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 that 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. Using 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 in the presence of the shape restriction. 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

The first draft: October 18, 2018. This version: September 13, 2020.
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 the Co-editor Xiaohong Chen, the Associate Editor, and two referees for their detailed and constructive comments. We are also grateful to Stephane Bonhomme, Yanqin Fan, Jinyong Hahn, Marc Henry, Yu-Chin Hsu, Hide Ichimura, Zheng Li, Jen-Che Liao, Essie Maasoumi, Eric Renault, Peter Robinson, Andres Santos, Le Wang, Ke-Li Xu and seminar/conference participants at Academia Sinica, Indiana University, NCSU, University of Maryland, New York Camp Econometrics XIV, 2019 North American Summer Meeting of the Econometric Society, 2019 CEME conference for Young Econometricians for many thoughtful suggestions and discussions. Yu gratefully acknowledges the support of JSPS KAKENHI Grant Number 19K13666. The usual disclaimer applies.

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 labor supply behavior of females (Heckman, 1974; Gronau, 1974), the determinants of schooling choices (Willis and Rosen, 1979), unionism status (Lee, 1978), and migration decisions (Borjas, 1987), among others. Recently, there has been revived interest in this classical topic by extending the framework to non-separable models (Arellano and Bonhomme, 2017; Chernozhukov, Fernandez-Val, and Luo, 2018; Maasoumi and Wang, 2019) and settings with discrete excluded variables (Brinch, Mogstad, and Wiswall, 2017) or without any exclusion restriction (Honoré and Hu, 2020), further broadening its scope.

A prototypical sample selection model consists of the 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', W_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 $\lambda_0(W_i'\gamma_0) = \mathbb{E}[\varepsilon_i|\nu_i > -W_i'\gamma_0, W_i]$ corrects for the sample selection bias and is known as the control function[1] (Heckman and Robb, 1985). Since the seminal work of Heckman (1979), his two-step method has been the default choice for estimating the sample selection model (1.1). Its original setup assumes the joint normality on the error terms (ε, ν). As a result, the control function has a 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. 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 show 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. The right tail increasing (RTI)

[1]In some alternative formulation (Heckman and Vytlacil, 2007a,b), the control function $\kappa_0(\cdot)$ 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 $\lambda_0(v) = \kappa_0(\bar{F}_\nu(-v))$ 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 κ_0 are equivalent up to a monotone transformation.

represents the positive selection, i.e., the probability that ε takes large values increases with the value of ν. This condition only depends on the copula function without restricting the marginal distributions of the latent errors in the outcome or selection equation. For instance, in the generalized selection model of Lee (1983), a non-negative correlation coefficient of the Gaussian copula entails a decreasing control function,[2] regardless of marginal distribution specifications. In practice, the choice between a positive and negative dependence can be determined by the researcher a priori, when it is possible to postulate whether one gets positive or negative sorting for specific applications. In addition, the comparison of the model fitting associated with imposing a decreasing or an increasing control function is informative about the direction.

Maintaining the monotonicity assumption of the control function, we propose a new semiparametric estimator and a new test for selectivity. Our method is fully data-driven and is free of any tuning parameter to be determined by practitioners. The resulting estimators of the finite-dimensional parameters β_0 and γ_0 are root-n consistent and asymptotically normal. Compared with existing semiparametric procedures using kernel or sieve, our method circumvents the choice of bandwidths in kernel smoothing, penalization parameters in smoothing splines, the order of polynomials in series estimation, and the trimming parameters. To the best of our knowlege, the only existing tuning-parameter-free method for the sample selection model was proposed by Cosslett (1991). His estimator is different from ours and its convergence rate and asymptotic distribution remain unknown[3].

Our estimation method consists of two stages. The first stage focuses on the binary choice data (D_i, W_i) in the selection equation and uses the likelihood function in terms of the coefficient γ and distribution function F_ν of the latent error ν:

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

to obtain estimates $(\hat{\gamma}_n, \hat{F}_{n\nu}(\cdot; \hat{\gamma}_n))$ following from 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 regressors $(W_i'\hat{\gamma}_n)_{i=1}^n$:

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

where λ is restricted to be either a decreasing or increasing function. Note that our estimation method uses two monotonicity restrictions: one on the distribution function F_ν and the other on the control function λ. The nonparametric estimation of F_ν and λ are based on

[2]Throughout this paper, a "decreasing function" refers to a "non-increasing function."
[3]See Remark 3.1 for a detailed comparison between our approach and Cosslett (1991).

the nonparametric maximum likelihood estimator (NPMLE) and isotonic regression. The resulting estimates $\hat{F}$ and $\hat{\lambda}$ are piecewise constant functions with random jump locations and sizes determined by the data itself. 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 of Robertson, Wright, and Dykstra (1988), i.e., their test statistic $\bar{E}_{01}^2$, to our setting. Extensions of our methodology to a panel selection model and to settings with convex restriction are also studied. Both the Monte Carlo simulation and the real data application demonstrate the robust performance of our procedures.

Our main contributions are three-fold. First, we find a sufficient condition for the monotone control function, which is related to an intuitive dependence concept of two latent errors. This demonstrates that the monotonicity of the inverse Mills ratio in the original Heckman model is shared by a much larger family without requiring any parametric assumptions. Therefore, our model with a monotone control function nests some existing parametric generalizations (Olsen, 1980; Lee, 1983; Smith, 2003; Marchenko and Genton, 2012) of the Heckman's model as special cases. Second, our estimation/testing method complements the existing semiparametric approaches (Ahn and Powell, 1993; Das, Newey, and Vella, 2003; Newey, 2009; Li and Wooldridge, 2002) in the sense that it frees practitioners from specifying tuning parameters. The proposed method will be particularly appealing in the scenarios where researchers have certain prior knowledge regarding the dependence between latent errors. For example, Zhou and Xie (2019) noted that individuals who benefit more from college are more motivated than their peers to attend college, exhibiting a pattern of positive selection. Last but not least, from a theoretical perspective, our work contributes to the literature of two-stage estimation and testing that involves shape-restricted nonparametric components. A distinction from the statistics literature (Huang, 2002; Cheng, 2009; Groeneboom and Hendrickx, 2018) is that we have to handle a two-step estimation with the generated regressors $(W_i'\hat{\gamma}_n)_{i=1}^n$. Unlike the sieve or kernel approach adopted by Newey (2009) or Li and Wooldridge (2002), our estimator for the control function is only a piecewise constant function. As a consequence, the estimated control function cannot be simply differentiated to determine the asymptotic influence of $\hat{\gamma}_n$ from the first stage estimation. Taking aim at those challenges, our proofs, which make novel use of the empirical process theory and the characterization of isotonic regression, are also of independent interest. Another notable feature is that we only impose mild moment restrictions on the latent errors without requiring sub-Gaussian or sub-exponential tails, in line with the recent development on the nonparametric least squares estimation with heavy-tailed errors (Han and Wellner, 2018, 2019; Kuchibhotla and Patra, 2019).

Referring to the kernel-based estimation of the sample selection model, which consists of estimating a single-index model (the selection equation) and a partially linear model (the outcome equation), the crucial tuning parameter is the kernel bandwidth. We summarize the existing concepts of optimal bandwidths in three categories; also see the discussion in Section 6.3 of Ichimura and Todd (2007). First, Härdle, Hall, and Ichimura (1993) proposed to *jointly* minimize the finite dimensional parameter and the kernel bandwidth with respect to the sample criterion function such as the sum of squared residuals. For sample selection models, Escanciano and Zhu (2015) emphasized the optimality of tuning parameters in the sense that the estimation error for the semiparametric conditional mean function in (1.2) is minimized. Second, one can directly aim for the nonparametric components such as $\mathbb{E}[Y|D=1,W'\gamma_0], \mathbb{E}[X|D=1,W'\gamma_0]$ in the intermediate step of Robinson (1988), after plugging in some first-stage estimate for γ_0 (Powell, 2001). Even for the same oracle optimal bandwidth, its feasible estimator can be obtained either by the plug-in or cross-validation method. Both approaches have their advantages and limitations, which explains the great variety of choices that are found in the literature (Härdle, Wolfgang, Hall, and Marron, 1988; Ruppert, Sheather, and Wand, 1995; Li and Racine, 2007). Third, when the finite dimensional parameter is of primary interest, a higher-order expansion is necessary, since the bandwidth does not affect the first-order root-n asymptotics. Targeting the mean squared error of the linear coefficient in a partial linear model, Linton (1995) shows that the optimal bandwidth is of order $O(n^{-2/9})$, which differs from the usual one as $O(n^{-1/5})$ in a plain nonparametric model. In light of various bandwidth selectors mentioned above, for practicing empirists, making the right choice among them can be a skillful task. The paucity of this skill may have impeded a broader appreciation and adoption of semiparametric methods in applications. We are not claiming any theoretical superiority of the proposed approach over the existing methods in which tuning parameters are carefully chosen; instead, our work provides applied researchers with an alternative path to circumvent such a delicate choice.

1.1. Related Literature

This paper joins a rich and evolving literature on shape-restricted estimation and inference, where the underlying criterion function can be meaningfully maximized (or minimized) without additional penalization or smoothing under the maintained shape restriction (Cosslett, 1983; Matzkin, 1991; Banerjee, Mukherjee, and Mishra, 2009; Groeneboom and Hendrickx, 2018; Horowitz and Lee, 2017). See Groeneboom and Jongbloed (2014) and Chetverikov, Santos, and Shaikh (2018) for comprehensive reviews in statistics and econometrics. In this paper, we take a major step to introduce the shape restriction to the

control function in sample selection models by converting an intuitive concept regarding the dependence between latent errors into a precise condition known as right tail increasing (decreasing).

The copula function, which we utilize to establish the monotonicity of control functions, has been an essential tool in the sample selection or generalized Roy models. It has been employed to estimate sample selection models with non-normal errors (Lee, 1983; Smith, 2003), obtain bounds on distributional treatment effects (Fan and Wu, 2010; Fan, Guerre, and Zhu, 2017), and aid in identification and inference for non-separable models (Arellano and Bonhomme, 2017; Chernozhukov, Fernandez-Val, and Luo, 2018; Maasoumi and Wang, 2019). The copula function plays a different role in our context: the proposed sufficient condition for the monotonicity of the control function only depends on the copula function of the latent errors. In addition, this condition is easy to check for commonly used copula families.

The joint normality assumption on (ε, ν) is more for convenience than necessity for the sample selection model. Indeed, misspecification of distributions leads to inconsistent estimates and invalid inference, motivating the development of non-normal parametric selection models (Olsen, 1980; Lee, 1983; Smith, 2003; Marchenko and Genton, 2012) and more flexible semi and 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; Ahn and Powell, 1993; Andrews and Schafgans, 1998; Chen and Lee, 1998; Das, Newey, and Vella, 2003). 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 semiparametric methods have been plagued by issues of sensitivity of estimates to choices of smoothing parameters, trimming parameters, and the like."

Our paper improves on Cosslett (1991) who was among the first to propose a tuning-parameter-free semiparametric estimator for the sample selection model. The work of Cosslett (1991) has been highlighted by a number of influential reviews (Heckman, 1990; Vella, 1998; Pagan and Ullah, 1999; Heckman and Vytlacil, 2007a) and its convenient feature continues to attract applied researchers in empirical studies (Francesconi and Nicoletti, 2006; Berman, Rebeyrol, and Vicard, 2019). However, Cosslett (1991) only presented the consistency proof based on a sample-splitting argument, whereas the rate of convergence and the asymptotic distribution remain unknown. Our work is inspired by Cosslett (1991), but the key distinction between our approach and his is that by imposing the monotonicity restriction on the control function, our second stage sets a well-posed minimization problem with fully data determined jump locations in the nonparametric estimator of the control

function. Furthermore, building on Groeneboom and Hendrickx (2018) and Sen and Meyer (2017), we establish the complete asymptotic theory.

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 test for the presence of sample selection bias. Section 4 establishes the asymptotic results. Section 5 extends our methodology to a two-period panel selection model and to the convex restriction. 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 the Appendix; detailed analysis of examples, proofs of technical lemmas, and additional simulation results are collected in the supplementary notes.

Throughout the paper, we work with the i.i.d. data $Z_i = (Y_i, D_i, X_i', W_i')$ 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 (Andrews and Schafgans, 1998). Let q denote the dimensionality of covariates W and we write $\gamma_0 \equiv (\gamma_{01}, \gamma_{02}, ..., \gamma_{0q})'$. Following Ichimura (1993) and Klein and Spady (1993), a normalization by taking $\gamma_{01} = 1$ is adopted, and the vector W is partitioned as $W = (W_1, W_{-1}')'$ accordingly. We also write $\gamma_0 = (1, \gamma_{0-}')'$ and denote our estimator by $\hat{\gamma}_n = (1, \hat{\gamma}_{n-}')'$. We use the standard empirical process notations as follows. For a function $f(\cdot)$ of a random vector Z 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 $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n - P)f$. To simplify the notations for the outcome equation, we write $\mathbb{P}_{1n} f = \mathbb{P}_n(Df(Z))$, $P_1 f = P(Df(Z))$, and $\mathbb{G}_{1n} f = \mathbb{G}_n(Df(Z))$.

2. Monotonicity of the Control Function

The sample selection bias arises when the latent error terms ν and ε in selection and outcome equations are dependent, which leads to a non-constancy control function $\lambda(W'\gamma_0) = \mathbb{E}[\varepsilon|\nu > -W'\gamma_0, W]$. In Heckman's original set up (Heckman, 1974, 1979), the control function is proportional to the well-known inverse Mills ratio $\phi(\cdot)/\Phi(\cdot)$, which is a decreasing function due to the log-concavity of normal distribution. It is natural to ask whether monotonicity of the control function is a more general feature shared by other families of distributions of (ε, ν). This section provides an affirmative answer by showing that a specific positive (negative) dependence condition between the latent error terms ν and ε

is sufficient to induce a monotone control function. Beyond the standard correlation coefficient, there exists a wealth of notions characterizing the positive (negative) dependence between two random variables (Nelsen, 2006, Chapter 5.2). Among them, the dependence condition for our purpose is *right tail increasing (decreasing)*, first defined by Esary and Proschan (1972) as follows.

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

Intuitively, $RTI(\varepsilon|\nu)$ is a positive dependence condition in the sense that ε is more likely to take on large values as ν increases.The following result establishes the precise link between the positive (negative) dependence and the monotonicity of the control function.

Theorem 2.1. *If ε is right tail increasing (decreasing) in ν, then the control function $\lambda(t) = \mathbb{E}[\varepsilon|\nu > -t]$ is decreasing (increasing).*

In the sample selection problem, it is natural to expect certain positive (negative) dependence between errors in the selection and outcome equations. Consider a simple Roy model (Heckman and Vytlacil, 2007a). The outcomes Y_1 and Y_0 are wages earned in two sectors (for example, formal and informal sectors) with following specifications:

$$Y_1 = X'\beta_1 + u_1 \text{ and } Y_0 = X'\beta_0 + u_0.$$

Assume there is a switching cost $C = \tilde{W}'\beta_C$, and let u_1 and u_0 be independent.[4] An optimizing agent self-selects into the sector with a higher wage net of the switching cost:

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

This gives rise to a sample selection model where the researcher only observes the wage in sector 1 (the formal sector), i.e., $Y = D \times Y_1$. Using the notation of model (1.1), two latent errors are $\varepsilon = u_1$ and $\nu = u_1 - u_0$. It is intuitive that the positive dependence between ε and ν stems from the common part u_1. Heuristically speaking, when $u_1 - u_0$ is larger, it is more likely that u_1 is large as well. The supplementary notes (Section S1.1) verify that $RTI(\varepsilon|\nu)$ holds for various continuous distributions of original errors u_1 and u_0.

Next, we present several parameterized joint distributions of (ε, ν) that yield monotone control functions. By Theorem 5.2.5 of Nelsen (2006), RTI/RTD is a dependence concept that relates the ranks of random variables, so it only depends on the copula function $C(\cdot, \cdot)$ of (ε, ν), regardless of marginal distributions. To avoid repetition, we focus on the version of positive dependence and thus a decreasing control function.

[4]The dependence structure of (u_1, u_0) is only partially identified in the Roy model (Fan and Wu, 2010). The independence assumption herein is merely for illustration purpose.

Example 2.1. (Gaussian Copula) The monotonicity of the control function in Heckman's model resides in the Gaussian copula $C(u, v; \rho) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$, where ρ denotes the correlation coefficient. Without restricting the marginal distribution to be Gaussian, Lee (1983) proposed a generalized selection model with arbitrary (but known) marginal distributions coupled with the Gaussian copula. For Gaussian copula models, it is straightforward to check that RTI is equivalent to $\rho \geq 0$, see the supplementary notes (Section S1.2) for details.

Example 2.2. (Archimedean Copula) Smith (2003) embedded Archimedean copulas to the sample selection model and developed the corresponding maximum likelihood estimation. An Archimedean copula $C(u, v)$ can be expressed as

$$C(u, v) = \psi^{[-1]}\left(\psi(u) + \psi(v)\right),$$

where ψ is the generator function and $\psi^{[-1]}$ represents its generalized inverse. Spreeuw (2014) showed that the RTI of an Archimedean copula is equivalent to the cross-ratio function[5] being greater or equal to 1. A popular family of Archimedean copula is the Clayton copula:

$$C(u, v; \alpha) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad 0 \leq u, v \leq 1, \tag{2.1}$$

where the parameter $\alpha \geq 0$. Its generator function is $\psi(u; \alpha) = u^{-\alpha} - 1$ and the cross-ratio function becomes $\alpha + 1$. Hence, the whole Clayton copula family satisfies RTI.

Example 2.3. (Generalized FGM Copula) A copula function belongs to the generalized Farlie-Gumbel-Morgenstern (FGM) family if $C(u, v; \tau) = uv + \tau\varphi(u)\varphi(v)$ for a scale parameter $\tau \in [-1, 1]$ (Amblard and Girard, 2002), where φ is the generator function. By Theorem 3 of Amblard and Girard (2002), when $\tau > 0$, this family is RTI if and only if $\varphi(u)/(u-1)$ is monotone. The original FGM copula specifies $\varphi(u) = u(1-u)$ so that $C(u, v; \tau) = uv + \tau uv(1-u)(1-v)$. Then $\varphi(u)/(u-1) = u$, which confirms RTI for the FGM copula when $\tau > 0$.

Theorem 2.1 states $RTI(\varepsilon|\nu)$ as a sufficient condition for the monotonicity of the control function $\lambda(t)$. On some occasions, it might be easier to directly verify the monotonicity of $\lambda(t)$ rather than going through the sufficient condition; the following normal mixture model, which was used by Cosslett (1991) in the Monte Carlo simulation, serves as an example.

---

[5] The cross-ratio function is defined by $CR(u) = -u\frac{\psi^{(2)}(u)}{\psi^{(1)}(u)}$ for $u \in [0, 1]$, where $\psi^{(j)}$ denotes the j-th order derivative of the generator ψ, for $j = 1, 2$.

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

$$f_{\varepsilon,\nu}(s,t) = \pi g(s,t;\sigma_1,\sigma_2,\rho) + (1-\pi)g(s,t;k\sigma_1,k\sigma_2,\rho), \tag{2.2}$$

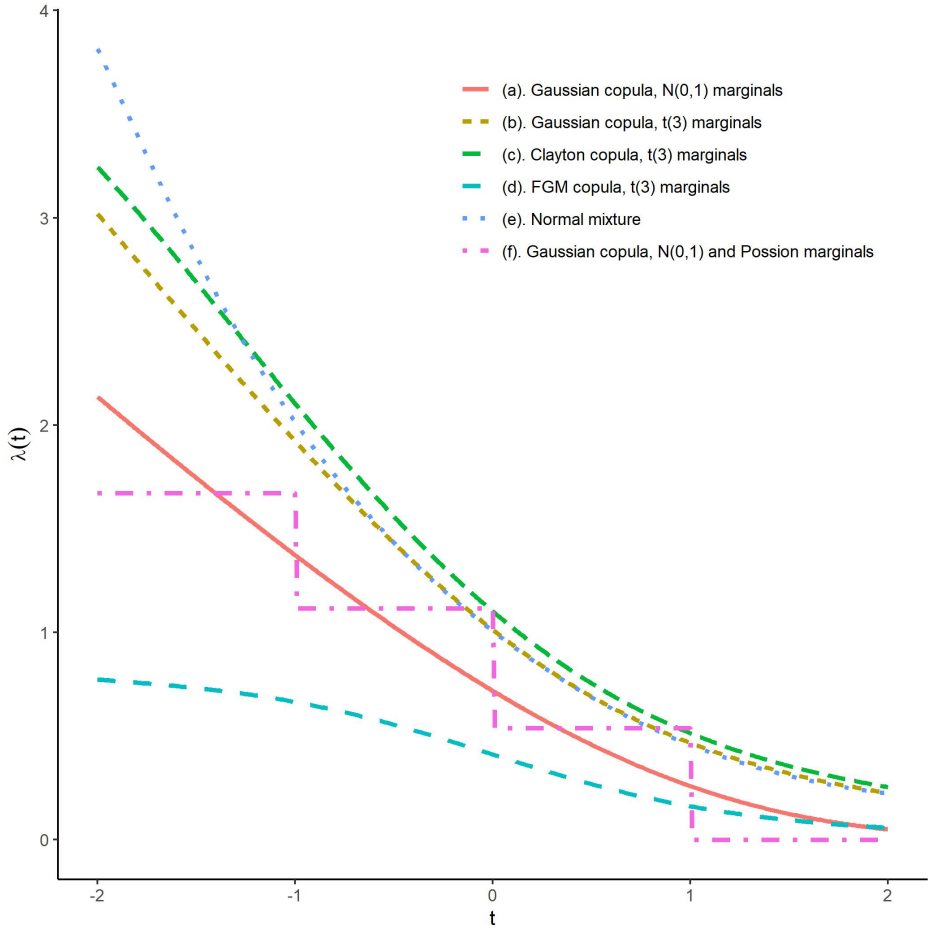
where the second normal component has a covariance matrix amplified by $k > 1$. The supplementary notes (Section S1.2) prove that the resulting control function is decreasing in t.

Figure 1 displays profile plots of control functions $\lambda(t) = \mathbb{E}[\varepsilon|\nu \geq -t]$ for various joint distributions of the error terms (ε,ν); (a) and (b) represent the Gaussian copula (Example 2.1 with $\rho = 0.9$): (a) has standard normal marginals and (b) has $t(3)$ marginals (a t-distribution with the degree of freedom equal to 3). Fixing the $t(3)$ margin, (c) corresponds to the Clayton copula (Example 2.2, equation (2.1) with $\alpha = 50$) and (d) to the FGM copula (Example 2.3 with $\tau = 1$). In addition, (e) represents a normal mixture described in Example 2.4 with $\rho = \pi = 0.9$, $\sigma_1 = \sigma_2 = 1$, and $k = 5$. Our framework also allows for discrete errors that lead to non-smooth control function, as (f) corresponds to the scenario where ν has a Poisson distribution and ε remains standard normal. All the control functions depicted in Figure 1 are decreasing by construction, but their shapes differ by copula functions and by marginal distributions. Apart from the aforementioned cases, the sample selection model with an imposed linear control function (Olsen, 1980; Newey, 1999) and the Heckman t-selection model from Marchenko and Genton (2012) (see their Figure 1) also satisfy our monotone restriction.

Remark 2.1. *Besides RTI, the stochastic increasing (SI) is another popular measure of positive dependence that has been used in sample selection models.*[6] *Recently, Honoré and Hu (2020) show that $SI(\varepsilon|\nu)$ sharpens the partial identification results of a selection model without any exclusion restriction. SI is stronger than RTI (Nelsen, 2006, Theorem 5.2.12). The following scenario illustrates the connection and difference between RTI and SI. Consider a typical selection problem in the labor market where ε stands for the unobservable in the wage equation (call it "productivity") and ν stands for the willingness to work (call it "motivation"). Suppose that the motivation positively contributes to the productivity in a linear additive form $\varepsilon = \theta\nu + u$ (for an independent random error u). For a positive constant coefficient θ, both $RTI(\varepsilon|\nu)$ and $SI(\varepsilon|\nu)$ hold, following Lehmann (1966, Example*

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

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



5). However, if θ (the rate that motivation transforms to productivity) is decreasing with ν, then $SI(\varepsilon|\nu)$ is likely to fail, but $RTI(\varepsilon|\nu)$ can still hold. A numerical example is provided in the supplementary notes (Section S1.3). In addition, SI requires differentiability of the copula function; see Nelsen (2006, Exercise 5.32) for another example that satisfies RTI but not SI.

3. Shape-restricted Estimation and Testing

In this section, we propose a shape-restricted two-stage semiparametric estimation method of $(\beta, \gamma, \lambda(\cdot), F_\nu(\cdot))$ that sidesteps any user-specified tuning parameter. We also develop a new sensitivity test for the presence of sample selection bias.

3.1. A Shape-restricted Two-stage Estimator

Our approach is inspired by Cosslett (1991) in the sense that we obtain a two-stage semi-parametric estimation method making use of a shape-restricted estimation of nonparametric components, i.e., the estimators $\hat{F}_{n\nu}$ and $\hat{\lambda}_n$ are step functions with data-determined jump locations, so that users do not need to provide tuning parameters. 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 asymptotically 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 use the isotonic regression technique (Robertson, Wright, and Dykstra, 1988). Although Cosslett (1991) did not require the monotonicity of λ_0, his method restricts the jump locations in the estimated control function to be the same as those of NPMLE from the first stage.

The detailed procedure is described as follows. Fixing the parameter γ, we consider the values of $\bar{V}_1^{(\gamma)} = -W_1'\gamma, \cdots, \bar{V}_n^{(\gamma)} = -W_n'\gamma$. Denote $\bar{V}_{(1)}^{(\gamma)} \leq \cdots \leq \bar{V}_{(n)}^{(\gamma)}$ as the order statistics with corresponding indicators $\bar{D}_{(i)}^{(\gamma)}$ for $i = 1, \cdots, n$. Let $\mathcal{D}$ and $\mathcal{I}$ be the spaces of decreasing and increasing functions, respectively.

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 resulting estimate $\hat{F}_{n\nu}(\cdot;\gamma)$ is the left derivative of the convex minorant of a cumulative sum diagram consisting of the points $(0,0)$ and $\left(i, \sum_{j=1}^{i} \bar{D}_{(j)}^{(\gamma)}\right)$ for $i = 1, \cdots, n$.

Stage 1(ii). Given $\hat{F}_{n\nu}(\cdot;\gamma)$, our estimator $\hat{\gamma}_n$ for the coefficients is the zero-crossing point of the estimating equation:[7]

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

Stage 2. Given $\hat{\gamma}_n$ from Stage 1, we estimate β and $\lambda(\cdot)$ by the least squares estimator:

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

[7]For the identification purpose, we simply normalize the first coordinate of $\hat{\gamma}_n$ to be 1 and solve for the other coordinate values $\hat{\gamma}_{n-}$ from the estimating equations.

where $\mathcal{G} = \mathcal{D}$ or $\mathcal{I}$, depending on the monotonicity restriction on λ_0 is decreasing or increasing.

The computation of each step is straightforward and can be implemented through the existing R packages. The NPMLE $\hat{F}_{n\nu}(\cdot;\gamma)$ in Stage 1(i) is a piece-wise constant function and it is obtained via the standard pool adjacent-violators algorithm (PAVA); see Chapter 1 of Robertson, Wright, and Dykstra (1988). Despite the discreteness of $\hat{F}_{n\nu}(\cdot;\gamma)$, Groeneboom and Hendrickx (2018) showed the global uniqueness of the zero-crossing point of the estimating equation (3.2) asymptotically, which facilitates the computation of Stage 1(ii). Here we resort to the (modified) Barzilai-Borwein (BB) method, which is efficient in solving the zero-crossing points in large-scale nonlinear systems of equations. The corresponding R package BB (Varadhan and Gilbert, 2009) is also available. The optimization problem (3.3) in Stage 2 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 available in the R package "coneproj" (Liao and Meyer, 2014) can be directly applied to obtain $(\hat{\beta}_n, \hat{\lambda}_n)$.

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). Within the context of binary choices models, the NPMLE is used by Cosslett (1983) to define the profile maximum likelihood estimator. However, only the consistency result is available for Cosslett's estimator. The key modification to achieve a root-n consistency and asymptotic normality for γ_0 while maintaining the tuning-parameter-free feature is to use the Z-estimator from Groeneboom and Hendrickx (2018) in Stage 1 (ii); see their discussion on [p.1420] about the difficulty of Cosslett's profile MLE. Essentially, one makes use of the population-level moment condition:

$$\mathbb{E}[W_{-1}(\bar{D} - F_\nu(-W'\gamma_0))] = 0, \tag{3.4}$$

and plug in the first-step estimator $\hat{F}_{n\nu}(\cdot;\gamma)$ in the sample analog. In comparison, the efficient score function derived from the smoothed maximum likelihood estimator of Klein and Spady (1993) is

$$\mathbb{E}\left[\frac{f_\nu(-W'\gamma_0)}{F_\nu(-W'\gamma_0)(1-F_\nu(-W'\gamma_0)))}W_{-1}(\bar{D} - F_\nu(-W'\gamma_0))\right] = 0, \tag{3.5}$$

which involves the additional weighting factor $\frac{f_\nu}{F_\nu(1-F_\nu)}$. In practice, the trimming is inevitable (as specified in Condition C.7 of Klein and Spady (1993)) for the Klein-Spady estimator or the efficient shape-restricted estimator in Section 4.2 of Groeneboom and Hendrickx (2018) due to the instability of the denominator $F_\nu(1-F_\nu)$, whereas it can be

dispensed with in the estimating equation (3.2).[8] Referring to Stage 2, assuming a monotone control function, it becomes straightforward to run the partial linear isotonic regression (Huang, 2002) after the inclusion of $W'\hat{\gamma}_n$ to control for the selection bias. In analog with the sieve type estimator (Newey, 2009) that searches for the best possible fit of the model once the approximating basis is determined, the NPMLE or isotonic regression also seeks to minimize the estimation error within the monotone class. The imposed monotonicity is sufficiently regular to make the maximization/minimization problem well-defined without additional smoothing or penalization.

Remark 3.1. *Cosslett (1991) proposed an ingenious two-step procedure in which no tuning parameter is needed and proved the consistency property. He first estimated γ_0 and $F_{\nu 0}(\cdot)$ by the profile maximum likelihood estimator from Cosslett (1983). His 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) estimated the outcome equation by approximating the control function $\lambda(\cdot)$ based on K_n indicator variables $\{\mathbb{I}(W'\tilde{\gamma}_n \in I_j)\}_{j=1}^{K_n}$. The optimization in his second stage can be viewed as finding the sieve type approximation by step-wise functions with predetermined window widths $(I_j)_{j=1}^{K_n}$. There is arguably certain degree of arbitrariness to force the set of jump locations (or window widths) to be the same for nonparametric components in both stages, which also complicates the subsequent asymptotic analysis. The most important distinction of our method is that we impose the monotonicity restriction on the control function $\lambda(\cdot)$, so that the second stage minimization problem is well posed without specifying the jump locations a priori. Although our estimated $\hat{\lambda}_n(\cdot)$ is also a piecewise constant function, it is monotone and the jump locations are automatically determined by the shape-restricted optimization in the second stage. The resulting estimate $\hat{\lambda}_n$ (together with the linear part $x'\hat{\beta}_n$) provides the best possible fit among the monotone class for the semiparametric model (1.2) in the same spirit of Härdle, Hall, and Ichimura (1993); Escanciano and Zhu (2015).*

3.2. A Shape-restricted Test for Selectivity

Under the null hypothesis of no selectivity bias, Heckman (1979) proposed a t-test on the regression coefficient attached to the inverse Mill's ratio. The t-test in Heckman (1979) is the Lagrange multiplier (LM) test statistic in this context (Vella, 1998).

[8]On one hand, if a fixed trimming is specified using an a priori chosen set, it yields a loss of efficiency. On the other hand, a data-dependent trimming that expands the whole support asymptotically requires additional assumptions about the tail behavior of ν and there is no consensus on the *optimal* trimming; see Section 6.4 of Ichimura and Todd (2007).

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. Based on that, we have developed a new test to detect the sample selection, which does not require user-determined tuning parameters . To focus on the main idea, we consider the case where there is 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 $\mathcal{S}_0 = \mathcal{X} \oplus \mathcal{C}$ and the alternative space as $\mathcal{S}_1 = \mathcal{X} \oplus \mathcal{D}$. 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.[9]

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

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

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, $\mathbf{Y} - \Pi(\mathbf{Y}|\mathcal{S}_0)$ is simply the residual vector from the ordinary least squares (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)). Sen and Meyer (2017) shows that the null critical value for this type of test statistic can be approximated by the bootstrap method. Because the control function boils down to a constant term under H_0 in our context, 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 $\tilde{\beta}_n$

[9]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.

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

$$T_{n,b}^* = \frac{\| \Pi(\mathbf{Y}_b^*|\mathcal{S}_0) - \Pi(\mathbf{Y}_b^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D}^2}{\| \mathbf{Y}_b^* - \Pi(\mathbf{Y}_b^*|\mathcal{S}_0) \|_{n,D}^2}. \tag{3.7}$$

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

Remark 3.2. *The monotone restriction on $\lambda_0(\cdot)$ is testable in principle, as one can compare the global difference (say in terms of the L_1-norm) between our shape-restricted estimator $\hat{\lambda}_n$ versus an unrestricted kernel-type estimator $\bar{\lambda}_{nh}$ with its smoothing bandwidth denoted by h. Section 13.2 of Groeneboom and Jongbloed (2014) developed the asymptotic theory of this type of test for purely nonparametric problems. It is interesting to observe that the test statistic can be asymptotically normal with proper centering and rescaling. It is expected that the centering and normalization terms might change due to additional complications from the first-stage estimation in our setup for this test statistic. We leave a rigorous investigation for future work.*

4. Main Theoretical 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 3.2 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 $U_i = D_i(X_i - \mathbb{E}[X_i|D_i = 1, V_i])$. We assume $H_\beta \equiv \mathbb{E}[U_i U_i']$ is non-singular. Moreover, we define the centered error term as

$$\epsilon_i = D_i(Y_i - X_i'\beta_0 - \lambda_0(V_i)) \tag{4.1}$$

with $\Sigma \equiv \mathbb{E}[\epsilon_i^2 U_i U_i']$ and $H_\gamma \equiv \mathbb{E}[U_i \frac{\partial \lambda_0(v_i)}{\partial v_i} W_{i,-1}]$. Regarding the first-stage estimation, the NPMLE $\hat{F}_{n\nu}$ in Cosslett (1983) provides an estimate of

$$F_\nu(u;\gamma) \equiv P\left\{\bar{D}^{(\gamma)}| - V^{(\gamma)} = u\right\} = \int F_{\nu 0}(u - w'(\gamma_0 - \gamma)) f_{W|W'\gamma}(w| - W'\gamma = u)dw, \tag{4.2}$$

for any fixed γ; see Groeneboom and Hendrickx (2018). In the sequel, we also denote its density by $f_\nu(u;\gamma)$ and its inverse by $Q_\nu(t;\gamma)$. $F_{\nu 0}(u)$ and $f_{\nu 0}(u)$ are used for $F_\nu(u;\gamma_0)$ and $f_\nu(u;\gamma_0)$.

The following regularity conditions will be assumed throughout the paper.

Condition 1. The latent error terms (ε, ν) are independent of (X, W). The probability $\Pr\{D = 1\}$ is bounded away from zero.

Condition 2. The random variable Y and each coordinate of X have bounded second moments. We also assume the centered error term ϵ satisfies the moment bound $\mathbb{E}[\epsilon^2|X, W] \leq \sigma_0^2$ a.s., and there exists a finite $r > 3$ such that $\mathbb{E}|\epsilon|^r \leq \infty$.

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 $\mathcal{R}^{q-1}$, respectively.

Condition 5. The true monotone control function λ_0 is continuously differentiable with its derivative denoted by $\dot{\lambda}_0(\cdot)$.

Condition 6. The function $F_\nu(\cdot;\gamma)$ has a continuous positive derivative for all γ. 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. The density $f_\nu(u;\gamma)$ and conditional expectations $\mathbb{E}[W_{-1}|W'\gamma = u]$ and $E[W_{-1}W'_{-1}|W'\gamma = u]$ are twice continuously differentiable with respect to u. The functions $\gamma \mapsto f_\nu(u;\gamma)$, $\gamma \mapsto \mathbb{E}[W_{-1}|W'\gamma = u]$, and $\gamma \mapsto \mathbb{E}[W_{-1}W'_{-1}|W'\gamma = u]$ are continuous functions for u in the definition domain and all γ in the parameter space. The matrices $E[XX'|D = 1]$ and H_γ are of full rank.

Condition 7. Denote the conditional mean functions by $\chi(u) \equiv \mathbb{E}[X|D = 1, W'\gamma_0 = u]$ and $\varpi(u;\gamma) \equiv \mathbb{E}[W_{-1}|W'\gamma = u]$. Assume that $\chi \circ \lambda_0^{-1}(s)$ is Lipschitz continuous and $\Psi(t;\gamma) = \varpi \circ Q_\nu(t;\gamma)$ is Lipschitz continuous for any γ with the Lipschitz constant independent of γ.

Most of our 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 main novelty is that we do not require sub-exponential tails on the centered errors, compared with Mammen and Yu (2007) and Cheng (2009). The weaker moment restriction in Condition (2) brings two extra steps in establishing the convergence rate. First, we rely on the Montgomery-Smith inequality to establish the stochastic boundedness of $\hat{\lambda}_n$ (Han and Wellner, 2018) instead of using the rough

union bound under sub-exponential tails. Second, employ a truncation argument similar to Chen and Shen (1998) and Kuchibhotla and Patra (2019) in the peeling argument for the empirical process involved. In order to handle the unbounded part, we need a precise control on the local envelope for the monotone function from Giné and Koltchinskii (2006); Han and Wellner (2018). The uniform bound on the conditional heteroskedasticity in Condition (2) allows us to convert the entropy integral for the underlying partial linear function to the one that works for the process involving the multiplier ϵ. This condition cannot be easily dispensed with, considering the impossibility result in Han and Wellner (2019). The imposed (higher than the third-order) moment condition on the error term guarantees a cubic root (modulo the logarithm term) rate for the monotone function, otherwise the rate can be slower depending on the tail; see Han and Wellner (2018); Kuchibhotla and Patra (2019). Another condition that we want to emphasize concerns the exclusion restriction of W in Condition (3). 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 S6.

As our first step estimation follows from Groeneboom and Hendrickx (2018), we state asymptotic results of $\hat{\gamma}_n$ and $\hat{F}_{n\nu}(\cdot;\gamma)$ in the following lemma . We provide a separate proof in the supplementary notes that allows for the large support of $V = W'\gamma_0$. This is made possible under the global Lipschitz condition, which is similar to condition (A3) in Huang (2002). The large support condition is important for the identification purpose regarding the first-stage binary choice model[10].

Lemma 4.1. *Under Conditions 1 to 7, we have*

$$n^{1/2}\left(\hat{\gamma}_{n-} - \gamma_{0-}\right) \Rightarrow \mathbb{N}(0, V_\gamma), \tag{4.3}$$

where V_γ is equal to $A^{-1}BA^{-1}$ with

$$A = \mathbb{E}\left[f_{\nu 0}(-W'\gamma_0)\left\{W_{-1} - E[W_{-1}|W'\gamma_0]\right\}^{\otimes 2}\right] \text{ and} \tag{4.4}$$

$$B = \mathbb{E}\left[\left\{(F_{\nu 0}(-W'\gamma_0) - \bar{D})(W_{-1} - E[W_{-1}|W'\gamma_0])\right\}^{\otimes 2}\right]. \tag{4.5}$$

Also, one gets

$$\left(\int \left(\hat{F}_{n\nu}(-w'\hat{\gamma}_n;\hat{\gamma}_n) - F_{\nu 0}(-w'\gamma_0)\right)^2 dF_W(w)\right)^{1/2} = O_p(\log n \times n^{-1/3}). \tag{4.6}$$

[10]We want to thank one anonymous referee for suggesting this extension.

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:

$$\| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| \equiv \left(\int \left(\hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \right)^2 f_{W|D=1}(w) dw \right)^{1/2}, \tag{4.7}$$

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

Theorem 4.1. *Suppose Conditions 1 to 7 hold, then one has the following result:*

$$|\hat{\beta}_n - \beta_0| + \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| = O_p(n^{-1/3} \log n). \tag{4.8}$$

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 $\hat{\gamma}_n$, so it implies the uniform convergence of $\hat{\lambda}_n$ to λ_0 over a given compact set within the interior of the support. The proof follows from the same argument as in Corollary 5.3 of Baladbaoui, Durot, and Jankowski (2019) with proper adjustment on the range for ω_n, as the first-stage estimator $\hat{\gamma}_n$ converges at the root-n rate.

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

$$\left(\int_{\underline{v}+\omega_n}^{\bar{v}-\omega_n} \left(\hat{\lambda}_n(v) - \lambda_0(v) \right)^2 dv \right)^{1/2} = O_p(n^{-1/3} \log n) \tag{4.9}$$

for all sequences of 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) focused 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, 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) relied on the sample-splitting trick in which the selection and outcome equations are estimated using separate subsamples. A rigorous proof based on the full sample is absent in Cosslett (1991).*

Remark 4.2. *The cubic-root rate in Lemma 4.2 seems to suggest that the control function is estimated with a slower rate than the kernel competitor; refer to Pagan and Ullah (1999) and Li and Racine (2007). In standard sample selection problems, the primary interest is on the regression coefficient β_0, whereas the control function λ_0 is included only to control for the selection bias, as its name suggests. We emphasize that the cubic-root rate for the control function is obtained without assuming its second-order differentiability, which is needed for kernel/sieve estimators to achieve the rate of $O_p(n^{-2/5})$. In fact, the shape-restricted estimators match the minimax rate for the first-order differentiable monotone functions; see Chapter 6 in Groeneboom and Jongbloed (2014). Also, the smoothness conditions on the marginal distribution F_ν or the control function λ_0 in Conditions 5 and 6 are only used for establishing the asymptotic normality of finite dimensional parameters and they are not required for the consistency of our estimator. In contrast, the smoothness assumption is the foundation of any kernel smoothing method. Finally, we show that a rate like $O_p(n^{-2/5})$ can be achieved under the convex constraint for the extension in Section 5.2.*

The next corollary states that the SSR associated with a decreasing or increasing control function is informative about the direction of the selection bias.[11]The comparison of two SSRs therefore provides a data-driven way to check whether the control function is decreasing or increasing, which complements the prior knowledge of the researcher about the selection direction. Naturally, one would pick the model that fits the data better. When the selection pattern is completely unknown, the kernel or sieve approach without shape restrictions is probably more suitable.

Corollary 4.1. *For the true control function being decreasing, i.e., $\lambda_0 \in \mathcal{D}$, let $(\tilde{\beta}_n^{\mathcal{I}}, \tilde{\lambda}_n^{\mathcal{I}})$ be the estimated regression coefficient and control function obtained from (3.3) with $(\beta \in \mathbf{B}, \lambda \in \mathcal{I})$, i.e., λ is restricted to be an increasing function in that optimization problem. Let $SSR_{\mathcal{D}}$ and $SSR_{\mathcal{I}}$ be*

$$SSR_{\mathcal{D}} \equiv \sum_{i=1}^{n} D_i[Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)]^2,\ SSR_{\mathcal{I}} \equiv \sum_{i=1}^{n} D_i[Y_i - X_i'\tilde{\beta}_n^{\mathcal{I}} - \tilde{\lambda}_n^{\mathcal{I}}(W_i'\hat{\gamma}_n)]^2,$$

under the correct specification and misspecification, respectively. Then we have $SSR_{\mathcal{I}} \geq SSR_{\mathcal{D}}$ with probability approaching 1 as $n \to \infty$.

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 restrictions, the control function is estimated using the monotonicity restriction in the outcome equation for our model. As a consequence, the estimated control function $\hat{\lambda}_n(\cdot)$ is piecewise constant with random jump locations and it

[11]We would like to thank Yu-chin Hsu for this point.

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; Han and Wellner, 2018).

Theorem 4.2 (Asymptotic Normality)**.** *Suppose Conditions 1 to 7 hold, then we get*

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \Rightarrow \mathbb{N}(0, V_\beta), \tag{4.10}$$

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 4.3. *The asymptotic variance matrix for $\hat{\beta}_n$ takes the generic form of the 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 (the shape restriction on λ_0 does not alter this part), whereas $H_\beta^{-1}H_\gamma V_\gamma H_\gamma' H_\beta^{-1}$ captures the effect from estimating γ_0 in the first stage. The use of Groeneboom and Hendrickx (2018) in the first stage estimation results in a larger V_γ than the efficient Klein-Spady estimator. Apart from the trimming issue related to the efficient score function (3.5), our Monte Carlo results in Section S.4 show that the efficiency gain from the Klein-Spady estimator only actualizes with the cross-validated bandwidth from Härdle, Hall, and Ichimura (1993) in very large sample, which unfortunately incurs significant computation burden. Note that the generic two-step procedures fail to achieve the semiparametric efficiency bound for sample selection models. A fully efficient method has to be based on maximizing the (smoothed) joint likelihood function combining both selection and outcome equations or the one-step update of the efficient score function using a pilot root-n consistent estimator (Chen and Lee, 1998). Nonetheless, the inefficiency does not hinder the popularity of two-step procedures in applications.*

4.2. Validity of the Selectivity Test

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 $(Y_i, D_i, X_i', W_i')_{i=1}^n$. Furthermore, we define the vector $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)'$. For two distribution functions, F_1 and F_2, the Lévy distance d_L is defined as

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

The Lévy distance metrizes the weak convergence (Shorack and Wellner, 2009).

Theorem 4.3. *Assume Conditions 1 to 7 hold. Also, suppose the centered error term ϵ satisfies that $\sigma_1^2 \equiv \mathbb{E}[\epsilon^2|D=1] > 0$ and the moment bound in Condition 2 holds with $r \geq 4$, then we have*

$$d_L(H_n, H_n^*) \to 0 \quad a.s.. \tag{4.11}$$

A direct consequence of the above theorem is the validity of using the 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 4.2. *Under the null hypothesis, for any $\alpha \in (0,1)$, we have*

$$\Pr\{T_n > c_{n,1-\alpha}\} \to \alpha, \ \text{as } n \to \infty. \tag{4.12}$$

We analyze the power property of our test against the alternative hypothesis $H_1 : \lambda_0 \in \mathcal{D} \backslash \mathcal{C}$. 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.

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

$$\lim_{n\to\infty} \frac{\| \mathbf{Y} - \boldsymbol{\xi}_{\mathcal{S}_0} \|_{n,D}^2}{n} = c_0 \quad \text{and} \quad \lim_{n\to\infty} \frac{\| \boldsymbol{\xi}_{\mathcal{S}_0} - \boldsymbol{\xi}_{\mathcal{S}_1} \|_{n,D}^2}{n} = c_1, \tag{4.13}$$

for some positive constant terms c_0 and c_1, then

$$\Pr\{T_n > c_{n,1-\alpha}\} \to 1, \ \text{as } n \to \infty. \tag{4.14}$$

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 demonstrated by Example 3 from Kuchibhotla and Patra (2019) when the underlying function is (piecewise) 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 the constant in a non-trivial way. The latter condition is formalized by the second equation in (4.13), which is also needed in studying the power properties of related tests in Sen and Meyer (2017).

5. Extensions

To demonstrate the versatility of the methodology, we discuss extensions in two directions: a two-period panel selection model and the convex restriction on the control function. Both extensions can be implemented by the algorithm of Groeneboom and Hendrickx (2018) as Stage 1 and that of Meyer (2013) as Stage 2. Therefore, no additional complication occurs in terms of the computation. The precise regularity conditions and proofs are relegated to the supplementary notes for space restriction.

5.1. The Panel Selection Model

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

$$Y_{it}^{*} = X_{it}'\beta_0 + \alpha_i + \varepsilon_{it}; \tag{5.1}$$
$$D_{it} = \mathbb{I}\{W_{it}'\gamma_0 + \eta_i + \nu_{it} > 0\}.$$

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 $t = 1, 2$. There is no parametric assumption on any unobserved error term in our model. In order to use 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 8. For $t = 1, 2$, $RTI(\varepsilon_{it}|\nu_{it})$ holds for all $i = 1, 2, ..., n$. The individual heterogeneity η_i in the selection equation is independent of W_i and ν_{it}. The latent error ε_{it} is independent of $\nu_{it'}$ given ν_{it} for $t \neq t'$.

Thereafter, we have the following identity:

$$\mathbb{E}[Y_{i1} - Y_{i2}|D_{i1} = 1, D_{i2} = 1, W_i] = (X_{i1} - X_{i2})'\beta_0 + \lambda_{01}(W_{i1}'\gamma_0) - \lambda_{02}(W_{i2}'\gamma_0), \tag{5.2}$$

where

$$\lambda_{0t}(W_{it}'\gamma_0) = \int \mathbb{E}[\varepsilon_{it}|\nu_{it} > -W_{it}'\gamma_0 - \eta_i] dF_\eta(\eta_i) \quad \text{for} \quad t = 1, 2, \tag{5.3}$$

where $F_\eta(\cdot)$ stands for the distribution of η_i. By Condition 8 and Theorem 2.1, $\mathbb{E}[\varepsilon_{it}|\nu_{it} > -W_{it}'\gamma_0 - \eta_i]$ is decreasing in $W_{it}'\gamma_0$ 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), but 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. 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).

Our estimation procedure can be easily adapted to the panel data setting as follows. In the first stage, we once again apply the method from Groeneboom and Hendrickx (2018) to selection equations in both time periods separately to obtain $\hat{\gamma}_{nt}$ with $t = 1, 2$. Given

$\hat{\gamma}_{nt}$, we estimate β_0 and $\lambda_{0t}(\cdot)$ under the shape restriction for λ_t:

$$(\hat{\beta}_n, \hat{\lambda}_{n1}, \hat{\lambda}_{n2}) = \arg \min_{\beta \in \mathbf{B}, \lambda_1, \lambda_2 \in \mathcal{D}} \sum_{D_{i1}=D_{i2}=1} \left[\Delta Y_i - \Delta X_i'\beta - \lambda_2(W_{i2}'\hat{\gamma}_{n2}) + \lambda_1(W_{i1}'\hat{\gamma}_{n1})\right]^2, \tag{5.4}$$

where $\Delta Y_i \equiv (Y_{i2} - Y_{i1})$ and $\Delta X_i \equiv (X_{i2} - X_{i1})$ denote differenced random variables.

The following theorem explores the additive structure in (5.2) to establish the root-n consistency and asymptotic normality of our estimator for the outcome equation, in line with the work on additive isotonic regression (Cheng, 2009; Mammen and Yu, 2007; Han and Wellner, 2018). Although the assumption about latent errors in Kyriazidou (1997) is weaker, her estimator converges slower than the standard root-n rate.

Theorem 5.1. *Suppose Conditions P.1-P.7 in the supplementary notes hold, then we get*

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \Rightarrow \mathbb{N}(0, V_\beta^P), \tag{5.5}$$

where the asymptotic covariance matrix V_β^P is found in Section S3.

5.2. The Convexity Restriction

Convexity (or concavity) frequently arises as a consequence of economic models (Matzkin, 1991; Chetverikov, Santos, and Shaikh, 2018). The class of convex control functions is more regular (in terms of the model complexity or its entropy bound) than the monotone class, which leads to the improved convergence rate. A complete characterization of the convexity of control functions is beyond our scope and deserves a thorough investigation in a separate paper. Nonetheless, leading examples of this class include the original Heckman selection model with joint normal errors, the t-selection model in Marchenko and Genton (2012, see their Figure 2), and a quadratic extension of Olsen (1980) in the spirit of Brinch, Mogstad, and Wiswall (2017); see Example 1 in Brinch, Mogstad, and Wiswall (2017).

Referring to the semiparametric estimation, we only have to modify Stage 2, i.e., equation (3.3) by running the least squares estimator under the convexity restriction for λ:

$$(\tilde{\beta}_n, \tilde{\lambda}_n) = \arg \min_{\beta, \lambda} \sum_{D_i=1} \left[Y_i - X_i'\beta - \lambda(W_i'\hat{\gamma}_n)\right]^2. \tag{5.6}$$

Although the algorithm for (5.6) exists in Meyer (2013), to the best of our knowledge, the asymptotic property for the convex restricted partial linear model has not been established even for the case without the generated regressors from Stage 1. Theorem 5.2 establishes the asymptotic result that takes into account the effect of Stage 1.

Theorem 5.2. *Suppose Conditions C.1-C.7 in the supplementary notes hold, then we get*

$$\sqrt{n}\left(\tilde{\beta}_n - \beta_0\right) \Rightarrow \mathbb{N}(0, V_\beta), \tag{5.7}$$

where the asymptotic variance V_β is the same as in Theorem 4.2.

Typically, the shape restriction on the nonparametric component, such as monotonicity or convexity, does not change the influence function or the efficiency bound for the finite dimensional parameter in the model (Tripathi, 2000). However, the characterization of the least squares estimator under the convex restriction is more complicated (Groeneboom, Jongbloed, and Wellner, 2001) and requires considerable effort in the proof. See Section S3.2 of the supplementary notes for details.

6. Monte Carlo Simulations

In this section, Monte Carlo simulations are conducted to evaluate the finite sample performances of the proposed estimator assuming a monotone control function. We refer to it as the "MCF" estimator. Two alternative procedures are considered for comparison: the Heckman's two-step estimator (Heckit) and a kernel-based estimator that treats the control function as completely unknown and does not impose any monotonicity restriction. We consider the kernel-based estimator that combines the estimator of Klein and Spady (1993) for the selection equation and Robinson (1988) for the outcome equation, as empirical researchers often do; see Schafgans (1998, 2000).

We consider the following simulation design:

$$Y_i^* = \beta X_i + \varepsilon_i,\ D_i = \mathbb{I}\{-1 + \tilde{W}_{1i} + \gamma_2 \tilde{W}_{2i} + \gamma_3 X_i + \nu_i > 0\},\ Y_i = Y_i^* D_i, \tag{6.1}$$

where $\beta = 1$, $\gamma_2 = -2$, and $\gamma_3 = 0.25$. Let X_i follow the standard normal distribution, $\tilde{W}_{1i}$ follow the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$, and $\tilde{W}_{2i}$ follow the exponential distribution with a unit variance. Three different joint distributions of (ε, ν) are considered: DGP I sets a bivariate normal distribution with standard normal margins and the correlation coefficient $\rho = 0.9$. DGP II and III specify (ε, ν) in the form of a normal mixture according to Example 2.4:

$$\begin{bmatrix} \varepsilon \\ \nu \end{bmatrix} \sim \pi N\left(\begin{bmatrix} \mu_1 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right) + (1-\pi) N\left(\begin{bmatrix} \mu_2 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \times 15^2 \right),$$

where $\pi = 0.9$, $\sigma = 0.25$, and $\rho = 0.9$. DGP II sets $\mu_1 = \mu_2 = 0$ while DGP III sets $\mu_1 = 0.1$ and $\mu_2 = -0.9$, which generates skewed distributions of ε and ν. Simulation results are based on $1,000$ replications with sample size $n = 1,000$, $2,000$, and $5,000$.

For the kernel-based estimator, we examine its performance using various bandwidth selectors. In the first stage, we implement Silverman's rule of thumb[12] and the cross-validation bandwidth selector (Härdle, Hall, and Ichimura, 1993). For the second stage, in addition to the rule of thumb and cross-validation bandwidth selectors (applied to local linear regressions to obtain the estimates of $\mathbb{E}[Y|D = 1, W'\gamma_0]$ and $\mathbb{E}[X|D = 1, W'\gamma_0]$ prior to the OLS step in Robinson (1988)), we also include Ruppert, Sheather, and Wand (1995)'s plug-in estimate of the MSE-optimal bandwidth (for local linear regressions) and Linton (1995)'s optimal bandwidth based on the second order approximation to the partial linear model. In terms of the computation time, the MCF estimator is considerably faster than the kernel estimator with the cross-validation bandwidth selector. For example, one replication in DGP II with sample size $n = 2,000$ takes about 6 seconds on a 2.10-GHz Intel Xeon-E5 processor with 32 GB of RAM when using the former. On the other hand, it takes 50 seconds when using latter. The respective computation time becomes about 35 and 320 seconds when the sample size increases to $5,000$. We refer interested readers to Section S.4 in the supplementary notes for more explanations regarding the computation gap.

Table 1 reports the bias, standard error (SE), and root mean square error (RMSE) of estimators for the second stage coefficient β, which is the main focus of the sample selection model.[13] The following observations are made. First, both MCF and kernel estimators perform reasonably well in all three DGPs. Heckman's two-step estimator dominates other methods when the joint normality assumption holds, as in DGP I. However, it yields substantially larger bias and MSE than the other estimators in DGPs II and III, once the error distributions deviate from the joint normality. Second, in DGP I, the finite sample SE and RMSE of the MCF estimator are only marginally larger than Heckman's two-step estimator, suggesting that the efficiency loss of MCF is fairly small when Heckit is correctly specified. Third, the MCF estimator yields smaller SE and RMSE than all kernel estimators in DGPs II and III. In DGP I, the MCF estimator yields similar SE and RMSE as most kernel estimators, except for the one using rule-of-thumb bandwidths. However, the performance of the MCF estimator dominates the rule-of-thumb in all aspects when it comes to DGPs II and III. Fourth, in terms of bias, the MCF estimator is comparable to the kernel estimator with the plug-in bandwidth in DGPs II and III. Its bias is also similar to the one using the cross-validation bandwidth when $n = 1,000$, whereas the bias of the latter greatly decreases for larger sample sizes. In DGP I, on the other hand, kernel estimators generally

[12]For (γ_2, γ_3) taking the value (c_2, c_3), the bandwidth is $2.78\hat{\sigma}_1 n^{-1/5}$, where $\hat{\sigma}_1$ is the sample standard deviation of $\{\tilde{W}_{1i} + c_2\tilde{W}_{1i} + c_3 X_i\}_{i=1}^n$.

[13]The supplementary notes (Section S4) discuss the first stage estimators for γ_2 and γ_3 and confirms the robust performance of the NPMLE-based estimator (Groeneboom and Hendrickx, 2018) used in the first stage of the MCF.

exhibit smaller bias than the MCF estimator. Finally, there is no clear-cut winner among different bandwidth selectors. For instance, the rule-of-thumb bandwidth outperforms the others in DGP I, but its performance deteriorates significantly when it comes to DGPs II and III. In particular, for DGPs II and III, rule-of-thumb bandwidth yields the largest MSE among all semiparametric estimators when sample sizes $n = 2,000$ and $5,000$. In addition, although the cross-validation bandwidth selector is frequently adopted in empirical studies (Schafgans, 1998, 2000), Table 1 finds it being outperformed by the plug-in bandwidth in terms of SE and RMSE on several occasions, such as in DGPs II and III when $n = 1,000$ and $2,000$. This is consistent with the evidence from Ruppert, Sheather, and Wand (1995) for estimating the conditional means. It is worthwhile highlighting that Linton (1995)'s second-order optimal bandwidth constantly produces smaller bias than all other estimators including our MCF, although its overall RMSE is on a par with other methods. One possible explanation is that the original focus of Linton (1995) is the partial linear model without generated regressors from the first stage estimation.

In sum, the simulation results demonstrate encouraging performances of the MCF estimator, which incorporates the monotonicity restriction into the estimation procedure. This echoes one of the main drives of embedding shape restrictions in estimation with improved finite sample behaviors (Chetverikov, Santos, and Shaikh, 2018). In the same vein, it is also beneficial to incorporate shape restrictions in addition to the smoothness conditions for kernel or sieve estimators; see Case 2.2 of Chen and Shen (1998) and Chapter 8 in Groeneboom and Jongbloed (2014). For instance, Coppejans (2007) found notable improvement in simulation performance when the monotonicity restriction was incorporated into the B-spline estimator for other semiparametric models. In comparison, the relative performance of the kernel-based estimator depends on the bandwidth selector and also varies with underlying DGPs and sample sizes. In theory, the root-n consistency and asymptotic normality for the finite dimensional parameters hold for a wide range of tuning parameters (Robinson, 1988; Newey, 2009; Li and Wooldridge, 2002); however, the set of tuning parameters obviously matters regarding the finite sample performance. There is no unanimously dominating choice when it comes to the optimal version. It will not be surprising if the ranking among different choices changes under some other DGPs. Free from this type of bandwidth selection, the reliable performance of the MCF estimator shows its potential as a viable tool for applied researchers.

7. An Empirical Application: US Female Wage Equation

This section applies our method to estimate the female wage equation, using the Merged Outgoing Rotation Groups (MORG) of the CPS for the year 2013. We focus on white

TABLE 1. Finite sample performances of the MCF, Heckman's two-step (Heckit), and kernel-based semiparametric estimators for the second stage coefficient β_0 (true value = 1). DGP I: jointly normal errors; DGP II & DGP III: non-normal errors.

	Methods (bandwidths)	Heckit	MCF	Kernel-based (Klein-Spady + Robinson)			
				(CV, CV)	(CV, PI)	$(CV, 2ord)$	(ROT, ROT)
DGP I							
$n = 1000$	Bias	.0024	.0076	-.0019	-.0026	.0005	.0007
	SE	.0487	.0502	.0508	.0512	.0507	.0495
	RMSE	.0487	.0507	.0509	.0513	.0507	.0495
$n = 2000$	Bias	-.0010	.0054	-.0019	-.0020	-.0006	.0007
	SE	.0337	.0346	.0352	.0353	.0348	.0342
	RMSE	.0337	.0350	.0353	.0354	.0348	.0342
$n = 5000$	Bias	.0008	.0037	-.0008	-.0008	-.0004	.0004
	SE	.0220	.0225	.0228	.0228	.0227	.0223
	RMSE	.0220	.0228	.0228	.0228	.0227	.0223
DGP II							
$n = 1000$	Bias	.0660	.0152	-.0146	-.0196	.0110	.0397
	SE	.1773	.1523	.1700	.1634	.1638	.1649
	RMSE	.1892	.1531	.1706	.1646	.1642	.1696
$n = 2000$	Bias	.0765	.0162	-.0040	-.0089	.0125	.0374
	SE	.1251	.1061	.1131	.1117	.1125	.1156
	RMSE	.1467	.1073	.1131	.1120	.1132	.1215
$n = 5000$	Bias	.0681	.0091	-.0040	-.0092	.0042	.0197
	SE	.0735	.0626	.0656	.0663	.0651	.0646
	RMSE	.1002	.0633	.0658	.0669	.0652	.0675
DGP III							
$n = 1000$	Bias	.0649	.0153	-.0187	-.0238	.0092	.0360
	SE	.1724	.1468	.1628	.1546	.1574	.1603
	RMSE	.1843	.1476	.1639	.1565	.1577	.1643
$n = 2000$	Bias	.0763	.0163	-.0038	-.0097	.0108	.0334
	SE	.1216	.1009	.1119	.1100	.1097	.1108
	RMSE	.1436	.1022	.1119	.1105	.1103	.1157
$n = 5000$	Bias	.0676	.0086	-.0046	-.0100	.0036	.0173
	SE	.0722	.0607	.0639	.0639	.0638	.0634
	RMSE	.0989	.0613	.0641	.0646	.0639	.0657

Note: CV: cross-validation bandwidth; PI: Ruppert, Sheather, and Wand (1995)'s plug-in estimate of the MSE-optimal bandwidth; $2ord$: Linton (1995)'s optimal bandwidth based on the second order approximation; ROT: rule-of-thumb bandwidth.

married women in the southern region of the United States who are between 25 and 54 years old and have at least a high school education. The dependent variable Y_i represents the i-th woman's hourly wage (in logarithms) and D_i indicates whether she works at least 35 hours a week (full-time worker). Exogenous variables, W_i, entering the selection equation

are: potential experience (divided by 10) and its square, four education level dummies for some college, associate degree, bachelor degree and advanced degree, and number of children in three age ranges (0 to 2, 3 to 5 and 6-13). The construction of variables follows from Huber and Melly (2015). The last three variables concerning number of children are excluded from the wage offer equation. The sample size is 8,327, among which 2,285 are full-time workers.[14]

Table 2 presents the estimated coefficients in the wage equation and the p-values for the selection bias tests using three methods: Heckit, MCF, and the kernel-based semiparametric estimator. For the kernel estimator, we consider bandwidths chosen by cross-validation (Kernel_{CV}) and plug-in (Kernel_{PI}) methods, respectively. In the current setup, it is reasonable to assume that the motivation to work positively contributes to the earnings. Then by the example in Remark 2.1, the errors in the selection and outcome equations are likely to satisfy RTI, which motivates the application of the MCF estimator with the decreasing control function. This assumption is also supported by inspecting Figure 2, which plots the estimates of the control function given by three approaches (the kernel estimator in Figure 2 refers to Kernel_{CV}). It shows that the kernel estimate of the control function exhibits a decreasing shape and lies very closely to the MCF estimate. In addition, the comparison of SSRs associated with the increasing and decreasing control functions (Corollary 4.1) also favors the decreasing direction. In particular, the SSR associated with an increasing control function is 680.0, whereas its decreasing counterpart is 674.6. In Table 2, the point estimates of the MCF approach are similar to the kernel-based estimators, and both differ from the Heckit to some extent. This suggests that the joint normality may be too restrictive for the dataset. To check the validity of Heckit, we conduct the LM-type test for the normality assumption in the Probit model (Bera, Jarque, and Lee, 1984) of the selection equation. It strongly rejects the marginal normality of ν (with the p-value ≈ 0), and hence refutes the Heckit specification. When comparing MCF with the two kernel estimates, on one hand, we find that in terms of the point estimation, MCF is closer to Kernel_{CV} (which exhibits smaller bias in large sample in our Monte Carlo) than to Kernel_{PI}, especially for the coefficients on Exp and Exp^2. On the other hand, where the length of the bootstrap confidence interval is concerned, MCF is similar to Kernel_{PI} and both are shorter than Kernel_{CV}. This agrees with our simulation findings with the relatively small SEs of MCF and Kernel_{PI}. In particular, the confidence interval of MCF is the shortest among the three semiparametric methods. The coefficients on Exp, Exp^2, and $Associate$ are significant at 5% level in the MCF and Kernel_{PI} estimates, which implies an increasingly concave experience-earning relationship and the significantly positive effects

[14]For our sample and variable choices, the independence assumption between (ε_i, ν_i) and W_i in model (1.1) is not rejected by the quantile-based test of Huber and Melly (2015).

TABLE 2. Wage equation for married women. Total number of observed = 8,327; number of working observed =2,285. 95% confidence intervals are in the brackets.

	Heckit	MCF	Kernel$_{CV}$	Kernel$_{PI}$
Exper	-.107	.281	.300	.434
	[-.392, .178]	[.089, .408]	[-.023, .776]	[.142, .555]
$Exper^2$	.026	-.051	-.059	-.085
	[-.041, .093]	[-.079, -.010]	[-.168 , .030]	[-.110, -.023]
Some college	.275	.214	.247	.212
	[.122, .428]	[.161, .284]	[.027 .444]	[.149, 288]
Associate	-.011	.153	.151	.164
	[-.191, .168]	[.083, .222]	[-.032, .473]	[.084, .239]
Bachelor	.271	.295	.299	.290
	[.109, .434]	[.229, .366]	[.054, .530]	[.202, .374]
Advanced	.394	.284	.300	.265
	[.230, .559]	[.179, .375]	[.054, .617]	[.086, .395]
Selection bias test (p-value)	.002	.023	.041	.021

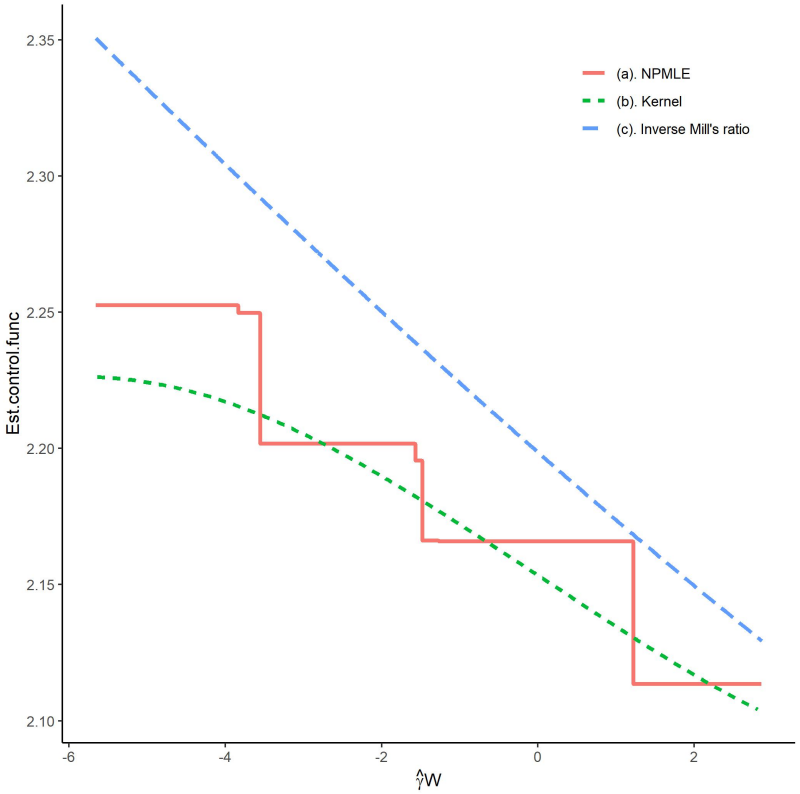
Note: The MCF estimation and testing assume a decreasing control function. Bandwidths for the kernel estimators are chosen by cross-validation in Kernel$_{CV}$, and by plug-in method (Ruppert, Sheather, and Wand, 1995) in Kernel$_{PI}$. The 95% confidence intervals in the last two columns and the p-value for the MCF are calculated from $1,000$ bootstrap replications.

of all four education level dummies. To test for the presence of labor market selection, we conduct the t-test based on Heckman's selection model, our selectivity test in Section 3.2, and a kernel-based test in the spirit of Christofides, Li, Liu, and Min (2003). As the last row of Table 2 shows, all methods unanimously reject the null hypothesis of no sample selection. Overall, this example illustrates that the bandwidth choice matters in the estimation and inference using the kernel-based method. In the same spirit of the literature on shape restriction (Chetverikov, Santos, and Shaikh, 2018), our method yields a fitted model that is more amenable to the interpretation of a positive selection bias in women's labor market participation.

8. Conclusion

This paper proposes a semiparametric sample selection model with a monotonicity constraint on the selection correction function. Non-random 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 32,200 Google Scholar

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



citations at the time of writing. Lying between the original Heckman selection model and the semiparametric 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 proposal shares the generality of semiparametric approaches while keeping the main convenience of parametric methods as its implementation is free from any tuning parameter selection.

9. Appendix: Proofs of Main Results

For two sequences a_n, b_n, we write $a_n \lesssim b_n$ if $a_n \leq b_n$ for some positive finite M independent of n. We denote $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold.

Proof of Theorem 2.1. Let the conditional distribution $F_{\varepsilon|\nu>t}(s)$ and conditional survivor function $\bar{F}_{\varepsilon|\nu>t}(s)$ denote $\Pr\{\varepsilon \leq s|\nu > t\}$ and $\Pr\{\varepsilon > s|\nu > t\}$, respectively. The definition of RTI directly states that $\bar{F}_{\varepsilon|\nu>t}(s)$ is an increasing function of t and $F_{\varepsilon|\nu>t}(s)$ is a decreasing function for any s. Therefore, $\int_0^{+\infty} \bar{F}_{\varepsilon|\nu>t}(s)ds$ is increasing and $\int_{-\infty}^{0} F_{\varepsilon|\nu>t}(s)ds$

is decreasing with respect to t. The following formula is the key for our purposes:

$$\mathbb{E}[\varepsilon|\nu > t] = \int_{-\infty}^{+\infty} s dF_{\varepsilon|\nu>t}(s) = \int_{0}^{+\infty} \bar{F}_{\varepsilon|\nu>t}(s)ds - \int_{-\infty}^{0} F_{\varepsilon|\nu>t}(s)ds. \tag{9.1}$$

See Shorack (2000, Chapter 6, Proposition 4.2) for the second equality. Combining the results attached to those two terms on the right hand side of equation (9.1), it is evident that the control function $\lambda(t) = \mathbb{E}[\varepsilon|\nu > -t, W]$ is decreasing with respect to t. □

Proof of Theorem 4.1. We use the short-hand notations $\theta_0 = (\beta_0, \lambda_0(\cdot))$ and $\hat{\theta}_n = (\hat{\beta}_n, \hat{\lambda}_n(\cdot))$. Regarding the outcome equation, we denote the function $f_{\theta,\gamma} = x'\beta + \lambda(w'\gamma)$, with $\beta \in \mathbf{B}$ and $\lambda \in \mathcal{D}$ and we abbreviate f_{θ_0,γ_0} as $f_0 = x'\beta_0 + \lambda_0(w'\gamma_0)$.

By definition of $(\hat{\beta}_n, \hat{\lambda}_n)$, we get $\mathbb{P}_{1n}[Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n)]^2 \leq \mathbb{P}_{1n}[Y - X'\beta_0 - \lambda_0(W'\hat{\gamma}_n)]^2$, which leads to

$$\begin{aligned} &P_1\left[(Y_i - X_i'\beta_0 - \lambda_0(W_i'\hat{\gamma}_n))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right] \\ \leq &(\mathbb{P}_{1n} - P_1)\left[(Y_i - X_i'\beta_0 - \lambda_0(W_i'\hat{\gamma}_n))^2 - \left(Y_i - X_i'\hat{\beta}_n - \hat{\lambda}_n(W_i'\hat{\gamma}_n)\right)^2\right]. \end{aligned} \tag{9.2}$$

Thereafter, we prove in Lemma S6 that the left-hand side (l.h.s.) of inequality (9.2) can be bounded below by (modulo some constant multiplier)

$$|\hat{\beta}_n - \beta_0|^2 + \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \|^2 - O_p(n^{-1}). \tag{9.3}$$

The right-hand side (r.h.s.) of inequality (9.2) is $o_p(1)$ by the Glivenko-Cantelli property of the corresponding empirical process. The combination of (9.2) and (9.3) concludes the consistency.

Considering the rate of convergence, we let $\mathbb{M}_n f_{\theta,\gamma} \equiv \frac{2}{n}\sum_{i=1}^n (f_{\theta,\gamma} - f_0)D_i\epsilon_i - \frac{1}{n}\sum_{i=1}^n (f_{\theta,\gamma} - f_0)^2 D_i$, and $\mathbb{M}f \equiv \mathbb{E}[\mathbb{M}_n f]$. Accordingly, we set $d(\theta, \theta_0; \gamma) \equiv -\mathbb{M}f_{\theta,\gamma} = \mathbb{E}[D(f_{\theta,\gamma} - f_0)^2]$ to be the pseudo-norm employed in the peeling argument. Also, we denote

$$\zeta_n = \mathbb{P}_{1n}\left[2\epsilon(\lambda_0(W'\gamma_0) - \lambda_0(W'\hat{\gamma}_n)) + (\lambda_0(W'\gamma_0) - \lambda_0(W'\hat{\gamma}_n))^2\right]. \tag{9.4}$$

It is clear that $\zeta_n = O_p(n^{-1})$ by the root-n consistency of $\hat{\gamma}_n$ and standard argument, given the first-order differentiability of λ_0.

For any $j \in \mathbb{N}$, let $\mathcal{F}_{n,j} \equiv \{f_{\theta,\gamma} : 2^{j-1}t\eta_n \leq d(\theta, \theta_0; \gamma) < 2^j t\eta_n\} \cap \mathcal{B}_n$, where $\mathcal{B}_n$ is a properly localized set that contains the estimates $(\hat{\beta}_n, \hat{\lambda}_n, \hat{\gamma}_n)$ with probability approaching 1; see equation (S.8) in the supplementary notes. Then by the peeling/slicing argument, we get

$$\Pr\left\{d(\hat{\theta}_n, \theta_0; \hat{\gamma}_n) \geq t\eta_n\right\} \leq \sum_{j\geq 1} \Pr\{\sup_{f\in\mathcal{F}_{n,j}} (\mathbb{M}_n(f) - M(f)) \geq 2^{2j-2}t^2\eta_n^2 - |\zeta_n|\} + o(1).$$

We rely on an appropriate truncation device that depends on the envelope of the empirical process. For this purpose, we denote the local envelope that varies on each slice by

$$\sup_{f\in\mathcal{B}_n:d(\theta,\theta_0;\gamma)\leq\eta}|2\epsilon(f-f_0)-(f-f_0)^2|\leq U(Z;\eta). \tag{9.5}$$

Given the negligibility of ζ_n, we can focus on the sum of each probability term into two parts:

$$P_{I,n}\equiv\sum_{j\geq1}\Pr\{\sup_{f\in\mathcal{F}_{n,j}}\left|(\mathbb{M}_n-\mathbb{M})(f\mathbb{I}\{U(Z;2^tj\eta_n)\leq B_j\})\right|\geq 2^{2j-4}t^2\eta_n^2\},$$
$$P_{II,n}\equiv\sum_{j\geq1}\Pr\{\sup_{f\in\mathcal{F}_{n,j}}\left|(\mathbb{M}_n-\mathbb{M})(f\mathbb{I}\{U(Z;2^tj\eta_n)\geq B_j\})\right|\geq 2^{2j-4}t^2\eta_n^2\},$$

in which the bounded part can be dealt with by the maximal inequality from Lemma 3.4.2 of Van Der Vaart and Wellner (1996):

$$P_{I,n}\lesssim\sum_{j=1}^{\infty}\frac{(\log n)^{3/2}}{n^{1/2}(2^jt\eta_n)^{3/2}}+\sum_{j=1}^{\infty}\frac{B_j\log n}{n(2^jt\eta_n)^3};$$

see Lemma S4. An elementary Markov inequality in Lemma S5 is applied to the unbounded remainder with a precise control on the local envelope function by the argument in Giné and Koltchinskii (2006) and Han and Wellner (2018):

$$P_{II,n}\lesssim\frac{n^{r/9}}{t^2\eta_n^2}\sum_{j=1}^{\infty}\frac{(2^jt\eta_n\log n)^r\log^{r/2}(1/2^jt\eta_n)}{2^{2j}B_j^{r-1}}.$$

By taking $\eta_n\asymp\log n\times n^{-1/3}$, $B_j\asymp 2^jt$, and $r>3$, one gets

$$\begin{aligned}&\Pr\left\{d(\hat\theta_n,\theta_0;\hat\gamma_n)\geq t\eta_n\right\}\\ \lesssim& t^{-3/2}\sum_{j=1}^{\infty}2^{-j/2}+t^{-2}(\log n)^{-2}\sum_{j=1}^{\infty}2^{-2j}+\frac{\log^{(5r/2-2)}(n)}{n^{2r/9-2/3}}\sum_{j=1}^{\infty}\log^{r/2}(2^{-j}t^{-1})2^{-j}t^{-1}.\end{aligned}$$

The stated rate of convergence follows by letting $t\to+\infty$ together with Lemma S6. □

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) or Groeneboom and Jongbloed (2014). For our purposes, 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 piecewise constant function that has the same jump locations with $\hat{\lambda}_n(\cdot)$. Therefore, we start with the following characterization condition for our estimator $(\hat{\beta}_n, \hat{\lambda}_n)$:

$$\mathbb{P}_n \left[D(Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)]) \right] = 0, \tag{9.6}$$

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

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

given the fact that $\mathbb{E}[\varepsilon|W, D=1] = 0$. Regarding the r.h.s. of Equation (9.7), we use the P-Donsker property in Lemma S7 to show that

$$\begin{aligned} &\mathbb{G}_n \left[D(Y - X'\hat{\beta}_n - \hat{\lambda}_n(W'\hat{\gamma}_n))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)]) \right] \\ &\mathbb{G}_n \left[D(Y - X'\beta_0 - \lambda_0(W'\gamma_0))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \lambda_0(W'\gamma_0)]) \right] + o_p(1) \\ &= \mathbb{G}_n \left[\epsilon(X - \mathbb{E}[X|D=1, W'\gamma_0]) \right] + o_p(1). \end{aligned} \tag{9.8}$$

Furthermore, we decompose the l.h.s. of Equation (9.7) into two terms, J_{1n} and J_{2n}, defined as follows:

$$J_{1n} = \sqrt{n}\mathbb{E}\left[D(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)])X' \right](\hat{\beta}_n - \beta_0), \tag{9.9}$$

$$J_{2n} = \sqrt{n}\left[D(\hat{\lambda}_n(W'\hat{\gamma}_n) - \lambda_0(W'\gamma_0))(X - \mathbb{E}[X|D=1, \lambda_0^{-1} \circ \hat{\lambda}_n(W'\hat{\gamma}_n)]) \right]. \tag{9.10}$$

In our Lemmas S8 and S9, we prove that

$$J_{1n} = \mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])X' \right] \sqrt{n}(\hat{\beta}_n - \beta_0) + o_p(1 + \sqrt{n}|\hat{\beta}_n - \beta_0|), \tag{9.11}$$

and

$$J_{2n} = \mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])\dot{\lambda}_0(W'\gamma_0)W'_{-1} \right] \sqrt{n}(\hat{\gamma}_{n-} - \gamma_{0-}) + o_p(1). \tag{9.12}$$

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

$$\begin{aligned} &\mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])X' \right] \sqrt{n}(\hat{\beta}_n - \beta_0) \\ =&\mathbb{G}_n \left[\epsilon(X - \mathbb{E}[X|D=1, W'\gamma_0]) \right] - \mathbb{E}\left[D(X - \mathbb{E}[X|D=1, W'\gamma_0])\dot{\lambda}_0(W'\gamma_0)W'_{-1} \right] \sqrt{n}(\hat{\gamma}_{n-} - \gamma_{0-}) \\ &+ o_p(1 + \sqrt{n}|\hat{\beta}_n - \beta_0|). \end{aligned} \tag{9.13}$$

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 (9.13) are uncorrelated, which gives rise to the particular form of the asymptotic covariance matrix in Theorem 4.2. □

The proof of Theorem 4.3 uses the p-Mallows' distance d_p for $p = 1, 2$ between two distribution functions F_1 and F_2, which is defined by:

$$d_p(F_1, F_2) \equiv \inf_J \left\{ [\mathbb{E}_{(S,T)\sim J} |S - T|^p]^{1/p} : \quad J \quad \text{has marginal distributions} \quad (F_1, F_2) \right\}.$$

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 follows the route in Mammen (1993) for bootstrapping the F-type statistics. In particular, we show the numerators of test statistics in (3.6) and (3.7) converging by the $d_1(\cdot, \cdot)$ (1-Mallows' distance), which implies the convergence in the Lévy metric. Also, their denominators converge in probability to the same limit.

Let $\hat{G}_n$ be a sequence of random distribution functions of the bootstrap residuals ϵ^*, where ϵ^* is obtained by re-sampling the centered residual $\hat{\epsilon}$. Under the null hypothesis, the centered error term ϵ has zero conditional mean given X and $D = 1$. By Condition 2 with $r \geq 4$, we have

$$\| \boldsymbol{\epsilon} - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) \|^2_{n,D} / n_1 = \mathbb{E}[\epsilon^2 | D = 1] + O_p(n_1^{-1/2}), \tag{9.14}$$

$$\| \boldsymbol{\epsilon}^* - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) \|^2_{n,D} / n_1 = \mathbb{E}[\epsilon^2 | D = 1] + O_p(n_1^{-1/2}), \tag{9.15}$$

where (9.15) is obtained by the fact $\mathbb{E}^*[\boldsymbol{\epsilon}^*|D = 1] = \sum_{i=1}^{n_1} \hat{\epsilon}^2 / n_1$, with $\mathbb{E}^*$ being the conditional expectation given the data. 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. To emphasize the dependence on the residual terms, we write the numerators of statistics (3.6) and (3.7), our test statistics, as $W_n(\epsilon)$ and $W_n(\epsilon^*)$, so that we have

$$W_n(\epsilon) = \| \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_{1,\hat{\gamma}_n}) \|^2_{n,D} / n_1 \text{ and } W_n(\epsilon^*) = \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|^2_{n,D} / n_1.$$

Thereafter, one can bound $n_1^{1/2} W_n(\epsilon)^{1/2}$ by

$$\begin{aligned} &\| \Pi(\boldsymbol{\epsilon}|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) \|_{n,D} + \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_0) - \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} + \| \Pi(\boldsymbol{\epsilon}^*|\mathcal{S}_{1,\hat{\gamma}_n}) - \Pi(\boldsymbol{\epsilon}|\mathcal{S}_{1,\hat{\gamma}_n}) \|_{n,D} \\ &\leq 2 \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_{n,D} + n_1^{1/2} W_n(\epsilon^*)^{1/2}. \end{aligned}$$

The analogous argument yields $n_1^{1/2} W_n(\epsilon^*)^{1/2} \leq 2 \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_{n,D} + n_1^{1/2} W_n(\epsilon)^{1/2}$. Therefore,

$$\left| W_n^{1/2}(\epsilon) - W_n^{1/2}(\epsilon^*) \right| \leq 2 \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^* \|_{n,D} / n_1^{1/2}. \tag{9.17}$$

Letting $\mathcal{L}(\cdot)$ denote the distribution, we have

$$\begin{aligned} d_1(\mathcal{L}(W_n), \mathcal{L}(W_n^*)) &\leq \mathbb{E}|W_n(\epsilon) - W_n(\epsilon^*)| \\ &\leq \sqrt{\mathbb{E}\left|W_n^{1/2}(\epsilon) - W_n^{1/2}(\epsilon^*)\right|^2 \mathbb{E}\left|W_n^{1/2}(\epsilon) + W_n^{1/2}(\epsilon^*)\right|^2} \\ &\leq C\sqrt{\mathbb{E} \parallel \boldsymbol{\epsilon} - \boldsymbol{\epsilon^*} \parallel_{n,D}^2 / n_1} = C d_2^{1/2}(\hat{G}_n, G) \to 0, \quad a.s., \end{aligned} \tag{9.18}$$

where the first inequality follows from the definition of the 1-Mallows' distance (see page 64 of Shorack and Wellner (2009)), the third inequality from (9.17) and Condition 2, the equality from Shorack and Wellner (2009, equation (2) on page 63 and Theorem 2 on page 65 therein), and the last convergence from Freedman (1981, Lemma 2.6). Note that (9.18) implies $d_L(\mathcal{L}(W_n), \mathcal{L}(W_n^*)) \to 0$, which in turn leads to the desired result when combined with (9.14) and (9.15). □

Proof of Theorem 4.4. Under the null hypothesis that the control function is a constant term, one can combine the proof of our Theorem (4.1) and Example 3 from Kuchibhotla and Patra (2019) to get $T_n = O_p(\log n/n)$, which leads to $c_{n,\alpha} = o(1)$. 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, one has

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

by the Glivenko-Cantelli property of the corresponding functional, which leads to $T_n \to_p c_1/c_0$, as $n \to +\infty$, combining with the two conditions stated in Theorem 4.4. □

References

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.

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

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 (2019): "Least squares estimation in the monotone single index model," *Bernoulli*, 25, 3276–3310.

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.

Bera, A. K., C. M. Jarque, and L.-F. Lee (1984): "Testing the normality assumption in limited dependent variable models," *International Economic Review*, 25, 563–578.

Berman, N., V. Rebeyrol, and V. Vicard (2019): "Demand learning and firm dynamics: evidence from exporters," *Review of Economics and Statistics*, 101, 91–106.

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, 985–1039.

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

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

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

Chen, X., and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, 66, 289–314.

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

Chernozhukov, V., I. Fernandez-Val, and S. Luo (2018): "Distribution regression with sample selection, with an application to wage decompositions in the UK," *arXiv preprint*, arXiv:1811.11603.

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

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.

Coppejans, M. (2007): "On efficient estimation of the ordered response model," *Journal of Econometrics*, 137, 577–614.

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.

Escanciano, J. C., and L. Zhu (2015): "A simple data-driven estimator for the semiparametric sample selection model," *Econometric Reviews*, 34, 734–762.

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 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.

Francesconi, M., and C. Nicoletti (2006): "Intergenerational mobility and sample selection in short panels," *Journal of Applied Econometrics*, 21, 1265–1293.

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.

Giné, E., and V. Koltchinskii (2006): "Concentration inequalities and asymptotic results for ratio type empirical processes," *The Annals of Probability*, 34, 1143–1216.

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., G. Jongbloed, and J. Wellner (2001): "Estimation of a convex function: characterizations and asymptotic theory," *The Annals of Statistics*, 29, 1653–1698.

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

Han, Q., and J. Wellner (2018): "Robustness of shape-restricted regression estimators: An envelope perspective," *working paper*.

——— (2019): "Convergence rates of least squares regression estimators with heavy-tailed errors," *The Annals of Statistics*, 47, 2286–2319.

Härdle, Wolfgang, P. Hall, and J. S. Marron (1988): "How far are automatically chosen regression smoothing parameters from their optimum?," *Journal of the American*

Statistical Association, 83, 86–95.

Härdle, W., P. Hall, and H. Ichimura (1993): "Optimal smoothing in single-index models," *The Annals of Statistics*, 21, 157–178.

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 R. Robb (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of Econometrics*, 30, 239–267.

Heckman, J. J., and E. Vytlacil (2007a): "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," *Handbook of Econometrics*, 6B, 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*, 6B, 4875–5143.

Honoré, B. E., and L. Hu (2020): "Selection without exclusion," *Econometrica*, 88, 1007–1029.

Horowitz, J. L., and S. Lee (2017): "Nonparametric estimation and inference under shape restrictions," *Journal of Econometrics*, 201, 108–126.

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

Huber, M., and B. Melly (2015): "A test of the conditional independence assumption in sample selection models," *Journal of Applied Econometrics*, 30, 1144–1168.

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58, 71–120.

Ichimura, H., and P. E. Todd (2007): "Implementing nonparametric and semiparametric estimators," *Handbook of Econometrics*, 6B, 5369–5468.

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

Kuchibhotla, A. K., and R. K. Patra (2019): "On Least Squares Estimation under Heteroscedastic and Heavy-Tailed Errors," *arXiv preprint*, arXiv:1909.02088.

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

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

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.

Linton, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica*, 63, 1079–1112.

Maasoumi, E., and L. Wang (2019): "The gender earnings gap: Measurement and analysis," *Journal of Political Economy*, forthcoming.

Mammen, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *The Annals of Statistics*, 21, 255–285.

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.

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

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

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

Newey, W. (1999): "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 63, 129–132.

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

Olsen, R. (1980): "A least squares correction for selectivity bias," *Econometrica*, 48, 1815–1820.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press.

Powell, J. L. (2001): "Semiparametric estimation of censored selection models," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, vol. 13, pp. 165–196. Cambridge University Press.

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.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995): "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, 90, 1257–1270.

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.

Shorack, G. R., and J. A. Wellner (2009): *Empirical Processes with Applications to Statistics*. SIAM.

Smith, M. D. (2003): "Modelling sample selection using Archimedean copulas," *The Econometrics Journal*, 6, 99–123.

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

Tripathi, G. (2000): "Local semiparametric efficiency bounds under shape restrictions," *Econometric Theory*, 16, 729–739.

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.

Varadhan, R., and P. Gilbert (2009): "BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function," *Journal of Statistical Software*, 32, 1–26.

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

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.

Zhou, X., and Y. Xie (2019): "Marginal treatment effects from a propensity score perspective," *Journal of Political Economy*, 127, 3070–3084.