Received: 16 August 2018 | Revised: 4 March 2019

DOI: 10.1002/jae.2727

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Estimation in a generalization of bivariate probit models with dummy endogenous regressors

Sukjin Han[1] | **Sungwon Lee**[2]

[1]Department of Economics, University of Texas at Austin, Austin, Texas, United States

[2]Global Asia Institute, National University of Singapore, Singapore

Correspondence
Sukjin Han, Sukjin Han, Department of Economics, University of Texas at Austin, Austin, TX 78712.
Email: sukjin.han@austin.utexas.edu

Summary

The purpose of this paper is to provide guidelines for empirical researchers who use a class of bivariate threshold crossing models with dummy endogenous variables. A common practice employed by the researchers is the specification of the joint distribution of unobservables as a bivariate normal distribution, which results in a *bivariate probit model*. To address the problem of misspecification in this practice, we propose an easy-to-implement semiparametric estimation framework with parametric copula and nonparametric marginal distributions. We establish asymptotic theory, including root-n normality, for the sieve maximum likelihood estimators that can be used to conduct inference on the individual structural parameters and the average treatment effect (ATE). In order to show the practical relevance of the proposed framework, we conduct a sensitivity analysis via extensive Monte Carlo simulation exercises. The results suggest that estimates of the parameters, especially the ATE, are sensitive to parametric specification, while semiparametric estimation exhibits robustness to underlying data-generating processes. We then provide an empirical illustration where we estimate the effect of health insurance on doctor visits. In this paper, we also show that the absence of excluded instruments may result in identification failure, in contrast to what some practitioners believe.

1 | INTRODUCTION

The purpose of this paper is to provide guidelines for empirical researchers who use a class of bivariate threshold crossing models with dummy endogenous variables. This class of models is typically written as follows. With the binary outcome Y and the observed binary endogenous treatment D, we consider

$$\begin{aligned} Y &= \mathbf{1}[X'\beta + \delta_1 D - \varepsilon \geq 0], \\ D &= \mathbf{1}[X'\alpha + Z'\gamma - \nu \geq 0], \end{aligned} \tag{1}$$

where X denotes a vector of exogenous regressors that determine both Y and D, and Z denotes a vector of exogenous regressors that directly affect D, but not Y (i.e., instruments for D). Since Y does not appear in the equation for D, this model forms a triangular model, as a special case of a simultaneous equations model, with the binary endogenous variables. In this paper, we investigate the consequences of the common practices employed by empirical researchers who use this class of models. As an important part of this investigation, we conduct a sensitivity analysis on the specification of the joint distribution of the unobservables (ε, ν). This is the component of the model that practitioners have the least knowledge about, and thus typically impose a parametric assumption. To address the problem of misspecification,

we propose a semiparametric estimation framework with parametric copula and nonparametric marginal distributions. The semiparametric specification is an attempt to ensure robustness while achieving point identification and efficient estimation.

The parametric class of models (Equation (1)) includes the *bivariate probit model,* in which the joint distribution of (ε, ν) is assumed to be a bivariate normal distribution. This model has been widely used in empirical research, including the works of Evans and Schwab (1995), Neal (1997), Goldman et al. (2001), Altonji, Elder, and Taber (2005), Bhattacharya, Goldman, and McCaffrey (2006), Rhine, Greene, and Toussaint-Comeau (2006), and Marra and Radice (2011), to name just a few. The distributional assumption in this model, however, is made out of convenience or convention, and is hardly justified by underlying economic theory and thus susceptible to misspecification. With binary endogenous regressors, the objects of interest. In model (1) are the mean treatment parameters, in addition to the individual structural parameters. Because the outcome variable is also binary, the mean treatment parameters such as the average treatment effect (ATE) are expressed as the differential between the marginal distributions of ε. Therefore, the problem of misspecification when estimating these treatment parameters can be even more severe than that when estimating individual parameters.

At one extreme, a nonparametric joint distribution of (ε, ν) can be used in a bivariate threshold crossing model, as in Shaikh and Vytlacil (2011). Their results, however, suggest that the ATE is only partially identified in this fully flexible setting. Instead of sacrificing point identification, we impose a parametric assumption on the dependence structure between the unobservables using copula functions that are known up to a scalar parameter. At the same time, in order to ensure robustness, we allow the marginal distribution of ε (and ν), which is involved in the calculation of the ATE, to be unspecified. Our class of models encompasses both parametric and semiparametric models with parametric copula and either parametric or nonparametric marginal distributions. This broad range of models allows us to conduct a sensitivity analysis on the specification of the joint distribution of (ε, ν).

Identification of the individual parameters and the ATE in this class of models is established in Han and Vytlacil (2017; hereafter, HV17). They show that when the copula function for (ε, ν) satisfies a certain stochastic ordering, identification is achieved in both parametric and semiparametric models under an exclusion restriction and mild support conditions. Building on these results, we consider estimation and inference in the same setting. For the semiparametric class of models (Equation (1)) with parametric copula and nonparametric marginal distributions, the likelihood contains infinite-dimensional parameters (i.e., the unknown marginal distributions). To estimate this model, we consider the sieve maximum likelihood (ML) estimation method for the finite- and infinite-dimensional parameters of the model, as well as their functionals. Estimation of the parametric model, on the other hand, is within the standard ML framework.

The contributions of this paper can be summarized as follows. Through these contributions, this paper is intended to provide a guideline to empirical researchers. First, we establish the asymptotic theory for the sieve ML estimators in a class of semiparametric copula-based models. This result can be used to conduct inference on the functionals of the finite- and infinite-dimensional parameters, such as inference on the individual structural parameters and the ATE. We show that the sieve ML estimators are consistent and that their smooth functionals are root-n asymptotically normal.

Second, in order to show the practical relevance of the theoretical results for empirical researchers, we conduct a sensitivity analysis via extensive Monte Carlo simulation exercises. We find that the parametric ML estimates, especially those for the ATE, can be highly sensitive to the misspecification of the marginal distributions of the unobservables. On the other hand, the sieve ML estimates perform well in terms of the mean squared error (MSE) as they are robust to the underlying data-generating process (DGP). Moreover, their performance is comparable to that of the parametric estimates under a correct specification. We also show that copula misspecification does not have a substantial effect in estimation, as long as the true copula is within the stochastic ordering class of the identification. As copula misspecification is a problem common to both parametric and semiparametric models considered in this paper, our sensitivity analysis suggests that a semiparametric consideration may be more preferable in estimation and inference.

Third, we provide an empirical illustration of the sieve estimation and the sensitivity analysis of this paper. We estimate the effect of health insurance on decisions to visit doctors using the the Medical Expenditure Panel Survey data combined with the National Compensation Survey data by matching industry types. We compare the estimates of parametric and semiparametric bivariate threshold crossing models with the Gaussian copula. We show that the estimates differ, especially so for the estimated ATEs, which suggest the misspecification of the marginal distribution of the unobservables, consistent with the simulation results. In other words, the estimates of the bivariate probit model can be misleading in this example.

Fourth, we formally show that identification may fail without the exclusion restriction, in contrast to the findings of Wilde (2000). The bivariate probit model is sometimes used in applied work without instruments (Rhine et al., 2006;

White & Wolaver, 2003). We show, however, that this restriction is not only sufficient but also necessary for identification in parametric and semiparametric models when there is a single binary exogenous variable common to both equations. We also show that under joint normality of the unobservables the parameters are, at best, weakly identified when there are common (and possibly continuous) exogenous variables. [1] We also note that another source of identification failure is the absence of restrictions on the dependence structure of the unobservables, as mentioned above.

The sieve estimation method is a useful nonparametric estimation framework that allows for a flexible specification, while guaranteeing the tractability of the estimation problem; see Chen (2007) for a survey of sieve estimation in semi-nonparametric models. The estimation method is also easy to implement in practice. The sieve ML estimation has been used in various contexts: Chen, Fan, and Tsyrennikov (2006; hereafter, CFT06) considered the sieve estimation of semiparametric multivariate distributions that were modeled using parametric copulas; Bierens (2008) applied the estimation method to the mixed proportional hazard model; and Hu and Schennach (2008) and Chen, Hu, and Lewbel (2009) used the method to estimate nonparametric models with nonclassical measurement errors. The asymptotic theory developed in this paper is based on the results established in the sieve extremum estimation literature (e.g., Bierens, 2014; CFT06; Chen, 2007). A semiparametric version of bivariate threshold crossing models was also considered in Marra and Radice (2011) and Ieva, Marra, Paganoni, and Radice (2014). In contrast to our setting, however, they introduced flexibility for the index function of the threshold, and not for the distribution of the unobservables.

The remainder of this paper is organized as follows. The next section reviews the identification results of HV17, and then discusses the lack of identification in the absence of exclusion restrictions and in the absence of restrictions on the dependence structure of the unobservables. Section 3 introduces the sieve ML estimation framework for the semiparametric class of models defined in Equation (1), and Section 4 establishes the large-sample theory for sieve ML estimators. Sensitivity analysis is conducted in Section 5 by investigating the finite-sample performance of the parametric ML and sieve ML estimates under various specifications. Section 6 presents the empirical example, and Section 7 concludes.

2 | IDENTIFICATION AND FAILURE OF IDENTIFICATION

2.1 | Identification results in Han and Vytlacil (2017)

We first summarize the identification results in HV17. In model (1), let $\underset{(k+1)\times 1}{X} \equiv (1, X_1, \ldots, X_k)'$ and $\underset{l\times 1}{Z} \equiv (Z_1, \ldots, Z_l)'$, and conformably, let $\alpha \equiv (\alpha_0, \alpha_1, \ldots, \alpha_k)'$, $\beta \equiv (\beta_0, \beta_1, \ldots, \beta_k)'$, and $\gamma \equiv (\gamma_1, \gamma_2, \ldots, \gamma_l)'$.

Assumption 1. X and Z satisfy that $(X, Z) \perp (\varepsilon, \nu)$, where "$\perp$" denotes statistical independence.

Assumption 2. (X', Z') does not lie in a proper linear subspace of $\mathbb{R}^{k+l}$ a.s.[2]

Assumption 3. There exists a copula function $C : (0,1)^2 \to (0,1)$ such that the joint distribution $F_{\varepsilon\nu}$ of (ε, ν) satisfies $F_{\varepsilon\nu}(\varepsilon, \nu) = C[F_\varepsilon(\varepsilon), F_\nu(\nu)]$, where F_ε and F_ν are the marginal distributions of ε and ν, respectively, that are strictly increasing and absolutely continuous with respect to Lebesgue measure.[3]

Assumption 4. As scale and location normalizations, $\alpha_1 = \beta_1 = 1$ and $\alpha_0 = \beta_0 = 0$.

A model with alternative scale and location normalizations, $\text{var}(\varepsilon) = \text{var}(\nu) = 1$ and $E[\varepsilon] = E[\nu] = 0$, can be viewed as a reparametrized version of the model with the normalizations given in Assumption 4; see, for example, the reparametrization (Equation (2)) below. For $x \in \text{supp}(X)$ and $z \in \text{supp}(Z)$, write a one-to-one map (by Assumption 3) as

$$s_{xz} \equiv F_\nu(x'\alpha + z'\gamma), \quad r_{0,x} \equiv F_\varepsilon(x'\beta), \quad r_{1,x} \equiv F_\varepsilon(x'\beta + \delta_1). \tag{2}$$

[1] HV17 only showed the sufficiency of this restriction for identification. Mourifié and Méango (2014) showed the necessity of the restriction, but their argument does not exploit all information available in the model; see Section 2.2 of the present paper for further details.

[2] A proper linear subspace of $\mathbb{R}^{k+l}$ is a linear subspace with a dimension strictly less than $k+l$.The assumption is that if M is a proper linear subspace of $\mathbb{R}^{k+l}$,then $\Pr[(X', Z') \in M] < 1$.

[3] Sklar's theorem (e.g., Nelsen, 1999) guarantees the existence of such a copula, which is, in fact, unique becauseF_ε and F_ν are continuous.

Take (x, z) and $(x, \tilde{z})$, for some $x \in \text{supp}(X|Z = z) \cap \text{supp}(X|Z = \tilde{z})$, where $\text{supp}(X|Z)$ is the conditional support of X, given Z. Then, by Assumption 1, model (1) implies that the fitted probabilities are written as

$$\begin{aligned} p_{11,xz} &= C(r_{1,x}, s_{xz}), & p_{11,x\tilde{z}} &= C(r_{1,x}, s_{x\tilde{z}}), \\ p_{10,xz} &= r_{0,x} - C(r_{0,x}, s_{xz}), & p_{10,x\tilde{z}} &= r_{0,x} - C(r_{0,x}, s_{x\tilde{z}}), \\ p_{01,xz} &= s_{xz} - C(r_{1,x}, s_{xz}), & p_{01,x\tilde{z}} &= s_{x\tilde{z}} - C(r_{1,x}, s_{x\tilde{z}}), \end{aligned} \tag{3}$$

where $p_{yd,xz} \equiv \Pr[Y = y, D = d|X = x, Z = z]$ for $(y, d) \in \{0, 1\}^2$. Equation (3) serves as the basis for the identification and estimation of the model. Depending upon whether one is willing to impose an additional assumption on the dependence structure of the unobservables (ε, ν) via $C(\cdot, \cdot)$, the underlying parameters of the model are either point identified or partially identified.

We first consider point identification. The results for point identification can be found in HV17, which we adapt here given Assumption 4. The additional dependence structure can be characterized in terms of the stochastic ordering of the copula parametrized with a scalar parameter.

Definition 1 (Strictly more SI or less SD)**.** Let $C(u_2|u_1)$ and $\tilde{C}(u_2|u_1)$ be conditional copulas, for which $1 - C(u_2|u_1)$ and $1 - \tilde{C}(u_2|u_1)$ are either increasing or decreasing in u_1 for all u_2. Such copulas are referred to as stochastically increasing (SI) or stochastically decreasing (SD), respectively. Then, $\tilde{C}$ is strictly more SI (or less SD) than C if $\psi(u_1, u_2) \equiv \tilde{C}^{-1}[C(u_2|u_1)|u_1]$ is strictly increasing in u_1,[4] which is denoted by $C \prec_S \tilde{C}$.

This ordering is equivalent to having a ranking in terms of the first-order stochastic dominance. Let $(U_1, U_2) \sim C$ and $(\tilde{U}_1, \tilde{U}_2) \sim \tilde{C}$. When $\tilde{C}$ is strictly more SI (less SD) than C is, then $\Pr[\tilde{U}_2 > u_2|\tilde{U}_1 = u_1]$ increases even more than $\Pr[U_2 > u_2|U_1 = u_1]$ does as u_1 increases.[5]

Assumption 5. The copula in Assumption 3 satisfies $C(\cdot, \cdot) = C(\cdot, \cdot; \rho)$ with a scalar dependence parameter $\rho \in \Omega$, is twice differentiable in u_1, u_2 and ρ, and satisfies

$$C(u_1|u_2; \rho_1) \prec_S C(u_1|u_2; \rho_2) \text{ for any } \rho_1 < \rho_2. \tag{4}$$

The meaning of the last part of this assumption is that the copula is ordered in ρ in the sense of the stochastic ordering defined above. This requirement defines the class of copulas that we allow for identification. Many well-known copulas satisfy Equation (4): the normal copula, Plackett copula, Frank copula, Clayton copula, among many others; see HV17 for the full list of copulas and their expressions. Under these assumptions, we first discuss the identification in a fully parametric model.

Assumption 6. F_ε and F_ν are *known* up to means $\mu \equiv (\mu_\varepsilon, \mu_\nu)$ and variances $\sigma^2 \equiv (\sigma_\varepsilon^2, \sigma_\nu^2)$.

Given this assumption, $F_\nu(\nu) = F_{\tilde{\nu}}(\tilde{\nu})$ and $F_\varepsilon(\varepsilon) = F_{\tilde{\varepsilon}}(\tilde{\varepsilon})$, where $F_{\tilde{\nu}}$ and $F_{\tilde{\varepsilon}}$ are the distributions of $\tilde{\nu} \equiv (\nu - \mu_\nu)/\sigma_\nu$ and $\tilde{\varepsilon} \equiv (\varepsilon - \mu_\varepsilon)/\sigma_\varepsilon$, respectively. Define

$$\mathcal{X} \equiv \bigcup_{\substack{z'\gamma \neq \tilde{z}'\gamma \\ z, \tilde{z} \in \text{supp}(Z)}} \text{supp}(X|Z = z) \cap \text{supp}(X|Z = \tilde{z}).$$

Theorem 1. *In model (1), suppose Assumptions 1 -6 hold. Then, $(\alpha', \beta', \delta_1, \gamma, \rho, \mu, \sigma)$ are point identified in an open and convex parameter space if (i) γ is a nonzero vector, and (ii) $\mathcal{X}$ does not lie in a proper linear subspace of $\mathbb{R}^k$ a.s.*

The proof of this theorem is a minor modification of the proof of theorem 5.1 in HV17.

Although the parametric structure on the copula is necessary for the point identification of the parameters, HV17 showed that the parametric assumption for F_ε and F_ν was not necessary. In addition, if we make a large support assumption, we can also identify the nonparametric marginal distributions F_ε and F_ν.

[4]Note that $\psi(u_1, u_2)$ is increasing in u_2 by definition.

[5]In the statistics literature, the SI dependence ordering is also referred to as the (strictly) "more regression dependent" or "more monotone regression dependent" ordering; see Joe (1997) for details.

Assumption 7. (i) The distributions of X_j (for $1 \leq j \leq k$) and Z_j (for $1 \leq j \leq l$) are absolutely continuous with respect to Lebesgue measure. (ii) There exists at least one element X_j in X such that its support conditional on $(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$ is $\mathbb{R}$ and $\alpha_j \neq 0$ and $\beta_j \neq 0$, where, without loss of generality, we let $j = 1$.

Theorem 2. *In model (1), suppose Assumptions 1 -5 and 7(i) hold. Then, $(\alpha', \beta', \delta_1, \gamma, \rho)$ are point identified in an open and convex parameter space if (i) γ is a nonzero vector and (ii) $\mathcal{X}$ does not lie in a proper linear subspace of $\mathbb{R}^k$ a.s. In addition, if Assumption 7(ii) holds, $F_\varepsilon(\cdot)$ and $F_\nu(\cdot)$ are identified up to additive constants.*

An interesting function of the underlying parameters that are point identified under the parametric and semiparametric distributional assumptions is the conditional ATE:

$$\text{ATE}(x) = E[Y_1 - Y_0 | X = x] = F_\varepsilon(x'\beta + \delta_1) - F_\varepsilon(x'\beta). \tag{5}$$

2.2 | Extension of Han and Vytlacil (2017): Identification under conditional independence

The identification analysis of Han and Vytlacil (2017) relies on the full independence assumption (Assumption 1) for (X, Z). The analysis, however, can be easily extended to a case where conditional independence is alternatively assumed. Since this is a more empirically relevant situation, we explore this case in detail here. In the empirical section below, we impose the conditional independence. Let M be a vector of (potentially endogenous) covariates in supp(M).

Assumption 1′. X and Z satisfy that $(X, Z) \perp (\varepsilon, \nu)|M$.

Similarly, we modify Assumptions 2,3 and 5-7 accordingly. Then the following theorems immediately hold by applying the same proof strategies as in Theorems 1 and 2. Let $C_m(u_1, u_2) \equiv C(u_1, u_2|M = m)$ be the conditional copula, and $F_{\varepsilon\nu|m}(\varepsilon, \nu) \equiv F_{\varepsilon\nu|M=m}(\varepsilon, \nu)$, $F_{\varepsilon|m}(\varepsilon) \equiv F_{\varepsilon|M=m}(\varepsilon)$ and $F_{\nu|m}(\nu) \equiv F_{\nu|M=m}(\nu)$ be the conditional distributions.

Theorem 3. *In model (1), suppose Assumptions 1′ and 4 hold. Also, suppose Assumption 2 holds conditional on M, and Assumptions 3 and 5,6 hold with $C_m(u_1, u_2)$, $F_{\varepsilon\nu|m}(\varepsilon, \nu)$, $F_{\varepsilon|m}(\varepsilon)$ and $F_{\nu|m}(\nu)$ instead, for all $m \in$ supp(M). Then, $(\alpha', \beta', \delta_1, \gamma, \rho, \mu, \sigma)$ are point identified in an open and convex parameter space if (i) γ is a nonzero vector and (ii) $\mathcal{X}$ does not lie in a proper linear subspace of $\mathbb{R}^k$ a.s. conditional on M.*

Theorem 4. *In model (1), suppose Assumptions 1′ and 4 hold. Also, suppose Assumptions 2 and 7(i) hold conditional on M, and Assumptions 3 and 5 hold with $C_m(u_1, u_2)$, $F_{\varepsilon\nu|m}(\varepsilon, \nu)$, $F_{\varepsilon|m}(\varepsilon)$ and $F_{\nu|m}(\nu)$ instead, for all $m \in$ supp(M). Then $(\alpha', \beta', \delta_1, \gamma, \rho)$ are point identified in an open and convex parameter space if (i) γ is a nonzero vector; and (ii) $\mathcal{X}$ does not lie in a proper linear subspace of $\mathbb{R}^k$ a.s. In addition, if Assumption 7(ii) holds conditional on M, $F_{\varepsilon|m}(\cdot)$ and $F_{\nu|m}(\cdot)$ are identified up to additive constants for all $m \in$ supp(M).*

2.3 | The failures of identification

In this section, we discuss two sources of identification failure in the class of models (1): the absence of exclusion restrictions and the absence of restrictions on the dependence structure of the unobservables (ε, ν).

2.3.1 | No exclusion restrictions

There are empirical works where Equation (1) is used without excluded instruments; see, for example, N. E. White and Wolaver (2003) and Rhine et al. (2006). Identification in these papers relies on the results of Wilde (2000), who provide an identification argument by counting the number of equations and unknowns in the system. Here, we show that this argument is insufficient for identification. We show that without the excluded instruments (i.e., when $\gamma = 0$) the structural parameters are not identified, even with a full parametric specification of the joint distribution (Assumptions 5 and 6). The existence of common exogenous covariates X in both equations is not very helpful for identification, in a sense that becomes clear below.

Before considering the lack of identification in a general case with possibly continuous X_1 in $X = (1, X_1)$, we start the analysis with binary X_1. Mourifié and Méango (2014) show the lack of identification when there is no excluded instrument in a bivariate probit model with binary X_1. They only provide, however, a numerical counterexample. Moreover, their analysis does not consider the full set of observed fitted probabilities, and hence possibly neglects information that could have contributed to the identification. Here, we provide an analytical counterexample in a more general parametric class

of model (Equation (1)) that nests the bivariate probit model. We show that $(\delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ are not identified, even if the full set of probabilities are used. Note that the reduced-form parameters (μ_v, σ_v) are always identified from the equation for D, and $\alpha = \beta = (0,1)'$ as a normalization using scalar X_1.

Theorem 5. *In model (1) with $X = (1, X_1)$, where $X_1 \in supp(X_1) = \{0,1\}$, suppose that the assumptions in Theorem 1 hold, except that $\gamma = 0$. Then, there exist two element-wise distinct sets of $(\delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ that generate the same observed data.*

In showing this lack-of-identification result, we find a counterexample where the copula density induced by $C(u_1, u_2)$ is symmetric around $u_2 = u_1$ and $u_2 = 1 - u_1$, and the density induced by F_ε is symmetric. Note that the bivariate normal distribution, namely the normal copula with normal marginals, satisfies these symmetry properties. That is, *in the bivariate probit model with a common binary exogenous covariate and no excluded instruments, the structural parameters are*not*identified.*

The proof of Theorem 5 proceeds as follows. Under Assumption 4, let

$$q_0 \equiv F_{\tilde{v}}(-\mu_v/\sigma_v), \qquad q_1 \equiv F_{\tilde{v}}[(1-\mu_v)/\sigma_v],$$
$$t_0 \equiv F_{\tilde{\varepsilon}}(-\mu_\varepsilon/\sigma_\varepsilon), \qquad t_1 \equiv F_{\tilde{\varepsilon}}[(1-\mu_\varepsilon)/\sigma_\varepsilon].$$

Then, we have

$$\tilde{p}_{11,0} = C\{F_{\tilde{\varepsilon}}[F_{\tilde{\varepsilon}}^{-1}(t_0) + \delta_1], q_0; \rho\}, \qquad \tilde{p}_{11,1} = C\{F_{\tilde{\varepsilon}}[F_{\tilde{\varepsilon}}^{-1}(t_1) + \delta_1], q_1; \rho\},$$
$$\tilde{p}_{10,0} = t_0 - C(t_0, q_0; \rho), \qquad \tilde{p}_{10,1} = t_1 - C(t_1, q_1; \rho),$$
$$\tilde{p}_{00,0} = 1 - t_0 - q_0 + C(t_0, q_0; \rho), \qquad \tilde{p}_{00,1} = 1 - t_1 - q_1 + C(t_1, q_1; \rho),$$

where $\tilde{p}_{yd,x} \equiv \Pr[Y = y, D = d | X_1 = x]$. We want to show that, given (q_0, q_1) which are identified from the reduced-form equation, there are two distinct sets of parameter values $(t_0, t_1, \delta_1, \rho)$ and $(t_0^*, t_1^*, \delta_1^*, \rho^*)$ (with $(t_0, t_1, \delta_1, \rho) \neq (t_0^*, t_1^*, \delta_1^*, \rho^*)$) that generate the same observed fitted probabilities $\tilde{p}_{yd,0}$ and $\tilde{p}_{yd,1}$ for all $(y,d) \in \{0,1\}^2$ under some choices of $C(u_1, u_2)$ and F_ε. A detailed proof can be found in the Supporting Information Appendix.

One might argue that the lack of identification in Theorem 5 is due to the limited variation of X. Although this is a plausible conjecture, this does not seem to be the case in the model considered here.[6] We now consider a general case with possibly *continuous* X_1, and discuss what can be said about the existence of two distinct sets of $(\beta, \delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ that generate the same observed data. To this end, define

$$q(x) \equiv F_{\tilde{v}}[(x'\alpha - \mu_v)/\sigma_v], \quad t(x) \equiv F_{\tilde{\varepsilon}}[(x'\beta - \mu_\varepsilon)/\sigma_\varepsilon].$$

Then,

$$p_{11,x} = C\left(F_{\tilde{\varepsilon}}\{F_{\tilde{\varepsilon}}^{-1}[t(x)] + \delta_1\}, q(x); \rho\right),$$
$$p_{10,x} = t(x) - C[t(x), q(x); \rho],$$
$$p_{00,x} = 1 - t(x) - q(x) + C[t(x), q(x); \rho].$$

Similar to the proof strategy for the binary X_1 case, we want to show that, given $(\alpha, \mu_v, \sigma_v)$, there are two distinct sets of parameter values $(\beta, \delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ and $(\beta^*, \delta_1^*, \rho^*, \mu_\varepsilon^*, \sigma_\varepsilon^*)$ that generate the same observed fitted probabilities $p_{yd,x}$ for all $(y,d) \in \{0,1\}^2$ and $x \in \text{supp}(X)$ under some choices of $C(u_1, u_2)$ and F_ε.

Let $t(x) \equiv F_{\tilde{\varepsilon}}(x'\beta) \in (0,1)$ for all x and for some β. Also, choose $\delta_1 = 0$ and some $\rho \in \Omega$. For $\rho^* > \rho$, we want to show that there exists (β^*, δ_1^*) such that, for $t^*(x) \equiv F_{\tilde{\varepsilon}}(x'\beta^*)$,

$$p_{10,x} = t(x) - C[t(x), q(x); \rho] = t^*(x) - C[t^*(x), q(x); \rho^*], \tag{6}$$

$$p_{11,x} = C\left(F_{\tilde{\varepsilon}}\{F_{\tilde{\varepsilon}}^{-1}[t(x)] + 0\}, q(x); \rho\right) = C(s^\dagger(x), q(x); \rho^*), \tag{7}$$

for all x, where

$$s^\dagger(x) = F_{\tilde{\varepsilon}}\{F_{\tilde{\varepsilon}}^{-1}[t^*(x)] + \delta_1^*\}. \tag{8}$$

[6] In fact, in Heckman's (1979) sample selection model under normality, although identification fails with binary exogenous covariates in the absence of the exclusion restriction, it is well known that identification is achieved with continuous covariates by exploiting the nonlinearity of the model (Vella, 1998).

The question is whether we find (β, δ_1, ρ) and $(\beta^*, \delta_1^*, \rho^*)$ such that Equations (6)–(8) hold simultaneously. First, note that, since $\rho^* > \rho$, we have $t^* > t$ and hence $\beta^* \neq \beta$ by the assumption that there is no linear subspace in the space of X. Now, choose $C(\cdot, \cdot; \rho)$ to be a normal copula and choose $\rho = 0$ and $\rho^* = 1$. Then, using arguments similar to those of the binary case (found in the Supporting Information Appendix), we obtain

$$t^*(x) = q(x) + [1 - q(x)]t(x) \tag{9}$$

and $s^\dagger(x) = q(x)t(x)$. Then, Equation (8) can be rewritten as

$$\delta_1^* = F_{\tilde{\varepsilon}}^{-1}[s^\dagger(x)] - F_{\tilde{\varepsilon}}^{-1}[t^*(x)] = F_{\tilde{\varepsilon}}^{-1}[q(x)t(x)] - F_{\tilde{\varepsilon}}^{-1}\{q(x) + [1 - q(x)]t(x)\}. \tag{10}$$

The complication here is to ensure that this equation is satisfied for all x. Note that Equations (9) and (10) are consistent with the definition of a distribution function of a continuous r.v.: $F_{\tilde{\varepsilon}}(+\infty) = 1, F_{\tilde{\varepsilon}}(-\infty) = 0$, and $F_{\tilde{\varepsilon}}(\varepsilon)$ is strictly increasing. We can then show numerically that a distribution function that is close to a normal distribution satisfies the conditions with a particular choice of (β^*, δ_1^*); see Figure 1. Although no formal derivation of the counterexample is given, this result suggests the following:

(i) In the bivariate probit model with continuous common exogenous covariates and no excluded instruments, the parameters will be, *at best,* weakly identified.
(ii) This also implies that, in the semiparametric model considered in Theorem 2, the structural parameters and the marginal distributions are not identified without an exclusion restriction, even if X_1 has large support.

2.3.2 | No restrictions on dependence structures

When the restriction imposed on $C(\cdot, \cdot)$ (i.e., Assumption 5) is completely relaxed, the underlying parameters of model (1) may fail to be identified, regardless of whether the exclusion restriction holds. That is, a structure describing how the unobservables (ε, v) are dependent on each other is necessary for identification. This is closely related to the results in the literature that the treatment parameters (which are lower dimensional functions of the individual parameters) in triangular models similar to Equation (1) are only partially identified without distributional assumptions; see Bhattacharya, Shaikh, and Vytlacil (2008), Chiburis (2010), Shaikh and Vytlacil (2011), and Mourifié (2015).

Suppose Assumptions 1,2,3,4 hold. Then the model becomes a semiparametric threshold crossing model in that the joint distribution is completely unspecified. Then, as a special case of Shaikh and Vytlacil (2011), one can easily derive bounds for the ATE $F_\varepsilon(x'\beta + \delta_1) - F_\varepsilon(x'\beta)$. The sharpness of these bounds is shown in their paper under a rectangular support assumption for (X, Z), which is, in turn, relaxed in Mourifié (2015). In addition, using Assumption 6, one can also derive bounds for the individual parameters $x'\beta$ and δ_1, as shown in Chiburis (2010). When there are no excluded

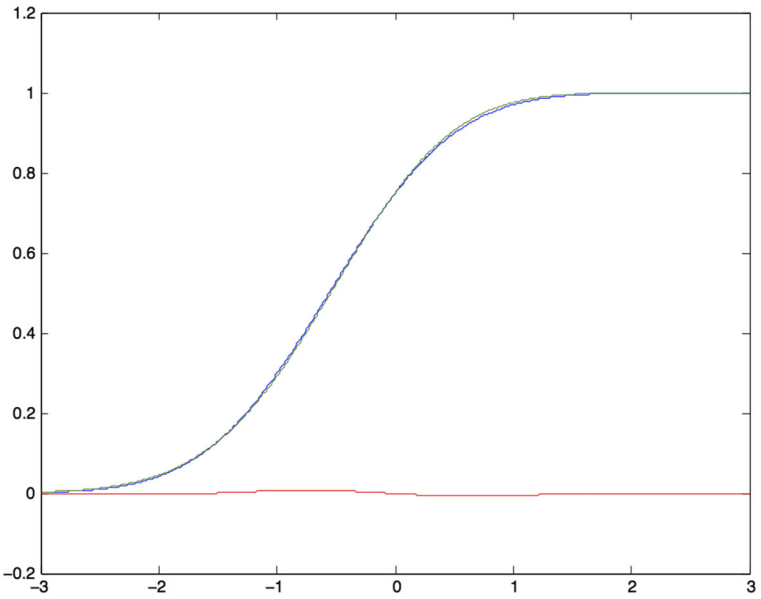

FIGURE 1 A numerical calculation of a distribution function under which identification fails (blue line), compared with a normal distribution function (green line) [Colour figure can be viewed at wileyonlinelibrary.com]

instruments in the model, Chiburis shows that the bounds on the ATE do not improve on the bounds of Manski (1990), whose argument applies to the individual parameters.

3 | SIEVE AND PARAMETRIC ML ESTIMATIONS

Based on the identification results, we now consider estimation. Let $\psi \equiv (\alpha', \beta', \delta_1, \gamma, \rho)$ denote the vector of the structural individual parameters. Let f_ϵ and f_ν be the density functions associated with the distribution functions F_ϵ and F_ν, respectively, of the unobservables. Then, $(\psi', f_\epsilon, f_\nu)'$ is the set of parameters in the *semiparametric version* of the model. The model becomes fully parametric, once the infinite-dimensional parameters f_ϵ and f_ν are fully characterized by some finite-dimensional parameters; that is, $f_\epsilon(\cdot;\eta_\epsilon)$ and $f_\nu(\cdot;\eta_\nu)$ for $\eta_\epsilon \in \mathbb{R}^{d_{\eta_\epsilon}}$ and $\eta_\nu \in \mathbb{R}^{d_{\eta_\nu}}$. This yields $(\psi', \eta_\epsilon', \eta_\nu')'$ to be the set of parameters in the *parametric version* of the model. For either case, the parameter of the model is denoted by θ for convenience; that is, $\theta \equiv (\psi', f_\epsilon, f_\nu)'$ in the semiparametric model and $\theta \equiv (\psi', \eta_\epsilon', \eta_\nu')'$ in the parametric model. For the rest of this paper, we explicitly express θ_0 to be the true parameter value for θ. This applies to all the other parameter expressions.

Let $\tilde{\Psi}$ be the parameter space for ψ. For the parametric model, the spaces for the finite-dimensional parameters η_ϵ and η_ν are denoted by $\mathbf{H}_\epsilon \subseteq \mathbb{R}^{d_{\eta_\epsilon}}$ and $\mathbf{H}_\nu \subseteq \mathbb{R}^{d_{\eta_\nu}}$, respectively. Then, the parameter space $\tilde{\Theta}$ for $\theta \equiv (\psi', \eta_\epsilon', \eta_\nu')'$ becomes a Cartesian product of $\tilde{\Psi}$, $\mathbf{H}_\epsilon$, and $\mathbf{H}_\nu$; that is, $\tilde{\Theta} \equiv \tilde{\Psi} \times \mathbf{H}_\epsilon \times \mathbf{H}_\nu \subseteq \mathbb{R}^{d_\psi + d_{\eta_\epsilon} + d_{\eta_\nu}}$, in the parametric model.[7] For the semiparametric model, we consider the following function spaces as the spaces for f_ϵ and f_ν:

$$\mathcal{F}_j \equiv \left\{ f = q^2 : q \in \mathcal{F}, \int \{q(x)\}^2 dx = 1 \right\}, \tag{11}$$

where $j \in \{\epsilon, \nu\}$ and $\mathcal{F}$ is a space of functions, which we specify later. Then, the parameter space $\tilde{\Theta}$ of $\theta \equiv (\psi', f_\epsilon, f_\nu)'$ can be written as $\tilde{\Theta} \equiv \tilde{\Psi} \times \mathcal{F}_\epsilon \times \mathcal{F}_\nu$ in the semiparametric model. Note that the function spaces $\mathcal{F}_\epsilon$ and $\mathcal{F}_\nu$ contain functions that are nonnegative.

We adopt the ML method to estimate the parameters in the model. Let $\{W_i = \{Y_i, D_i, X_i', Z_i'\} : i = 1, 2, \ldots, n\}$ be the random sample. For both parametric and semiparametric models with corresponding θ, we define the conditional density function of (Y_i, D_i) conditional on $(X_i', Z_i')'$ as

$$f(Y_i, D_i | X_i, Z_i; \theta) = \prod_{y,d=0,1} [p_{yd}(X_i, Z_i; \theta)]^{\mathbf{1}\{Y_i = y, D_i = d\}},$$

where $p_{yd}(x, z; \theta)$ abbreviates the right-hand-side expression that equates to $p_{yd,xz}$ in Equation (3). Then, the log of density $l(\theta, w) \equiv \log f(y, d|x, z; \theta)$ becomes

$$l(\theta, W_i) \equiv \sum_{y,d=0,1} \mathbf{1}_{yd}(Y_i, D_i) \cdot \log p_{yd}(X_i, Z_i; \theta), \tag{12}$$

where $\mathbf{1}_{yd}(Y_i, D_i) \equiv \mathbf{1}\{Y_i = y, D_i = d\}$. Consequently, the log-likelihood function can be written as $Q_n(\theta) = \frac{1}{n} \sum_{i=1}^n l(\theta, W_i)$.

Now, the ML estimator $\tilde{\theta}_n$ of $\theta_0 \equiv (\psi_0', \eta_{\epsilon 0}, \eta_{\nu 0})'$ in the parametric model is defined as

$$\tilde{\theta}_n \equiv \arg\max_{\theta \in \tilde{\Theta}} Q_n(\theta). \tag{13}$$

For the semiparametric model, let $\mathcal{F}_{\epsilon n}$ and $\mathcal{F}_{\nu n}$ be appropriate sieve spaces for $\mathcal{F}_\epsilon$ and $\mathcal{F}_\nu$, respectively, and let $f_{\epsilon n}(\cdot; a_{\epsilon n})$ and $f_{\nu n}(\cdot; a_{\nu n})$ be the sieve approximations of f_ϵ and f_ν on their sieve spaces $\mathcal{F}_{\epsilon n}$ and $\mathcal{F}_{\nu n}$, respectively. Then, we define the sieve ML estimator $\hat{\theta}_n$ of $\theta_0 \equiv (\psi_0', f_{\epsilon 0}, f_{\nu 0})'$ in the semiparametric model as follows:

$$\hat{\theta}_n \equiv \arg\max_{\theta \in \tilde{\Theta}_n} Q_n(\theta), \tag{14}$$

where $\tilde{\Theta}_n \equiv \tilde{\Psi} \times \mathcal{F}_{\epsilon n} \times \mathcal{F}_{\nu n}$ is the sieve space for θ.

[7] For example, if one imposes Assumption 6, then $\eta_\epsilon = (\mu_\epsilon, \sigma_\epsilon)'$ and $\eta_\nu = (\mu_\nu, \sigma_\nu)'$.

With the parameter spaces $\mathcal{F}_\epsilon$ and $\mathcal{F}_\nu$ in Equation (11), we are interested in a class of "smooth" univariate square root density functions. Specifically, we assume that $\sqrt{f_\epsilon}$ and $\sqrt{f_\nu}$ belong to the class of p-smooth functions and we restrict our attention to linear sieve spaces for $\mathcal{F}_\epsilon$ and $\mathcal{F}_\nu$.[8] In this case, the choice of sieve spaces for $\mathcal{F}_\epsilon$ and $\mathcal{F}_\nu$ depends on the supports of ϵ and ν. If the supports are bounded, then one can use the polynomial sieve, trigonometric sieve, or cosine sieve. When the supports are unbounded, then we can use the Hermite polynomial sieve or the spline wavelet sieve.

In this paper, we implicitly assume that the copula function is correctly specified. As mentioned earlier, using a parametric copula may lead to model misspecification. It is well known that when the model is misspecified, the ML estimator converges to a pseudo-true value which minimizes the Kullback–Leibler (KL) divergence (e.g., White, 1982). This result applies to a semiparametric model (e.g., Chen & Fan, 2006a, 2006b) as in our semiparametric case. We, however, do not investigate the asymptotic properties of the sieve estimators under copula misspecification, as it is beyond the scope of this paper. Instead, later in simulation, we investigate how the copula misspecification affects the performance of estimators.[9]

4 | ASYMPTOTIC THEORY FOR SIEVE ML ESTIMATORS

In this section, we provide the asymptotic theory for the sieve ML estimator $\hat{\theta}_n$ of $\theta \equiv (\psi', f_\epsilon, f_\nu)'$ in the semiparametric model. This theory will be useful for practitioners to conduct inference. The asymptotic theory for the ML estimator $\tilde{\theta}_n$ of $\theta \equiv (\psi', \eta_\epsilon, \eta_\nu)'$ in the parametric model is relatively standard and can be found in Newey and McFadden (1994), for example. The theory establishes that the parametric ML estimator is consistent, asymptotically normal, and efficient under some regularity conditions. To investigate the asymptotic properties of the sieve ML estimator, we slightly modify our model as follows.

Let $G(\cdot)$ be a strictly increasing function mapping from $\mathbb{R}$ to $[0, 1]$. We further assume that G is differentiable and that its derivative $g(x) \equiv \frac{dG(x)}{dx}$ is bounded away from zero on $\mathbb{R}$. Then, without loss of generality (e.g., Bierens, 2014), we consider the following transformation of $F_{\epsilon 0}$ and $F_{\nu 0}$ as

$$F_{\epsilon 0}(x) = H_{\epsilon 0}[G_\epsilon(x)], \quad F_{\nu 0}(x) = H_{\nu 0}[G_\nu(x)], \tag{15}$$

where $H_{\epsilon 0}(\cdot)$ and $H_{\nu 0}(\cdot)$ are unknown distribution functions on $[0, 1]$. For G, we can choose the standard normal distribution function or the logistic distribution function. Since we assume that the distribution functions of ϵ and ν admit density functions, we require that $H_{\epsilon 0}$ and $H_{\nu 0}(\cdot)$ be differentiable, and write their derivatives as $h_{\epsilon 0}(\cdot)$ and $h_{\nu 0}(\cdot)$, respectively. For each $j \in \{\epsilon, \nu\}$, let $\mathcal{H}_j \equiv \{h_j = q^2 : q \in \mathcal{F}\}$ for some function space $\mathcal{F}$. With this modification, we redefine the parameter as $\theta = (\psi', h_\epsilon, h_\nu)' \in \tilde{\Theta}^\dagger \equiv \tilde{\Psi} \times \mathcal{H}_\epsilon \times \mathcal{H}_\nu$. Note that, using the transformation of the distribution functions in Equation (15), the unknown infinite-dimensional parameters are defined on a bounded domain. In the Supporting Information Appendix, we show that the transformation does not affect the identification result.

We redefine the parameter space to facilitate developing the asymptotic theory. The identification requires that the space of the finite-dimensional parameter $\tilde{\Psi}$ be open and convex (see Theorems 1 and 2), and thus $\tilde{\Psi}$ cannot be compact. We introduce an "optimization space" that contains the true parameter ψ_0 and consider it as the parameter space of ψ. Formally, we restrict the parameter space for estimation in the following way.

Assumption 9. There exists a compact and convex subset $\Psi \subseteq \tilde{\Psi}$ such that $\psi_0 \in int(\Psi)$, where $int(A)$ is the interior of the set A.

With the optimization space, we define the parameter space as $\Theta \equiv \Psi \times \mathcal{H}_\epsilon \times \mathcal{H}_\nu$, and the corresponding sieve space is denoted by $\Theta_n \equiv \Psi \times \mathcal{H}_{\epsilon n} \times \mathcal{H}_{\nu n}$. Then, the sieve ML estimator in Equation (14) is also redefined as follows:

$$\hat{\theta}_n \equiv \underset{\theta \in \Theta_n}{argmax} Q_n(\theta). \tag{16}$$

[8] A definition of p-smooth functions can be found in Chen (2007, p. 5570) or CFT06 (p. 1230). We give the formal definition of p-smooth functions in Section 4.

[9] For related issues of copula misspecification, refer to Chen and Fan, 2006a, or Liao and Shi, 2017, for example. In particular, Chen and Fan (2006a) propose a test procedure for model selection that is based on the test of Vuong (1989). Liao and Shi (2017) extend Vuong's test to cases where models contain infinite-dimensional parameters and propose a uniformly asymptotically valid Vuong test for semi/nonparametric models. Their setting encompasses those models that can be estimated by the sieve ML as a special case.

4.1 | Consistency of the sieve ML estimators

We begin by showing the consistency of the sieve ML estimator. Since the parameter involves both finite- and infinite-dimensional objects, we establish the consistency of the sieve ML estimators with respect to a pseudo distance function d_c on $\Theta\times\Theta$.[10] All of the norms and the definitions of function spaces in this paper are provided in the Supporting Information Appendix.

We present the following assumptions, under which the sieve ML estimator in Equation (16) is consistent with respect to the pseudo-metric $d_c(\cdot,\cdot)$.

Assumption 10. There exists a measurable function $\underline{p}(X,Z)$ such that for all $\theta\in\Theta$ and for all $y,d=0,1$, $p_{yd,XZ}(\theta)\geq \underline{p}(X,Z)$, with $E|\log(\underline{p}(X,Z))|<\infty$ and $E\left[\frac{1}{\underline{p}(X,Z)^2}\right]<\infty$.

Assumption 11. $\{W_i : i=1,2,\ldots,n\}$ is a random sample, with $E\left[||(X_i',Z_i')'||_E^2\right]<\infty$.

Assumption 12. (i) $\sqrt{h_{\epsilon 0}},\sqrt{h_{v0}}\in\Lambda_R^p([0,1])$, with $p>\frac{1}{2}$ and some $R>0$; (ii) $\mathcal{H}_\epsilon=\mathcal{H}_v=\mathcal{H}$ where $\mathcal{H}\equiv\left\{h=q^2 : q\in\Lambda_R^p([0,1]),\int_0^1 q=1\right\}$, with R being defined as in (i) and $\Lambda_R^p([0,1])$ being a Hölder ball with radius R; (iii) the density functions $h_{\epsilon 0}$ and h_{v0} are bounded away from zero on $[0,1]$.

Assumption 13. (i) $\mathcal{H}_{\epsilon n}=\mathcal{H}_{vn}\equiv\{h\in\mathcal{H} : h(x)=p^{k_n}(x)'a_{k_n},a_{k_n}\in\mathbb{R}^{k_n},||h||_\infty<2R^2\}$, where $k_n\to\infty$ and $k_n/n\to 0$ as $n\to\infty$; (ii) for all $j\geq 1$, we have $\Theta_j\subseteq\Theta_{j+1}$, and there exists a sequence $\{\pi_j\theta_0\}_j$ such that $d_c(\pi_j\theta_0,\theta_0)\to 0$ as $j\to\infty$.

Assumption 14. For $j=1,2$, let $C_j(u_1,u_2;\rho)\equiv\frac{\partial C(u_1,u_2;\rho)}{\partial u_j}$ and $C_\rho(u_1,u_2;\rho)\equiv\frac{\partial C(u_1,u_2;\rho)}{\partial\rho}$. The derivatives $C_j(\cdot,\cdot;\cdot)$ and $C_\rho(\cdot,\cdot;\cdot)$ are uniformly bounded for all $j=1,2$.

Assumption 10 guarantees that the log-likelihood function $l(\theta,W_i)$ is well defined for all $\theta\in\Theta$ and that $Q_0(\theta_0)>-\infty$. Assumption 11 restricts the DGP, and assumes the existence of moments of the data. Assumption 12 defines the parameter space and implies that the infinite-dimensional parameters are in some smooth class called a Hölder class. Note that conditions (i) and (ii) in Assumption 12 together imply that $h_{\epsilon 0}$ and h_{v0} belong to $\Lambda_{\tilde{R}}^p([0,1])$, where $\tilde{R}\equiv 2^{m+1}R^2<\infty$.[11] Thus we may assume that $h_{\epsilon 0}$ and h_{v0} belong to a Hölder ball with smoothness p under Assumption 12.[12] The condition that $\mathcal{H}_\epsilon$ and $\mathcal{H}_v$ are the same can be relaxed, but it is imposed for simplicity. The first part of Assumption 13 restricts our choice of sieve spaces for $\mathcal{H}_\epsilon$ and $\mathcal{H}_v$ to linear sieve spaces with order k_n. This can be relaxed so that the choice of k_n is different for h_ϵ and h_v. The latter part of Assumption 13 requires that the sieve space be chosen appropriately so that the unknown parameters can be well approximated. Because the unknown infinite-dimensional parameters belong to a Hölder ball and are defined on bounded supports, we can choose the polynomial sieve, trigonometric sieve, cosine sieve, or spline sieve.[13] For example, if we choose the polynomial sieve or the spline sieve, then one can show that $d_c(\pi_{k_n}\theta_0,\theta_0)=O(k_n^{-p})$ (e.g., Lorentz, 1966). Assumption 14 imposes the boundedness of the derivatives of the copula function.

The following theorem demonstrates that under the above assumptions the sieve estimator $\hat{\theta}_n$ is consistent with respect to the pseudo metric, d_c.

Theorem 6. *Suppose that Assumptions 1-5 and 7 hold. If Assumptions 9-14 are satisfied, then* $d_c(\hat{\theta}_n,\theta_0)\xrightarrow{p}0$.

[10] It is important to choose appropriate norms to ensure the compactness of the original parameter space, as compactness plays a key role in establishing the asymptotic theory. Since the parameter space is infinite dimensional, it may be compact under certain norms but not under other norms. An infinite-dimensional space that is closed and bounded is not necessarily compact, and thus it is more demanding to show that the parameter space is compact under certain norms. To overcome this difficulty, we take the approach introduced by Gallant and Nychka (1987), which uses two norms to obtain the consistency. Their idea is to use the strong norm to define the parameter space as a ball, and then to ensure the compactness of the parameter space using the consistency norm. In our setting, the Hölder norm is the strong norm and $||\cdot||_c$ is the consistency norm. Related to this issue, Freyberger and Masten (2019) recently extend the idea to more cases and present compactness results for several parameter spaces.

[11] See the Supporting Information Appendix for details.

[12] These conditions implicitly define the strong norm (Hölder norm).

[13] Refer to Chen (2007) or CFT06 for details on the choice of sieve spaces.

4.2 | Convergence rates

In this section, we derive the convergence rate of the sieve ML estimator. The convergence rate provides information on how fast the estimator converges to the true parameter value. Heuristically, the faster the convergence rate, the larger the effective sample size is for estimation. The next theorem demonstrates the convergence rate of the sieve ML estimator with respect to the L^2-norm $||\cdot||_2$.

Theorem 7. *Suppose that Assumptions 1-5, 7, and 9-14 hold. If Assumption B.1 in the Supporting Information Appendix additionally holds, then we have* $||\hat{\theta}_n - \theta_0||_2 = O_p\left(\max\left\{\sqrt{k_n/n}, k_n^{-p}\right\}\right)$. *Furthermore, if we choose* $k_n \propto n^{\frac{1}{2p+1}}$, *then we have* $||\hat{\theta}_n - \theta_0||_2 = O_p\left(n^{-\frac{p}{2p+1}}\right)$.

The former convergence rate is standard in the literature, where the first term corresponds to variance, which increases in k_n, and the second term corresponds to the approximation error $||\theta_0 - \pi_k\theta_0||_2$, which decreases in k_n. The choice of $k_n \propto n^{\frac{1}{2p+1}}$ yields the optimal convergence rate, which is slower than the parametric rate ($n^{-1/2}$). Note that this rate increases with the degree of smoothness, p.

4.3 | Asymptotic normality of smooth functionals

We now establish the asymptotic normality of smooth functionals. The parameters in our model contains both finite- and infinite-dimensional parameters, and many objects of interest are written as functionals of both types of the parameters. The results of this section can be used to calculate the standard error of the estimate of a functional of interest (including the individual finite-dimensional parameters), or to conduct inference (i.e., testing hypotheses and constructing confidence intervals) based on normal approximation.

Before proceeding, we strengthen the smoothness condition in Assumption 5. Let $C_{ij}(u_1, u_2; \rho)$ denote the second-order partial derivative of a copula function $C(u_1, u_2; \rho)$ with respect to i and j, for $i, j \in \{u_1, u_2, \rho\}$.

Assumption 15. The copula function $C(u_1, u_2; \rho)$ is twice continuously differentiable with respect to u_1, u_2, and ρ, and its first- and second- order partial derivatives are well defined in a neighborhood of θ_0.

Let $\mathbb{V}$ be the linear span of $\Theta - \{\theta_0\}$. For $t \in [0, 1]$, define the directional derivative of $l(\theta, W)$ at the direction $v \in \mathbb{V}$ as

$$\left.\frac{dl(\theta_0 + tv, W)}{dt}\right|_{t=0} \equiv \lim_{t\to 0}\frac{l(\theta_0 + tv, W) - l(\theta_0)}{t} = \frac{\partial l(\theta_0, W)}{\partial \psi'}v_\psi + \sum_{j\in\{\epsilon,\nu\}}\frac{\partial l(\theta_0, W)}{\partial h_j}[v_j], \tag{17}$$

where $\frac{\partial l(\theta_0,W)}{\partial \psi'}v_\psi$, $\frac{\partial l(\theta_0,W)}{\partial h_\epsilon}[v_\epsilon]$, and $\frac{\partial l(\theta_0,W)}{\partial h_\nu}[v_\nu]$ are given by Equations B.4–B.6 in the Supporting Information Appendix. If we denote the closed linear span of $\mathbb{V}$ under the Fisher norm $||\cdot||$ by $\bar{\mathbb{V}}$, then $(\bar{\mathbb{V}}, ||\cdot||)$ is a Hilbert space.

Let $T : \Theta \to \mathbb{R}$ be a functional. For any $v \in \mathbb{V}$, we write

$$\frac{\partial T(\theta_0)}{\partial \theta'}[v] \equiv \lim_{t\to 0}\frac{T(\theta_0 + tv) - T(\theta_0)}{t},$$

provided the right-hand-side limit is well defined. The following assumption characterizes the smoothness of the functional T.

Assumption 16. The following conditions hold: (i) There exist constants $w > 1 + \frac{1}{2p}$ and a small $\epsilon_0 > 0$ such that for any $v \in \mathbb{V}$ with $||v|| \le \epsilon_0$

$$\left|T(\theta_0 + v) - T(\theta_0) - \frac{\partial T(\theta_0)}{\partial \theta'}[v]\right| = O(||v||^w).$$

(ii) For any $v \in \mathbb{V}$, $T(\theta_0 + tv)$ is continuously differentiable in $t \in [0, 1]$ around $t = 0$, and

$$\left\|\frac{\partial T(\theta_0)}{\partial \theta'}\right\| \equiv \sup_{v\in\mathbb{V}, ||v||>0}\frac{\left|\frac{\partial T(\theta_0)}{\partial \theta'}[v]\right|}{||v||} < \infty.$$

Assumption 16 defines a smooth functional T and guarantees the existence of $v^* \in \bar{\mathbb{V}}$ such that $< v^*, v >= \frac{\partial T(\theta_0)}{\partial \theta'}[v]$ for all $v \in \mathbb{V}$ and $||v^*||^2 = \left\| \frac{\partial T(\theta_0)}{\partial \theta'} \right\|^2$. Here, we call v^* the Riesz representer for the functional T.

The next assumption requires that the Riesz representer be well approximated over the sieve space and that it converges at a rate with respect to the Fisher norm.

Assumption 17. There exists $\pi_n v^* \in \Theta_n - \{\theta_0\}$ such that $||\pi_n v^* - v^*|| = o(n^{-1/4})$.

The following proposition states that the plug-in sieve ML estimator $T(\hat{\theta}_n)$ of $T(\theta_0)$ is $\sqrt{n}$ -asymptotically normally distributed under certain conditions. The technical conditions (Assumptions B.1, B.2, and B.3) can be found in the Supporting Information Appendix.

Proposition 1. *Suppose that Assumptions 1–5, 7, 9-17, and B.1–B.3 are satisfied. If* $k_n \propto n^{\frac{1}{2p+1}}$, *then we have*

$$\sqrt{n}(T(\hat{\theta}_n) - T(\theta_0)) \xrightarrow{d} \mathcal{N}\left(0, \left\| \frac{\partial T(\theta_0)}{\partial \theta'} \right\|^2\right).$$

It is worth noting that, although the parameter $T(\theta_0)$ contains an infinite-dimensional object (i.e., the marginal distributions of ϵ and v), the sieve plug-in estimator is $\sqrt{n}$-estimable due to the fact that T is a smooth functional.

4.3.1 | Example 1: Asymptotic normality for the finite-dimensional parameter ψ_0

The finite-dimensional parameter ψ_0 is a special case of the smooth functionals. Here, we demonstrate the asymptotic normality of the sieve estimator of the finite-dimensional parameter ψ_0.

Theorem 8. *Suppose that Assumptions 1 -5, 7, 9-15, 17, and B1–B4 hold. Then, we have*

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_*(\psi_0)^{-1}\right), \tag{18}$$

and the form of $\mathcal{I}_*(\psi)$ *is given in the Supporting Information Appendix.*

The covariance matrix in Equation (18) needs to be estimated. To do so, CFT06 adopted the covariance estimation method proposed by Ai and Chen (2003). Since an infinite-dimensional optimization is involved in calculating S_{ψ_0}, we provide a sieve estimator of $\mathcal{I}_*(\psi_0)^{-1}$. The sieve spaces for b_ϵ and b_v can be the same as those for h_ϵ and h_v, respectively. As in Ai and Chen, we first estimate efficient score functions by solving the following minimization problem: for all $k = 1, 2, \ldots, d_\psi$

$$(\hat{b}_{\epsilon k}, \hat{b}_{vk}) \equiv \underset{(b_{\epsilon k}, b_{vk}) \in \mathcal{H}_{\epsilon n} \times \mathcal{H}_{vn}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial l(\hat{\theta}_n, W_i)}{\partial \psi_k} - \left(\frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\epsilon}[b_{\epsilon k}] + \frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_v}[b_{vk}] \right) \right\}^2.$$

Let $\hat{b}_j = (\hat{b}_{j1}, \hat{b}_{j2}, \ldots, \hat{b}_{jd_\psi})'$ for given $j \in \{\epsilon, v\}$ and compute

$$\hat{\mathcal{I}}_*(\hat{\psi}_n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \left[\frac{\partial l(\hat{\theta}_n, W_i)}{\partial \psi} - \left(\frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\epsilon}[\hat{b}_\epsilon] + \frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_v}[\hat{b}_v] \right) \right] \times \left[\frac{\partial l(\hat{\theta}_n, W_i)}{\partial \psi} - \left(\frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\epsilon}[\hat{b}_\epsilon] + \frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_v}[\hat{b}_v] \right) \right]' \right\}$$

to obtain a consistent estimator of $\mathcal{I}_*(\psi_0)$. We now summarize this result as follows:

Theorem 9. *Suppose that assumptions in Theorem 8 hold. Then,* $\hat{\mathcal{I}}_*(\hat{\psi}_n) = \mathcal{I}_*(\psi_0) + o_p(1)$.

The proof of the theorem can be found in theorem 5.1 in Ai and Chen (2003).

4.3.2 | Example 2: Asymptotic normality for the conditional ATE

We now consider the conditional ATE, $E[Y_1 - Y_0 | X = x] = F_{\epsilon 0}(x'\beta_0 + \delta_{10}) - F_{\epsilon 0}(x'\beta_0)$. From Proposition 1, we provide the asymptotic normality of the sieve plug-in estimator of the conditional ATE:

Theorem 10. *Let $x \in supp(X)$ be given. Suppose that the conditions in Proposition 1 hold with $T(\theta_0) = ATE(\theta_0; x)$. Then, we have*

$$\sqrt{n}(ATE(\hat{\theta}_n; x) - ATE(\theta_0; x)) \xrightarrow{d} \mathcal{N}\left(0, \left\| \frac{\partial ATE(\theta_0; x)}{\partial \theta'}[v] \right\|^2\right), \tag{19}$$

where $\left\| \frac{\partial ATE(\theta_0;x)}{\partial \theta'}[v] \right\|^2 = \sup_{v \in \mathbb{V}, ||v|| > 0} \frac{\left| \frac{\partial ATE(\theta_0;x)}{\partial \theta'}[v] \right|}{||v||}$, *and the form of* $\frac{\partial ATE(\theta_0;x)}{\partial \theta'}[v]$ *is given by Equation B.7 in the Supporting Information Appendix.*

Furthermore, the asymptotic variance in Equation (19) can be estimated as follows:

$$\hat{\sigma}^2_{ATE(\theta;x)} \equiv \max_{v \in \Theta_n} \left\| \frac{\partial ATE(\hat{\theta}_n; x)}{\partial \theta'}[v] \right\|^2.$$

4.4 | Weighted bootstrap

The asymptotic variances characterized in the previous subsection can be estimated using the sieve methods. In practice, estimating asymptotic variances may be sensitive to the choice of the number of sieve approximation terms. Furthermore, when the dimension of θ_0 is large, it is relatively cumbersome to estimate the asymptotic variance of the sieve estimator for the finite-dimensional parameter. In this subsection, we briefly discuss the weighted bootstrap as an alternative procedure.

For general semiparametric M-estimation, Ma and Kosorok (2005) and Cheng and Huang (2010) provide the validity of the weighted bootstrap for finite-dimensional parameters in a class of semiparametric models that includes our model. Related to these results, Chen and Pouzo (2009) provide the bootstrap validity in semiparametric conditional moment models. We do not pursue to prove the bootstrap validity in this paper, as these references sufficiently address it. In our empirical exercise, we use the weighted bootstrap scheme proposed in these papers to obtain the standard errors of the estimated functionals of interest. Let $T(\theta_0)$ be a smooth functional of interest and B be the number of bootstrap iterations. The weighted bootstrap is carried out as follows:

1. For each $b = 1, 2, \ldots, B$, let $\{B_i^{(b)} : i = 1, 2, \ldots, n\}$ be a random sample generated from a positive random variable B_i such that $EB_i = 1$, $\text{var}(B_i) = 1$, and is independent of $\{W_i : i = 1, 2, \ldots, n\}$.[14]
2. For each bootstrap iteration $b = 1, 2, \ldots, B$, define $\hat{\theta}_n^{*(b)}$ be a bootstrap estimate of θ_0:

$$\hat{\theta}_n^{*(b)} \equiv \arg\max_{\theta \in \tilde{\Theta}_n} Q_n^{*(b)}(\theta),$$

   where $Q_n^{*(b)}(\theta) \equiv \frac{1}{n}\sum_{i=1}^n B_i^{(b)} \cdot l(\theta, W_i)$. Obtain the bootstrap estimate of the functional of interest by using $\hat{\theta}_n^{*(b)}$ and denote it by $T(\hat{\theta}_n^{*(b)})$.
3. The bootstrap standard error of $T(\hat{\theta}_n)$ is given by $\sqrt{\frac{1}{B}\sum_{b=1}^B \left(T(\hat{\theta}_n^{*(b)}) - \bar{T}_B^*\right)}$, where $\bar{T}_B^* \equiv \frac{1}{B}\sum_{b=1}^B T(\hat{\theta}_n^{*(b)})$.

One may use the bootstrap standard errors to construct confidence intervals, and such confidence intervals rely on the normal approximation. As an alternative to the normal approximation, one can use percentile confidence intervals. For a small $p \in (0, 1)$, a $(1-p) \times 100\%$ percentile confidence interval for a functional $T(\theta_0)$ is constructed as follows:

$$\text{PCI}(p) \equiv \left[Q_T^*(p/2), \quad Q_T^*(1-p/2)\right],$$

where $Q_T^*(\tau)$ is the τth quantile of bootstrap estimates $\{T(\hat{\theta}_n^{*(b)}) : b = 1, 2, \ldots, B\}$. We suggest that practitioners use the percentile confidence intervals rather than the confidence intervals with the bootstrap standard errors.

5 | MONTE CARLO SIMULATION AND SENSITIVITY ANALYSIS

In this section, we conduct a sensitivity analysis via Monte Carlo simulation exercises to provide guidance for empirical researchers. To this end, we investigate the finite-sample performance of the sieve ML estimators of the finite-dimensional

[14]Note that the condition on the variance of B_i can be relaxed. In our empirical example, we use $B_i \sim \exp(1)$.

TABLE 1 Correct specification ($n = 500$) (true marginal: normal)

	γ	δ_1	ρ_{sp}	ATE		γ	δ_1	ρ_{sp}	ATE
Parametric estimation, Gaussian copula					*Semiparametric estimation, Gaussian copula*				
True values	0.8000	1.1000	0.5000	0.3643	True values	0.8000	1.1000	0.5000	0.3643
Estimate	0.8074	1.1469	0.4956	0.3657	Estimate	0.8070	1.1577	0.5037	0.3584
SD	0.0934	0.3954	0.1537	0.0897	SD	0.0940	0.4141	0.1528	0.0935
Bias	0.0074	0.0469	-0.0044	0.0014	Bias	0.0070	0.0577	0.0038	-0.0060
RMSE	0.0936	0.3982	0.1537	0.0897	RMSE	0.0943	0.4181	0.1528	0.0937
Parametric estimation, Frank copula					*Semiparametric estimation, Frank copula*				
True values	0.8000	1.1000	0.5000	0.3643	True values	0.8000	1.1000	0.5000	0.3643
Estimate	0.8027	1.1450	0.4909	0.3681	Estimate	0.8028	1.1556	0.4981	0.3598
SD	0.0936	0.3379	0.1310	0.0781	SD	0.0943	0.3588	0.1314	0.0829
Bias	0.0027	0.0450	-0.0091	0.0037	Bias	0.0028	0.0556	-0.0019	-0.0045
RMSE	0.0936	0.3409	0.1313	0.0781	RMSE	0.0944	0.3631	0.1314	0.0830
Parametric estimation, Clayton copula					*Semiparametric estimation, Clayton copula*				
True values	0.8000	1.1000	0.5000	0.3643	True values	0.8000	1.1000	0.5000	0.3643
Estimate	0.8024	1.1083	0.5075	0.3598	Estimate	0.8027	1.1275	0.5140	0.3504
SD	0.0942	0.3371	0.1368	0.0791	SD	0.0935	0.3719	0.1354	0.0816
Bias	0.0024	0.0083	0.0075	-0.0045	Bias	0.0027	0.0275	0.0139	-0.0139
RMSE	0.0942	0.3372	0.1370	0.0792	RMSE	0.0936	0.3729	0.1361	0.0828
Parametric estimation, Gumbel copula					*Semiparametric estimation, Gumbel copula*				
True values	0.8000	1.1000	0.5000	0.3643	True values	0.8000	1.1000	0.5000	0.3643
Estimate	0.8026	1.1339	0.5060	0.3605	Estimate	0.8035	1.1564	0.5102	0.3562
SD	0.0974	0.4002	0.1488	0.0894	SD	0.0994	0.4300	0.1535	0.0978
Bias	0.0026	0.0339	0.0060	-0.0038	Bias	0.0035	0.0564	0.0102	-0.0081
RMSE	0.0974	0.4016	0.1489	0.0895	RMSE	0.0995	0.4337	0.1539	0.0981

parameter ψ_0 and the ATE. We compare them with the performance of the parametric ML estimators under various DGPs and model specifications, and illustrate how the parametric estimators of ψ_0 and the ATE suffer from misspecification of the marginal distribution of ϵ. Note that the ATE involves ψ_0 and the marginal of ϵ.

5.1 | Simulation design

We compare the performance of the parametric and semiparametric estimators when the marginal distributions are misspecified in the parametric models. To calculate the parametric estimators, we specify the parametric models with normal distributions for the marginals of ϵ and ν, owing to their popularity. For the DGPs, we consider two marginals of ϵ and ν: the standard normal distribution (to reflect correct specification) and a mixture of normal distributions (to reflect misspecification).

The DGPs are as follows:

$$Y_i = \mathbf{1}\{X_i\beta + D_i\delta_1 \geq \epsilon\}, \quad D_i = \mathbf{1}\{X_i\alpha + Z_i\gamma \geq \nu\},$$

where

$$(\alpha, \gamma, \beta, \delta_1) = (-1, 0.8, -1, 1.1), (X, Z)' \sim \mathcal{N}\left((0,0)', \begin{pmatrix} 1 & -0.1 \\ -0.1 & 1 \end{pmatrix}\right)$$

and $(\epsilon, \nu)' \sim C(F_{\epsilon 0}(\cdot), F_{\nu 0}(\cdot); \rho)$. Here, $F_{\epsilon 0}$ and $F_{\nu 0}$ are normal or a mixture of normal.[15] For $C(\cdot, \cdot; \rho)$, we consider the Gaussian, Frank, Clayton, and Gumbel copulas, which satisfy the identifying assumption (Assumption 5). The dependence structure between ϵ and ν is characterized by a one-dimensional parameter ρ in all copulas considered, but the interpretation of the dependence parameter differs across the copulas. To resolve this issue, we report Spearman's ρ corresponding to the estimated dependence parameter in each copula specification. We estimate the models with several values of ρ to examine whether the performance of the estimators varies with the degree of dependence. Although we assume that the copula is correctly specified, economic theory does not provide a justification for the choice of copula. In this simulation study, we also examine the effect of copula misspecification on the performance of the estimators.[16]

[15] For the mixture of normal distributions, ϵ and ν are generated from $0.6\mathcal{N}(-1, \sigma^2) + 0.4\mathcal{N}(1.5, \sigma^2)$ for appropriate $\sigma > 0$, so that the mean is zero and the variance is one.

[16] Misspecification problems in copula-based models have been documented using Monte Carlo simulations in the statistics literature (e.g., Kim, Sivapulle, & Silvapulle, 2007a, 2007b; Lawless & Yilmaz, 2011). In particular, Lawless and Yilmaz (2011) compared the performance of the parametric and

TABLE 2 Misspecification of marginals ($n = 500$) (true marginal: mixture of normals)

	γ	δ_1	ρ_{sp}	ATE		γ	δ_1	ρ_{sp}	ATE
Parametric estimation, Gaussian copula					*Semiparametric estimation, Gaussian copula*				
True values	0.8000	1.1000	0.5000	0.1066	True values	0.8000	1.1000	0.5000	0.1066
Estimate	0.7994	1.0925	0.4496	0.2443	Estimate	0.8562	1.2696	0.4895	0.1241
SD	0.1281	0.6285	0.1651	0.1129	SD	0.1113	0.3728	0.1059	0.0653
Bias	-0.0006	-0.0075	-0.0504	0.1377	Bias	0.0562	0.1696	-0.0105	0.0174
RMSE	0.1281	0.6285	0.1726	0.1780	RMSE	0.1247	0.4096	0.1064	0.0675
Parametric estimation, Frank copula					*Semiparametric estimation, Frank copula*				
True values	0.8000	1.1000	0.5000	0.1066	True values	0.8000	1.1000	0.5000	0.1066
Estimate	0.8056	1.3088	0.3976	0.2894	Estimate	0.8377	1.2541	0.4829	0.1276
SD	0.1272	0.5093	0.1221	0.0883	SD	0.1141	0.3564	0.0963	0.0689
Bias	0.0056	0.2088	-0.1024	0.1827	Bias	0.0377	0.1541	-0.0171	0.0210
RMSE	0.1273	0.5504	0.1594	0.2030	RMSE	0.1202	0.3883	0.0978	0.0720
Parametric estimation, Clayton copula					*Semiparametric estimation, Clayton copula*				
True values	0.8000	1.1000	0.5000	0.1066	True values	0.8000	1.1000	0.5000	0.1066
Estimate	0.8099	1.1439	0.4236	0.2555	Estimate	0.8441	1.2234	0.4948	0.1192
SD	0.1309	0.5236	0.1412	0.0913	S.D	0.1134	0.3611	0.0999	0.0611
Bias	0.0099	0.0439	-0.0764	0.1488	Bias	0.0441	0.1234	-0.0053	0.0126
RMSE	0.1312	0.5254	0.1605	0.1746	RMSE	0.1217	0.3816	0.1001	0.0624
Parametric estimation, Gumbel copula					*Semiparametric estimation, Gumbel copula*				
True values	0.8000	1.1000	0.5000	0.1066	True values	0.8000	1.1000	0.5000	0.1066
Estimate	0.7892	1.0326	0.4650	0.2373	Estimate	0.8484	1.2692	0.4900	0.1259
SD	0.1333	0.5297	0.1338	0.0986	S.D	0.1142	0.3646	0.0986	0.0645
Bias	-0.0108	-0.0674	-0.0350	0.1307	Bias	0.0484	0.1692	-0.0099	0.0193
RMSE	0.1337	0.5340	0.1383	0.1637	RMSE	0.1241	0.4019	0.0991	0.0673

We impose a restriction that X has no constant for the location normalization, and fix α and β to -1 for the scale normalization. We use these normalizations in both parametric and semiparametric models, and it allows us to easily compare the performance of the parametric and semiparametric estimators. We consider two sample sizes—500 and 1,000—and all results are obtained from 2,000 Monte Carlo replications. As a performance measure of the estimators, we consider the root mean squared errors (RMSEs) in our simulation.

5.2 | Estimation of parametric and semiparametric models

The parametric models can be estimated by the standard ML method. Since bivariate probit models are commonly used in practice, we specify the model using the Gaussian copula and normal marginals. In addition to that, we also try different copulas and normal marginals.[17]

Consider semiparametric models. Recall that we assume that $\sqrt{h_j} \in \Lambda^p([0,1])$. Therefore, for each $j \in \{\epsilon, v\}$, we approximate h_j to

$$h_j(x) = \frac{\left(\sum_{k=0}^{k_{nj}} a_{jk}\psi_{jk}(x)\right)^2}{\int_0^1 \left(\sum_{k=0}^{k_{nj}} a_{jk}\psi_{jk}(x)\right)^2 dx}, \tag{20}$$

where $\{\psi_{jk}(\cdot)\}_{k=0}^{k_{nj}}$ is the set of approximating functions for $h_j(\cdot)$, and k_{nj} is the number of approximating functions. The approximation in Equation (20) guarantees that $\int_0^1 h_j(x)dx = 1$ by construction. We take the space of the polynomials as the sieve space for h_ϵ and h_v. The orders of the polynomials ($k_{n\epsilon}$ and k_{nv}) are set to be proportional to $n^{1/7}$. To incorporate the specification given in Equation (15), we choose the standard normal distribution function for G.

semiparametric ML estimators in a copula-based model and showed that the semiparametric two-step method outperformed the parametric estimation method when the copula function was misspecified.

[17]Such an estimation method in related parametric models can be found in Marra and Radice (2011). The R package (GJRM) used in their paper can be used to estimate our parametric model as well.

TABLE 3 Summary statistics

	Variable	Mean	SD	Min.	Max.
Y	Whether or not visit doctor	0.182	0.386	0	1
D	Whether or not have insurance	0.657	0.475	0	1
M	Age	42.591	10.574	25	64
	Years of education	13.433	2.892	0	17
	Income (hourly)	20.094	11.990	0.4	73.08
	Family size	2.932	1.595	1	14
	Living in MSA	0.868	0.338	0	1
	Male	0.500	0.500	0	1
	Region: NorthEast	0.141	0.348	0	1
	Region: MidWest	0.226	0.418	0	1
	Region: South	0.369	0.483	0	1
	Region: West	0.264	0.441	0	1
	Race: White	0.739	0.439	0	1
	Race: Black	0.170	0.376	0	1
	Race: Minority	0.010	0.099	0	1
	Race: Asian	0.081	0.273	0	1
	Ever married	0.782	0.413	0	1
	Physical health below Good†	0.095	0.293	0	1
	Mental health below Good†	0.036	0.186	0	1
Z	Number of employees	149.385	182.662	1	500
	Firm has multiple locations	0.682	0.466	0	1
X	sick 32	68.317	17.402	42	91
	sick 34	70.463	3.633	67	77

Number of observations = 7,555

† The original variables for these variables are coded into five groups: Excellent, Very Good, Good, Fair, and Poor. These variables show the proportion of individuals in the sample that consider their physical/mental health is below Good (i.e., Fair or Poor).

5.3 | Simulation results

We begin by examining the simulation results under correct specification (i.e., the true marginal distributions and the specified marginal distributions are both normal). Table 1 shows the simulation results for $n = 500$. We find that the ML estimators of ψ and the ATE perform well in the parametric models, with negligible biases and small variances.[18] The performance of the sieve ML estimators of ψ and the ATE in the semiparametric models is as good as that in the parametric models, even with this moderate sample size.

Now, we consider the cases where the marginal distributions are misspecified in the parametric models. Table 2 considers the case where the true marginal distributions are a mixture of normal distributions, but the researcher specifies them as normal distributions. In this table, the RMSEs of the parametric ML estimators are larger than those of the sieve ML estimators. This implies that the parametric ML estimators suffer from misspecification whereas the sieve ML estimators do not. Moreover, the parametric estimators of the ATE are substantially distorted under this misspecification, presumably because the ATE is a function of the misspecified distribution of ϵ. Note that the poor performance of the parametric estimators is attributed not only to large bias but also large variance. For instance, the bias of the parametric estimator of the ATE with the Gaussian copula is 0.1377, which is about eight times larger than that of the corresponding sieve estimator. These biases of the parametric estimators of the ATE are substantial in that they do not disappear with the increased sample size.[19] Therefore, the simulation results demonstrate that when the marginal distributions are misspecified the sieve estimators outperform the parametric estimators in terms of the RMSE. The Supporting Information Appendix also contains simulation results for the cases where both the copula and the marginal distributions are misspecified. The results show that, even under copula misspecification, the sieve ML estimators remain to outperform the parametric counterparts when the marginal distributions are misspecified.

Overall, the simulation results suggest that researchers are recommended to use the semiparametric models and the sieve ML estimation proposed in this paper when they are concerned about model misspecification. The following is a summary of the main findings from our simulation study:

[18]The ATE is evaluated at the mean of X.
[19]We provide simulation results with a larger sample size ($n = 1,000$), and they can be found in the Supporting Information Appendix.

TABLE 4 Estimates in selection equation

	Parametric	Semiparametric
Age†	0.130***	0.077***
	(0.018)	(0.038)
Years of education†	0.190***	0.098**
	(0.018)	(0.044)
Family size†	-0.120***	-0.041*
	(0.017)	(0.023)
Income†	0.268***	0.416***
	(0.028)	(0.089)
Male	0.193***	0.062
	(0.036)	(0.039)
Living in MSA	-0.090*	-0.040
	(0.047)	(0.056)
Ever married	-0.112***	-0.048
	(0.043)	(0.050)
Physical health very good	0.001	-0.024
	(0.050)	(0.042)
Physical health good	0.009	-0.011
	(0.053)	(0.043)
Physical health fair	-0.097	-0.066
	(0.071)	(0.060)
Physical health poor	0.080	0.039
	(0.155)	(0.126)
Mental health very good	0.004	-0.016
	(0.043)	(0.043)
Mental health good	-0.031	-0.029
	(0.049)	(0.038)
Mental health fair	-0.009	-0.041
	(0.095)	(0.067)
Mental health poor	0.135	0.113
	(0.287)	(0.399)
Days for sick leave† (T32)	0.119***	0.094***
	(0.020)	(0.025)
Days for sick leave† (T34)	0.113***	0.113
	(0.019)	(N/A)
Number of employees (Z_1)	0.228***	0.231**
	(0.020)	(0.116)
Firm has multiple locations (Z_2)	0.374***	0.173***
	(0.034)	(0.067)
Region and race dummies	Yes	Yes
Number of observations	7,555	7,555

Note. Standard errors in parentheses. *$p < 0.10$; **$p < 0.05$; ***$p < 0.01$. The coefficient on T34 in the semiparametric model is fixed for normalization. Gaussian copula is used. † indicates that the variable is standardized.

(i) When the model is correctly specified, the performance of the sieve ML estimators is comparable to that of the parametric ML estimators.
(ii) When the marginal distributions are misspecified, the sieve ML estimation is recommended in order to improve the performance.
(iii) The semiparametric ML estimators performs better than the parametric ML estimators under both copula and marginal misspecification. Therefore, the semiparametric models are preferred to the parametric models in such cases.
(iv) Especially for the ATE, whenever the marginal distributions are misspecified, the parametric ML estimates can be significantly distorted.

We provide additional simulation results in the Supporting Information Appendix, where we consider (a) a larger sample size, (b) both copula and marginal misspecification, (c) different degrees of dependence, (d) marginal density functions

of heavy tails, and (e) the coverage probabilities of bootstrap confidence intervals. Here is a summary. Across various simulation designs (a–c), our main findings remain the same. When the marginal distributions are believed to have fat tails, we recommend practitioners to use the transformation function G that has fat tails. Lastly, the percentile bootstrap works well with the coverage probabilities close to its nominal level.

6 | EMPIRICAL EXAMPLE

In this section, we illustrate in an application the practical relevance of the theoretical results developed in this paper. It is widely recognized that health insurance coverage can be an important factor for patients' decisions for making medical visits. At the same time, having insurances is endogenously determined by an individual's health status and socioeconomic characteristics. In our empirical application, we analyze how health insurance coverage affects an individual's decision to visit a doctor. In this example, Y is a binary outcome variable indicating whether an individual visited a doctor's office, and D is the endogenous treatment variable that indicates whether an individual has his or her own private insurance.

We use the 2010 wave of the Medical Expenditure Panel Survey (MEPS) as our main data source. We focus on all the visits that occurred in January, 2010. We restrict the sample to contain individuals with age between 25 and 64, and exclude individuals who have retained any kinds of federal or state insurance in 2010. For Z, we consider two instrumental variables that are used in Zimmer (2018)—the number of employees in the firm at which the individual works and a dummy variable that indicates whether a firm has multiple locations. These variables reflect how big the firm is, and the underlying rationale for using these variables as instruments is as follows: The bigger the firm is, the more likely it provides fringe benefits including health insurance. Therefore, it is likely that these instruments affect insurance status. We can argue, however, that they do not have direct effects on decisions to visit doctors.[20] We assume that these variables are exogenous conditional on covariates. For additional covariates M, we include age, gender, years of education, family size (the number of family members), income, region, race, marital status, subjective physical and mental health status evaluations, and whether living in a metropolitan statistical area. For the exogenous variable X in our model, we use information about the provision of paid sick leave, which is separately collected from the National Compensation Survey published by the US Bureau of Labor Statistics. We match the information for various industries with the primary data set we use. Conditional on the covariates listed above, we assume that the number of sick leave days and leave benefits are exogenous, by the same argument as for the instruments. Since X and Z are assumed to be exogenous only conditional on M, we rely on Assumption $1'$ instead of Assumption 1 for identification.

Since we include various control variables, one concern may be that the resulting estimators are imprecise with a moderate sample size. It is worth emphasizing, however, that our semiparametric estimators do not suffer from the curse of dimensionality as theoretically shown in Section 4. This is because of the parametric index structure in our model. Moreover, we do not attempt to estimate the distributions of the unobservables conditional on these covariates, but only estimate the marginal distributions.

Table 3 summarizes the variables used in estimation and shows their summary statistics. While 65.7% of individuals had private health insurances in January 2010, only 18.2% of them visited doctors during the period. We use two variables for the pay sick leave provision (i.e., X)—within each industry, the percentage of workers who are provided with paid sick leave benefits and the percentage of workers who are provided with a fixed number of days for sick leave per year. The summary statistics for these two variables show that there are sufficient variations across individuals in different industries. Note that all the continuous variables are standardized in order to ensure stability in estimation.[21]

Before estimating the parametric and semiparametric models, we run a first-stage ordinary least squares regression of D on X, M, and Z to see if the excluded instruments are weak. The F-statistic value is 167.19, and thus we assume that the instruments are strong.[22] For the normalization of the parametric model, we use the convention $E[\epsilon] = E[\nu] = 0$ and $\text{var}(\epsilon) = \text{var}(\nu) = 1$. On the other hand, for the semiparametric model, we impose the normalization used in our simulation studies; that is, we exclude the constant terms and the coefficients on *sick 34* are fixed to be corresponding parametric estimates. We choose the Gaussian copula to capture the dependence structure between ϵ and ν. In both models, the stan-

[20]Note that it is difficult to justify these instruments for individuals who are either self-employed or unemployed. To avoid this issue, we exclude those individuals from our analysis.

[21]That is, for a continuous random variable X, define $\tilde{X} = \frac{X - \bar{X}_n}{\hat{sd}(X)}$, where $\bar{X}_n$ and $\hat{sd}(X)$ are the sample average and standard deviation of X, respectively.

[22]The F-statistic in the first-stage linear regression may not be the best indicator for detecting weak instruments in nonlinear models. Han and McCloskey (2019) developed inference methods that were robust to weak identification for a class of nonlinear models, and considered bivariate probit models as one of the leading examples.

TABLE 5 Estimates in outcome equation

	Parametric	Semiparametric
Treatment (δ)	0.493***	0.368**
	(0.168)	(0.183)
Age†	0.055***	0.059
	(0.020)	(0.047)
Years of education†	0.142***	0.126*
	(0.028)	(0.066)
Family size†	-0.055***	-0.052*
	(0.021)	(0.030)
Income†	0.018	0.031
	(0.026)	(0.068)
Male	-0.398***	-0.373**
	(0.037)	(0.169)
Living in MSA	0.063	0.040
	(0.052)	(0.061)
Ever married	0.188***	0.179**
	(0.049)	(0.084)
Physical health very good	0.227***	0.201**
	(0.056)	(0.084)
Physical health good	0.395***	0.356***
	(0.059)	(0.130)
Physical health fair	0.691***	0.644***
	(0.077)	(0.224)
Physical health poor	0.978***	0.959*
	(0.163)	(0.492)
Mental health very good	-0.033	-0.040
	(0.048)	(0.057)
Mental health good	-0.066	-0.064
	(0.053)	(0.064)
Mental health fair	0.042	0.053
	(0.105)	(0.154)
Mental health poor	0.300	0.186
	(0.297)	(0.320)
Days for sick leave† (T32)	-0.026	-0.023
	(0.026)	(0.027)
Days for sick leave† (T34)	-0.049**	-0.049
	(0.025)	(N/A)
Region and Race dummies	Yes	Yes
Number of observations	7,555	7,555

Note. Standard errors in parentheses. *$p < 0.10$; **$p < 0.05$; ***$p < 0.01$. The coefficient on T34 in the semiparametric model is fixed for normalization. Gaussian copula is used. † indicates that the variable is standardized.

dard errors are obtained by the bootstrap procedure (Section 4.4), where the bootstrap weights are generated from the exponential distribution with the parameter value 1.

Tables 4 and 5 present the estimation results for the selection equation and the outcome equation, respectively. Between the parametric and semiparametric models, the magnitude and significance of the estimates differs, although, overall, the signs of the estimates are similar. Table 6 shows the ATE estimates evaluated at various values of X and M, as well as the estimates of the copula parameter ρ. The parametric estimate of ρ is statistically significant under the 5% level, whereas the semiparametric estimate is not. We can find that the parametric estimates of the ATE are different from the corresponding semiparametric estimates. For example, the parametric ATE estimate evaluated at the 50% quantile of $(X', M')'$ is about 0.129, which means that having private insurance increases the probability of visiting doctors by 12.9%. On the other hand, the corresponding semiparametric estimate shows that the effect is 10.4%. The discrepancy in the ATE estimates between the parametric and semiparametric models suggests the possible misspecification of the marginals, which is consistent with the premise of this paper.

TABLE 6 Estimated ATEs and Spearman's ρ

	Parametric	Semiparametric
ATE at the mean	0.114***	0.100**
	(0.037)	(0.048)
ATE at 50% quantile	0.129***	0.104*
	(0.045)	(0.054)
ATE at 25% quantile	0.121**	0.104
	(0.050)	(0.058)
ATE at 75% quantile	0.139***	0.105*
	(0.043)	(0.056)
Spearman's ρ	-0.200**	-0.154
	(0.105)	(0.134)
Number of observations	7,555	7,555

Note. Standard errors in parentheses. *$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

7 | CONCLUSIONS

In this paper, we propose semiparametric estimation and inference methods for generalized bivariate probit models. Specifically, we develop the asymptotic theory for the sieve ML estimators of semiparametric copula-based triangular systems with binary endogenous variables. We show that the sieve ML estimators are consistent and that their smooth functionals are $\sqrt{n}$ -asymptotically normal under some regularity conditions. This semiparametric estimation approach allows for flexibility in the models and thus provides robustness in estimation and inference.

We conduct a sensitivity analysis to examine how sensitive the estimation results are to model specifications. The results show that, overall, the semiparametric sieve ML estimators perform well in terms of both bias and variance. When the marginal distributions are misspecified, the sieve ML estimators substantially outperform the parametric ML estimators and the latter exhibit substantial bias. In particular, we find that the parametric estimates of the parameters involving the misspecified marginal distributions, such as the ATE, are highly misleading. When the model is correctly specified, we find that the performance of the sieve ML estimators is comparable to that of the parametric ones. When the copula is also misspecified, the distortion of the parametric estimates under misspecification of the marginals can become even more severe, whereas the semiparametric estimates do not seem to be affected by this misspecification as long as the copula of the true DGP is within the stochastic ordering class. A related and interesting question is how the results would change if the data are not generated from this class of copulas.

We also formally show that the exclusion restriction is not only sufficient, but is also necessary for identification. Without the exclusion restriction, the model parameters are not identified or, under the normality assumption, are, at best, weakly identified. Some empirical studies ignore the exclusion restriction when estimating the model, and our nonidentification result provides a caveat for practitioners.

ACKNOWLEDGMENTS

The authors thank Jason Abrevaya, Xiaohong Chen, Stephen Donald, Brendan Kline, Ed Vytlacil, and Haiqing Xu for valuable discussions and helpful comments. An earlier version of this paper has been circulated under the title "Sensitivity analysis in triangular systems of equations with binary endogenous variables."

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [http://qed.econ.queensu.ca/jae/datasets/han001/].

REFERENCES

Ai, C., & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, *71*(6), 1795–1843.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human Resources*, *40*(4), 791–821.

Bhattacharya, J., Goldman, D., & McCaffrey, D. (2006). Estimating probit models with self- selected treatments. *Statistics in Medicine*, *25*(3), 389–413.

Bhattacharya, J., Shaikh, A. M., & Vytlacil, E. (2008). Treatment effect bounds under monotonicity assumptions: An application to Swan–Ganz catheterization. *American Economic Review*, *98*(2), 351–356.

Bierens, H. J. (2008). Semi-nonparametric interval-censored mixed proportional hazard models: Identification and consistency results. *Econometric Theory*, *24*(3), 749–794.

Bierens, H. J. (2014). Consistency and asymptotic normality of sieve ML estimators under low-level conditions. *Econometric Theory*, *30*(5), 1021–1076.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman, & E. Leamer (Eds.), *Handbook of Econometrics*, Vol. *6B*. Amsterdam, Netherlands: Elsevier, pp. 5549–5632.

Chen, X., & Fan, Y. (2006a). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics*, *135*(1), 125–154.

Chen, X., & Fan, Y. (2006b). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, *130*(2), 307–335.

Chen, X., Fan, Y., & Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, *101*(475), 1228–1240.

Chen, X., Hu, Y., & Lewbel, A. (2009). Nonparametric identification and estimation of non- classical errors-in-variables models without additional information. *Statistica Sinica*, *19*(3), 949–968.

Chen, X., & Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, *152*(1), 46–60.

Cheng, G., & Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Annals of Statistics*, *38*(5), 2884–2915.

Chiburis, R. (2010). Semiparametric bounds on treatment effects. *Journal of Econometrics*, *159*(2), 267–275.

Evans, W. N., & Schwab, R. M. (1995). Finishing high school and starting college: Do Catholic schools make a difference? *Quarterly Journal of Economics*, *110*(4), 941–974.

Freyberger, J., & Masten, M. (2019). A practical guide to compact infinite dimensional parameter spaces. *Econometric Reviews*. *38*(9), 979–1006.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*(2), 363–390.

Goldman, D., Bhattacharya, J., Mccaffrey, D., Duan, N., Leibowitz, A., Joyce, G., & Morton, S. (2001). Effect of insurance on mortality in an HIV-positive population in care. *Journal of the American Statistical Association*, *96*(455), 883–894.

Han, S., & McCloskey, A. (2019). Estimation and inference with a (nearly) singular Jacobian. *Quantitative Economics*. *10*(3), 1019–1068.

Han, S., & Vytlacil, E. (2017). Identification in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Econometrics*, *199*(1), 63–73.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–162.

Hu, Y., & Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, *76*(1), 195–216.

Ieva, F., Marra, G., Paganoni, A. M., & Radice, R. (2014). A semiparametric bivariate probit model for joint modeling of outcomes in STEMI patients. *Computational and Mathematical Methods in Medicine*, *2014*, article no. 240435.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability. London, UK: Taylor and Francis.

Kim, G., Silvapulle, M. J., & Silvapulle, P. (2007a). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, *51*(6), 2836–2850.

Kim, G., Silvapulle, M. J., & Silvapulle, P. (2007b). Semiparametric estimation of the error distribution in multivariate regression using copulas. *Australian and New Zealand Journal of Statistics*, *49*(3), 321–336.

Lawless, J. F., & Yilmaz, Y. E. (2011). Comparison of semiparametric maximum likelihood estimation and two-stage semiparametric estimation in copula models. *Computational Statistics and Data Analysis*, *55*(7), 2446–2455.

Liao, Z., & Shi, X. (2017). A uniform model selection test for semi/nonparametric models. (*Working Paper*). Madison, WI: Department of Economics, University of Wisconsin–Madison.

Lorentz, G. (1966). *Approximation of functions*. New York, NY: Holt, Rinehart & Winston.

Ma, S., & Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, *96*(1), 190–217.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *The American Economic Review*, *80*(2), 319–323.

Marra, G., & Radice, R. (2011). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics*, *39*(2), 259–279.

Mourifié, I. (2015). Sharp bounds on treatment effects in a binary triangular system. *Journal of Econometrics*, *187*(1), 74–81.

Mourifié, I., & Méango, R. (2014). A note on the identification in two equations probit model with dummy endogenous regressor. *Economics Letters*, *125*(3), 360–363.

Neal, D. A. (1997). The effects of Catholic secondary schooling on educational achievement. *Journal of Labor Economics*, *15*(1), 98–123.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Berlin, Germany: Springer.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle, & D. L. McFadden (Eds.), *Handbook of Econometrics*, Vol. *4*. Berlin, Germany: Elsevier, pp. 2111–2245.

Rhine, S. L., Greene, W. H., & Toussaint-Comeau, M. (2006). The importance of check-cashing businesses to the unbanked: Racial/ethnic differences. *Review of Economics and Statistics*, *88*(1), 146–157.

Shaikh, A. M., & Vytlacil, E. J. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica*, *79*(3), 949–955.

Vella, F. (1998). Models with sample selection bias: A survey. *Journal of Human Resources*, *33*(1), 127–169.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307–333.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

White, N. E., & Wolaver, A. M. (2003). Occupation choice, information, and migration. *Review of Regional Studies*, *33*(2), 142–163.

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters*, *69*(3), 309–312.

Zimmer, D. (2018). Using copulas to estimate the coefficient of a binary endogenous regressor in a Poisson regression: Application to the effect of insurance on doctor visits. *Health Economics*, *27*(3), 545–556.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Han S, Lee S. Estimation in a generalization of bivariate probit models with dummy endogenous regressors. *J Appl Econ*. 2019;1–22. https://doi.org/10.1002/jae.2727