First of all, it is sufficient to bound the entropy number for the class consisting of the following functions:

$$f^2_{\theta,\gamma}(z)=d\left[(y-x'\beta-\lambda(w'\gamma))(x-\chi\circ\lambda_0^{-1}(\lambda(w'\gamma)))\right], \tag{10.18}$$

because the part after the minus sign in (10.16) does not involve any unknown parameter. We begin with the definitions of some subclasses:

$$\mathcal{F}^2_{K\delta,0} = \{\lambda(w'\gamma) : d(\theta,\theta_0) \leq \delta, \sup|\lambda| \leq K, |\gamma-\gamma_0| \leq Mn^{-1/2}\},$$
$$\mathcal{F}^2_{K\delta,1} = \{(x - \chi \circ \lambda_0^{-1}(\lambda(w'\gamma))) : d(\theta,\theta_0) \leq \delta, \sup|\lambda| \leq K, |\gamma-\gamma_0| \leq Mn^{-1/2}\},$$
$$\mathcal{F}^2_{K\delta,2} = \{d(y - x'\beta - \lambda(w'\gamma)) : d(\theta,\theta_0) \leq \delta, \sup|\lambda| \leq K, |\gamma-\gamma_0| \leq Mn^{-1/2}\}.$$

Hence, by Lemma 10.4, we get the following bound on the bracketing entropy

$$\log N_{[]}\left(\epsilon, \mathcal{F}^2_{K\delta,0}, \|\cdot\|_2\right) \lesssim \frac{\delta}{\epsilon}. \tag{10.19}$$

Essentially, the function in $\mathcal{F}^2_{K\delta,1}$ is a Lipschitz continuous transformation of the one in $\mathcal{F}^2_{K\delta,0}$ given our condition on $\chi \circ \lambda_0$. Thereafter, we resort to Theorem 2.7.11 in Van Der Vaart and Wellner (1996) which gives entropy bounds for classes of functions that are Lipschitz in the index parameter. The entropy is bounded by the one of the original index parameter space up to some constant. The same idea applies to the class $\mathcal{F}^2_{K\delta,2}$. Considering $\mathcal{F}^2_{K\delta}$ as the product of two subclasses $\mathcal{F}^2_{K\delta,1}$ and $\mathcal{F}^2_{K\delta,2}$, the desired conclusion follows from the result in Section 2.10.3 of Van Der Vaart and Wellner (1996). □

Lemma 10.6. *For the functional classes defined by (10.14) with $K_n = M_1 \log n$ and $\delta_n = M_2 \log n/n^{1/3}$ for some finite constant terms M_1 and M_2, we have the following stochastic equicontinuity results*

$$\| \mathbb{G}_n f^j_{\theta,\gamma} \|_{\mathcal{F}^j_{K_n\delta_n}} = o_p(1), \tag{10.20}$$

for $j = 1, 2$.

Proof. We only verify the statement for the class $\mathcal{F}^1_{K_n\delta_n}$ to avoid repetition. First of all, we define a rescaled functional class $\tilde{\mathcal{F}}^1_{K_n\delta_n} = K_n^{-1}\mathcal{F}^1_{K_n\delta_n}$ which consists of functions that are uniformly bounded. By the imposed Lipschitz continuity condition, for any function f within the class $\mathcal{F}^1_{K_n\delta_n}$, we have

$$\begin{aligned}
&P\left(d\left[(y - x'\beta_0 - \lambda_0(w'\gamma_0))^2 - (y - x'\beta - \lambda(w'\gamma))^2\right]^2\right) \\
\leq& 4P\left(d\left[\epsilon(x'(\beta-\beta_0) + \lambda(w'\gamma) - \lambda_0(w'\gamma_0))\right]\right)^2 \\
&+ 2P\left(d\left[(x'(\beta-\beta_0) + \lambda(w'\gamma) - \lambda_0(w'\gamma_0))\right]\right)^4 \\
\lesssim& |\hat{\beta}_n - \beta_0|^2 + \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \|_2^2
\end{aligned}$$

which leads to $Pf^2 \lesssim \delta_n^2/K_n^2$ for $f \in \tilde{\mathcal{F}}^1_{K_n\delta_n}$. One can also easily verify that $\| f \|_\infty \leq M$ for some finite constant M. Note that for any class $\mathcal{F}$, if the entropy integral is bounded by

$$\mathcal{J}_{[]}\left(\delta, \mathcal{F}, \|\cdot\|_2\right) \lesssim \int_0^\delta \sqrt{1 + M/\epsilon} d\epsilon, \tag{10.21}$$

then we have

$$\mathcal{J}_{[]}\left(\delta, \mathcal{F}, \|\cdot\|_2\right) \lesssim \delta + 2M^{1/2}\delta^{1/2}, \tag{10.22}$$

which follows from the elementary inequality that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$. Given the entropy bound in Lemma 10.5, we have

$$\mathcal{J}_{[]}\left(\delta_n/K_n, \tilde{\mathcal{F}}^j_{K_n\delta_n}, \|\cdot\|_2\right) \lesssim \sqrt{\delta_n}/\sqrt{K_n}. \tag{10.23}$$

By resorting to the maximal inequalities (10.7), we obtain the following bounds for both functional classes:

$$\begin{aligned}
&\mathbb{E}\left[\| \mathbb{G}_n f^j_{\theta,\gamma} \|_{\tilde{\mathcal{F}}^j_{K_n\delta_n}}\right] \\
&\lesssim \mathcal{J}_{[]}\left(\delta_n/K_n, \tilde{\mathcal{F}}^j_{K_n\delta_n}, \|\cdot\|_2\right)\left[1 + \frac{\mathcal{J}_{[]}\left(\delta_n/K_n, \tilde{\mathcal{F}}^j_{K_n\delta_n}, \|\cdot\|_2\right) K_0}{(\delta_n/K_n)^2\sqrt{n}}\right] \\
&\lesssim K_n^{1/2}\delta_n^{1/2},
\end{aligned}$$

for $j = 1, 2$. Hence, we get

$$\mathbb{E}\left[\| \mathbb{G}_n f^j_{\theta,\gamma} \|_{\mathcal{F}^j_{K_n\delta_n}}\right] \lesssim K_n^{3/2}\delta_n^{1/2}. \tag{10.24}$$

By taking $K_n = M_1 \log n$ and $\delta_n = M_2 \log n/n^{1/3}$ for some finite constant terms M_1 and M_2, we have

$$\mathbb{E}\left[\| \mathbb{G}_n f^j_{\theta,\gamma} \|_{\mathcal{F}^j_{K\delta}}\right] \lesssim \frac{(\log n)^2}{n^{1/6}}. \tag{10.25}$$

□

Now we are ready to verify the asymptotic negligibility of several terms in the proofs of our Theorem 4.1 and Theorem 4.2.

Lemma 10.7. *Under our conditions, we have*

$$J_{1n} = \mathbb{E}\left[D(X - \mathbb{E}[X|D = 1, W'\gamma_0])X'\right]\sqrt{n}(\hat{\beta}_n - \beta_0) + o_p(1 + \sqrt{n}|\hat{\beta}_n - \beta_0|). \tag{10.26}$$

Proof. We start with

$$\begin{aligned}
&J_{1n} - \mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])X'\right]\sqrt{n}(\hat{\beta}_n - \beta_0) \\
&= \mathbb{E}\left[D(\mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)] - \mathbb{E}[X|D=1, W'\gamma_0])X'\right]\sqrt{n}(\hat{\beta}_n - \beta_0) \\
&\lesssim \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| \sqrt{n}(\hat{\beta}_n - \beta_0),
\end{aligned}$$

where we have applied the Lipschitz continuity property of $\xi(\cdot)$. The desired result follows from the consistency result in Theorem 4.1. □

Lemma 10.8. *Under our conditions, we have*

$$J_{2n} = \mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])\dot{\lambda}_0(W'\gamma_0)W'\right]\sqrt{n}(\hat{\gamma}_n - \gamma_0) + o_p(1). \tag{10.27}$$

Proof. First of all, we have $\| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\hat{\gamma}_n) \| = O_p(n^{-1/3}\log n)$ by the root-n consistency $\hat{\gamma}_n$ and results in Theorem 4.1. Recall the definition of J_{2n}

$$\begin{aligned}
J_{2n} &= \sqrt{n}\left[D(\hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])\right] \\
&= \sqrt{n}\left[D(\hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))(X - \mathbb{E}[X|D=1, W'\hat{\gamma}_n])\right] \\
&+ \sqrt{n}\left[D(\hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))(\mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)] - \mathbb{E}[X|D=1, W'\hat{\gamma}_n])\right].
\end{aligned}$$

The second term on the r.h.s. of the equality can be bounded by the Cauchy-Schwarz inequality as

$$\begin{aligned}
&\left[D(\hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))(\mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)] - \mathbb{E}[X|D=1, W'\hat{\gamma}_n])\right] \\
&\lesssim \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| \times \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\hat{\gamma}_n) \| \\
&= O_p(n^{-2/3}\log^2 n) = o_p(n^{-1/2}).
\end{aligned}$$

Also, note that the following identity holds

$$\mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\hat{\gamma}_n])(\hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\hat{\gamma}_n))\right] = 0$$

by conditioning on $(D=1, W'\hat{\gamma}_n)$ and applying the law of iterated expectation. Thus, we have

$$J_{2n} = \sqrt{n}\mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\hat{\gamma}_n])(\lambda_0(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))\right] + o_p(1).$$

Moreover, it is straightforward to obtain

$$\sqrt{n}\mathbb{E}\left[D(\mathbb{E}[X|D=1, W'\gamma_0] - \mathbb{E}[X|D=1, W'\hat{\gamma}_n])(\lambda_0(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))\right] = o_p(1),$$

based on the Lipschitz continuity of $\chi(\cdot)$, the differentiability of $\lambda_0(\cdot)$, and the root-n consistency of $\hat{\gamma}_n$.

In sum, one arrives at

$$J_{2n} = \sqrt{n}\mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])(\lambda_0(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))\right] + o_p(1).$$

Finally, the claimed result follows from a standard Taylor expansion of $\lambda_0(\cdot)$ and the root-n consistency of $\hat{\gamma}_n$. □

Lemma 10.9. *Suppose our Conditions hold, then we have*

$$\begin{aligned}(10.28)\qquad & |\hat{\beta}_n - \beta_0|^2 + \parallel \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \parallel^2 - O_p(n^{-1/2}) \\ \lesssim & P\left[D_i\left\{(Y_i - X_i'\beta_0 - \lambda_0(W_i'\hat{\gamma}_n))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right\}\right].\end{aligned}$$

Proof. The following decomposition is straightforward.

(10.29)

$$\begin{aligned}& P\left[D_i\left\{(Y_i - X_i'\beta_0 - \lambda_0(W_i'\hat{\gamma}_n))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right\}\right] \\ &= P\left[D_i(X_i'(\hat{\beta}_n - \beta_0) + \hat{\lambda}_n(W_i'\hat{\gamma}_n) - \lambda_0(W_i'\gamma_0))^2\right] - P\left[D_i(\lambda_0(W_i'\hat{\gamma}_n) - \lambda_0(W_i'\gamma_0))^2\right] \\ &= P\left[D_i(X_i'(\hat{\beta}_n - \beta_0) + \hat{\lambda}_n(W_i'\hat{\gamma}_n) - \lambda_0(W_i'\gamma_0))^2\right] - O_p(n^{-1/2}),\end{aligned}$$

where in the last equality we have made use of the differentiability of $\lambda_0(\cdot)$ and the root-n consistency of $\hat{\gamma}_n$.

Now we apply Lemma (10.1) to get separated convergence for both $\hat{\beta}_n$ and $\hat{\lambda}_n$ as follows. We take $g_1 = X'(\hat{\beta}_n - \beta_0)$ and $g_2 = \hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0)$. We then have:

$$(\mathbb{E}[g_1 g_2])^2 = (\mathbb{E}[g_1\mathbb{E}[g_2|X]])^2 \leq \mathbb{E}[g_1^2]\mathbb{E}[(\mathbb{E}[g_2|X])^2],$$

by the law of iterated expectation and the Cauchy-Schwarz inequality. Also, given the non-degeneracy of g_2 conditional on X, we get

$$\mathbb{E}[(\mathbb{E}[g_2|X])^2] < \mathbb{E}[(\mathbb{E}[g_2|X])^2] + \mathbb{E}[(g_2 - \mathbb{E}[g_2|X])^2] = \mathbb{E}[g_2^2].$$

Thus, there exists a constant $c \leq 1$ such that

$$(\mathbb{E}[g_1 g_2])^2 \leq c\mathbb{E}[g_1^2]\mathbb{E}[g_2^2].$$

Applying Lemma (10.1) gives us

$$\begin{aligned}& (1 - \sqrt{c})\left(P[D_i(X_i'(\hat{\beta}_n - \beta_0))^2] + P[D_i(\hat{\lambda}_n(W_i'\hat{\gamma}_n) - \lambda_0(W_i'\gamma_0))^2]\right) \\ \leq & P\left[D_i(X_i'(\hat{\beta}_n - \beta_0) + \hat{\lambda}_n(W_i'\hat{\gamma}_n) - \lambda_0(W_i'\gamma_0))^2\right].\end{aligned}$$

Now given the full rank condition of $\mathbb{E}[XX'|D=1]$ and the fact that $\Pr\{D=1\}$ is bounded away from zero, we have

$$|\hat{\beta}_n - \beta_0|^2 + \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \|^2 \lesssim P\left[D_i(X_i'(\hat{\beta}_n - \beta_0) + \hat{\lambda}_n(W_i'\hat{\gamma}_n) - \lambda_0(W_i'\gamma_0))^2\right],$$

which leads to the desired conclusion given (10.29). □

Finally, we show the idea of Corollary 5.3 in Baladbaoui, Durot, and Jankowski (2016) applied to our context delivers the uniform convergence (within any compact set in the interior of the support) for the estimated control function.

Proof of Lemma (4.2). We denote the support of W by $\mathcal{W}$. By the normalization condition, both $\gamma_{0,1}$ and $\hat{\gamma}_{n,1}$ are equal to 1. Then, with $\underline{v}$ from Lemma (4.2) and upon change of variables, we get

$$\begin{aligned}\int_{\mathcal{W}} (\hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\hat{\gamma}_n))^2 f_{W|D=1}(w)dw &\geq \underline{v}\int_{\mathcal{W}} (\hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\hat{\gamma}_n))^2 dw \\ &= \underline{v}\int_{\mathcal{C}_{\hat{\gamma}_n}\times\mathcal{W}_{q-1}} (\hat{\lambda}_n(t_1) - \lambda_0(t_1))^2 dt_1\cdots dt_q,\end{aligned}$$

where $\mathcal{W}_{q-1} = \{(w_2,\cdots,w_q) : w \in \mathcal{W}\}$ and $\mathcal{C}_{\hat{\gamma}_n} = \{w'\hat{\gamma}_n : w \in \mathcal{W}\}$. Because $\int_{\mathcal{W}_{q-1}} dt_2\cdots dt_q > 0$, there exists another positive constant M such that

$$\begin{aligned}\int_{\mathcal{W}} (\hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\hat{\gamma}_n))^2 f_{W|D=1}(w)dw &\geq M\int_{\mathcal{C}_{\hat{\gamma}_n}} (\hat{\lambda}_n(v) - \lambda_0(v))^2 dv \\ &\geq M\int_{\underline{v}+\omega_n}^{\bar{v}-\omega_n} (\hat{\lambda}_n(v) - \lambda_0(v))^2 dv,\end{aligned}$$

with probability tending to 1, using the definition of ω_n and $\hat{\gamma}_n - \gamma_0 = O_p(n^{-1/2})$. Hence, it is straightforward to obtain

$$\begin{aligned}\left(\int_{\underline{v}+\omega_n}^{\bar{v}-\omega_n} (\hat{\lambda}_n(v) - \lambda_0(v))^2 dv\right)^{1/2} &\lesssim \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\hat{\gamma}_n) \| \\ &\leq \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| - \| \lambda_0(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| \\ &= O_p(n^{-1/3}\log n) - O_p(n^{-1/2}),\end{aligned}$$

which leads to the desired result. □

References

Abbring, J., and J. J. Heckman (2007): "Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete

choice, and general equilibrium policy evaluation," *Handbook of econometrics*, 6, 5145–5303.

Ahn, H., and J. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29.

Amblard, C., and S. Girard (2002): "Symmetry and dependence properties within a semiparametric family of bivariate copulas," *Journal of Nonparametric Statistics*, 14, 715–727.

Amemiya, T. (1984): "Tobit models: A survey," *Journal of Econometrics*, 24, 3–61.

Andrews, D. (1991): "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59, 307–345.

Andrews, D., and M. Schafgans (1998): "Semiparametric estimation of the intercept of a sample selection model," *Review of Economic Studies*, 65, 497–517.

Arabmazar, A., and P. Schmidt (1982): "An Investigation of the Robustness of the Tobit Estimator to Non-normality," *Econometrica*, 50, 1055–1063.

Arellano, M., and S. Bonhomme (2017): "Quantile selection models with an application to understanding changes in wage inequality," *Econometrica*, 85, 1–28.

Ayer, M., H. Brunk, G. Ewing, W. Reid, and E. Silverman (1955): "An empirical distribution function for sampling with incomplete information," *Annals of Mathematical Statistics*, 26, 641–647.

Baladbaoui, F., C. Durot, and H. Jankowski (2016): "Least squares estimation in the monotone single index model," *working paper*.

Baladbaoui, F., P. Groeneboom, and K. Hendrickx (2017): "Score estimation in the monotone single index model," *working paper*.

Banerjee, M., D. Mukherjee, and S. Mishra (2009): "Semiparametric binary regression models under shape constraints with an application to Indian schooling data," *Journal of Econometrics*, 149, 101–117.

Borjas, G. (1987): "Self-selection and the earnings of immigrants," *American Economic Review*, 77, 531–555.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 125(4), 985–1039.

Cameron, S. V., and J. J. Heckman (1998): "Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males," *Journal of Political Economy*, 106, 262–333.

Cattanoe, M., M. Farrell, and M. Jansson (2018): "Higher-Order refinements of small bandwidth asymptotics for kernel-based semiparametric estimators," *working paper*.

Chen, L. Y., S. Lee, and M. J. Sung (2014): "Maximum score estimation with non-parametrically generated regressors," *Econometrics Journal*, 17, 271–300.

Chen, S. (1997): "Semiparametric estimation of the Type-3 Tobit model," *Journal of Econometrics*, 80, 1–34.

Chen, S., and L.-F. Lee (1998): "Efficient semiparametric scoring estimation of sample selection models," *Econometric Theory*, 14, 423–462.

Chen, S., and Y. Zhou (2010): "Semiparametric and nonparametric estimation of sample selection models under symmetry," *Journal of Econometrics*, 157, 143–150.

Chen, S., Y. Zhou, and Y. Ji (2018): "Nonparametric identification and estimation of sample selection models under symmetry," *Journal of Econometrics*, 202, 148–160.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Cheng, G. (2009): "Semiparametric additive isotonic regression," *Journal of Statistical Planning and Inference*, 139, 1980–1991.

Chernozhukov, V., W. K. Newey, and A. Santos (2015): "Constrained conditional moment restriction models," *arXiv preprint*, arXiv:1509.06311.

Chetverikov, D., A. Santos, and A. Shaikh (2018): "The econometrics of shape restrictions," *Annual Review of Economics*, forthcoming.

Chiquiar, D., and G. H. Hanson (2005): "International Migration, SelfSelection, and the Distribution of Wages: Evidence from Mexico and the United States," *Journal of Political Economy*, 113, 239–281.

Christofides, L. N., Q. Li, Z. Liu, and I. Min (2003): "Recent two-stage sample selection procedures with an application to the gender wage gap," *Journal of Business & Economic Statistics*, 21, 396–405.

Cosslett, S. R. (1983): "Distribution-free maximum likelihood estimator of the binary choice model," *Econometrica*, 51, 765–782.

——— (1991): "Semiparametric estimation of regression model with sample selectivity," in *Nonparametric and semiparametric methods in econometrics and statistics*, pp. 175–197. Cambridge University Press.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric estimation of sample selection models," *Review of Economic Studies*, 70, 33–58.

Esary, J., and F. Proschan (1972): "Relationships among some concepts of bivariate dependence," *Annals of Mathematical Statistics*, 43, 651–655.

Fan, Y., E. Guerre, and D. Zhu (2017): "Partial identification of functionals of the joint distribution of potential outcomes," *Journal of Econometrics*, 197, 42–59.

Fan, Y., and Q. Li (1996): "Consistent model specification tests: omitted variables and semiparametric functional forms," *Econometrica*, 64, 865–890.

Fan, Y., and J. Wu (2010): "Partial identification of the distribution of treatment effects in switching regime models and its confidence sets," *Review of Economic Studies*, 77, 1002–1041.

Freedman, D. A. (1981): "Bootstrapping regression models," *The Annals of Statistics*, 9, 1218–1228.

Gallant, A., and D. Nychka (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica*, 55, 363–390.

Gao, J., and I. Gijbels (2008): "Bandwidth selection in nonparametric kernel testing," *Journal of the American Statistical Association*, 103, 1584–1594.

Grenander, U. (1956): "On the theory of mortality measurement," *Scandinavian Actuarial Journal*, 39, 125–153.

Groeneboom, P., and K. Hendrickx (2018): "Current status linear regression," *The Annals of Statistics*, 46, 1415–1444.

Groeneboom, P., and G. Jongbloed (2014): *Nonparametric estimation under shape constraints*. Cambridge University Press.

Groeneboom, P., and J. A. Wellner (1992): *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser.

Gronau, R. (1974): "Wage comparisons: a selectivity bias," *Journal of Political Economy*, 82, 119–143.

Han, A. K. (1987): "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics*, 35, 303–316.

Heckman, J. J. (1974): "Shadow prices, market wages and labor supply," *Econometrica*, 42, 679–694.

——— (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

——— (1990): "Varieties of selection bias," *The American Economic Review*, 80, 313–318.

Heckman, J. J., and B. E. Honore (1990): "The empirical content of the Roy model," *Econometrica: Journal of the Econometric Society*, 58, 1121–1149.

Heckman, J. J., and R. Robb (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of Econometrics*, pp. 239–267.

——— (1986): "Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes," in *Drawing Inferences from Self-selected Samples*, pp. 63–107. Springer.

Heckman, J. J., J. Tobias, and E. Vytlacil (2003): "Simple estimators for treatment parameters in a latent-variable framework," *Review of Economics and Statistics*, 85, 748–755.

Heckman, J. J., and E. J. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73, 669–738.

——— (2007a): "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," *Handbook of econometrics*, 6, 4779–4874.

——— (2007b): "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments," *Handbook of econometrics*, 6, 4875–5143.

Honoré, B. E., E. Kyriazidou, and C. Udry (1997): "Estimation of Type-3 Tobit models using symmetric trimming and pairwise comparisons," *Journal of Econometrics*, 76, 107–128.

Huang, J. (2002): "A note on estimating a partly linear model under monotonicity constraints," *Journal of Statistical Planning and Inference*, 107, 345–351.

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58.

Imbens, G., and J. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Klein, R. W., and R. H. Spady (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, 61(2), 387–421.

Kline, P. M., and C. R. Walters (2019): "On Heckits, LATE, and numerical equivalence," *Econometrica*, forthcoming.

Kyriazidou, E. (1997): "Estimation of a panel data sample selection model," *Econometrica*, 65(6), 1335–1364.

Lee, L. F. (1978): "Unionism and wage rates: a simultaneous equation model with qualitative and limited dependent variables," *International Economic Review*, 19, 415–433.

——— (1983): "Generalized econometric models with selectivity," *Econometrica*, 51, 507–512.

——— (1994): "Semiparametric two-stage estimation of sample selection models subject to Tobit-type selection rules," *Journal of Econometrics*, 61, 305–344.

Lee, T.-H., Y. Tu, and A. Ullah (2014): "Nonparametric and semiparametric regressions subject to monotonicity constraints: Estimation and forecasting," *Journal of Econometrics*, 182, 196–210.

Lehmann, E. (1966): "Some concepts of dependence," *Annals of Mathematical Statistics*, 37, 1137–1153.

Lemieux, T. (1998): "Estimating the effects of unions on wage inequality in a panel data model with comparative advantage and nonrandom selection," *Journal of Labor*

Economics, 16, 261–291.

Li, Q., and J. Racine (2007): *Nonparametric econometrics: theory and practice*. Princeton University Press.

Li, Q., and J. Wooldridge (2002): "Semiparametric estimation of partially linear models for dependent data with generated regressors," *Econometric Theory*, 18, 625–645.

Liao, X., and M. C. Meyer (2014): "coneproj: An R package for the primal or dual cone projections with routines for constrained regression," *Journal of Statistical Software*, 61, 1–22.

Mammen, E., and K. Yu (2007): "Additive isotone regression," in *Asymptotics: particles, processes and inverse problems*, pp. 179–195. Institute of Mathematical Statistics.

Marchenko, Y. V., and M. G. Genton (2012): "A Heckman selection-t model," *Journal of the American Statistical Association*, 107, 304–317.

Martins, M. F. O. (2001): "Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal," *Journal of Applied Econometrics*, 16, 23–39.

Matzkin, R. L. (1991): "Semiparametric estimation of monotone and concave utility functions for polychotomous choice models," *Econometrica*, 59, 1315–1327.

——— (1993): "Nonparametric identification and estimation of polychotomous choice models," *Journal of Econometrics*, 58, 137–168.

Melino, A. (1982): "Testing for sample selection bias," *Review of Economic Studies*, 49, 151–153.

Meyer, M. (2013): "Semi-parametric additive constrained regression," *Journal of Nonparametric Statistics*, 25, 715–730.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619.

Nelsen, R. B. (2006): *An Introduction to Copulas, 2nd Edition*. Springer.

Newey, W. (2009): "Twostep series estimation of sample selection models," *Econometrics Journal*, 12, 217–229.

Oakes, D. (1989): "Bivariate survival models induced by frailties," *Journal of the American Statistical Association*, 84, 487–493.

Oaxaca, R. (1973): "Male-female wage differentials in urban labor markets," *International Economic Review*, 14, 693–709.

Powell, J. L. (1987): "Semiparametric estimation of bivariate latent variable models," *Working paper*.

Robertson, T., F. Wright, and R. Dykstra (1988): *Order restricted statistical inference*. Wiley.

Robinson, P. (1988): "Root-n consistent semiparametric regression," *Econometrica*, 56, 931–954.

Roy, A. (1951): "Some thoughts on the distribution of earnings," *Oxford Economic Papers*, 3, 135–146.

Schafgans, M. M. (1998): "Ethnic wage differences in Malaysia: parametric and semiparametric estimation of the ChineseMalay wage gap," *Journal of Applied Econometrics*, 13, 481–504.

Schafgans, M. M. (2000): "Gender wage differences in Malaysia: parametric and semiparametric estimation," *Journal of Development Economics*, 63, 351–378.

Sen, B., and M. Meyer (2017): "Testing against a linear regression model using ideas from shape-restricted estimation," *Journal of Royal Statistical Society Series B*, 79, 423–448.

Shorack, G. (2000): *Probability for Statisticians*. Springer.

Spreeuw, J. (2014): "Archimedean copulas derived from utility functions," *Insurance: Mathematics and Economics*, 59, 235–242.

Ullah, A. (2004): *Finite sample econometrics*. Oxford University Press.

Van Der Vaart, A. (1998): *Asymptotic statistics*. Cambridge University Press.

Van Der Vaart, A., and J. A. Wellner (1996): *Weak convergence and empirical processes*. Springer.

Vella, F. (1998): "Estimating models with sample selection bias: a survey," *Journal of Human Resources*, 33, 127–169.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70(1), 331–341.

Willis, R., and S. Rosen (1979): "Education and self-selection," *Journal of Political Economy*, 87, 7–36.

Wooldridge, J. (1995): "Selection corrections for panel data models under conditional mean independence assumptions," *Journal of Econometrics*, 68, 115–132.

Zhang, C.-H. (2002): "Risk bounds in isotonic regression," *The Annals of Statistics*, 30, 528–555.