Sieve Maximum Likelihood Estimation of a Copula-Based Sample Selection Model

Jörg Schwiebert*

April 30, 2013

Abstract

We propose a sieve maximum likelihood estimation procedure for a copula-based sample selection model. That is, we impose a parametric copula assumption on the joint distribution function of error terms, but estimate the marginal distributions semiparametrically by the method of sieves. Compared with the semi-nonparametric maximum likelihood estimation procedure due to Gallant and Nychka (1987), our approach allows to incorporate parametric prior information into the estimation process, is less computationally demanding and makes it easy to test for the validity of parametric assumptions. Furthermore, we provide asymptotic distribution results for our estimator. Our maximum likelihood approach to estimating a sample selection model has advantages over least squares based techniques proposed in the literature if the distribution of the error terms of main and selection equation is of interest.

Keywords: Sample selection model, semiparametric estimation, copula, method of sieves, maximum likelihood estimation.
JEL codes: C21, C24.

*Leibniz University Hannover, Institute of Labor Economics, Königsworther Platz 1, 30167 Hannover, Germany, phone: +49 (0) 511/762-5657, e-mail: schwiebert@aoek.uni-hannover.de

1 Introduction

The sample selection model has become the standard econometric tool when dealing with sample selectivity. The model typically consists of a main equation (of interest) and a selection equation, where the latter determines the probability of being in the selected sample. If sample selectivity is present, ordinary least squares estimation of the main equation is likely to produce inconsistent estimates because the selected sample is a nonrandom sample from the overall population. Heckman (1979) showed that the sample selection problem can be interpreted as an omitted variable bias problem. He demonstrated that ordinary least squares estimation of the main equation including a selectivity correction term (known as the inverse Mills ratio) leads to consistent estimates of the parameters of interest. Besides estimating the model by ordinary (or weighted) least squares techniques, it is also possible to estimate the model by maximum likelihood.

Gallant and Nychka (1987) have proposed a semi-nonparametric maximum likelihood estimator for estimating the sample selection model. The virtue of their approach is that it is not necessary to assume a parametric (joint) distribution for the error terms of the underlying econometric model. Consequently, consistent estimates of the parameters of interest can be obtained under weak conditions. This is an important advantage over the model proposed by Heckman (1979) who assumed a bivariate normal distribution for the error terms of main and selection equation.

In this paper, we propose a sieve maximum likelihood estimator for the sample selection model. We make the crucial assumption that the joint distribution of the error terms of main and selection equation can be characterized by a specific copula, but we estimate the marginal distributions semiparametrically by the method of sieves along with the structural parameters of interest. Our estimation concept is thus sieve maximum likelihood estimation (Chen, 2007).

Our modeling and estimation approach has several advantages over the Gallant and Nychka (1987) procedure. First, our approach allows to incorporate prior information on the distribution of error terms into the estimation process. For example, the joint distribution of error terms may be characterized by fat tails, hence a Student t copula

may be an appropriate modeling choice (Heckman, 2003). Furthermore, the selection equation may reasonably be estimated by probit or logit, hence the marginal distribution of the selection equation's error term is normal or logistic, respectively. Since a copula couples two marginal distributions into a joint distribution, such prior information on the joint or marginal distributions can be easily incorporated into our econometric model. This is not possible in the Gallant and Nychka (1987) approach, who estimate the entire joint density function of error terms semi-nonparametrically by a series expansion.

Second, our method is less computationally demanding than the Gallant and Nychka (1987) procedure. In Gallant and Nychka (1987) a two-dimensional density function is approximated semi-nonparametrically by a series expansion (where the number of series term grows with the sample size). The coefficients of the series expansion are then estimated along with the parameters of interest. However, the approximation of a two-dimensional density function requires a considerable number of series terms, which leads to a computationally demanding estimation process. Our approach, on the other hand, requires only the approximation of the one-dimensional marginal distributions, which is far easier than approximating a (bivariate) joint distribution.[1]

Third, Gallant and Nychka (1987) have proved the consistency of their estimator, but no (asymptotic) distribution results have been provided. Yet, such distribution results are necessary for hypotheses testing and obtaining confidence intervals. Of course, one could obtain estimates under the assumption that the number of series terms is fixed rather than increasing with the sample size; in that case, distribution results would follow from standard (parametric) maximum likelihood theory. However, this procedure is in general not justified due to the semiparametric nature of the estimation problem. Concerning our proposed method, conditions under which a sieve maximum likelihood estimator is consistent and asymptotically normally distributed have been provided by Chen et al. (2006) and Chen (2007). As will be shown below, under suitable assumption these conditions are fulfilled in case of our estimator, hence we are able to provide distribution results.

Fourth, our approach offers an easy way to test for the validity of parametric assump-

[1]The same argument has been used by Chen et al. (2006).

tions. Incorporating correct parametric prior information into an econometric model is desirable since this typically leads to efficiency gains. However, prior information may not be correct, hence it is important to test for the validity of such assumptions. Our copula framework provides an easy way to do so because one can separately test for the validity of the assumed copula and for the validity of the assumed marginal distributions. Details are given in Section 4.

Besides Gallant and Nychka (1987), several other authors have developed semi-nonparametric estimators for the sample selection model which do not rely on strong parametric assumptions. Examples include Powell (1987), Ahn and Powell (1993), Das et al. (2003) and Newey (2009). These authors propose least-squares based estimation procedures to consistently estimate the structural parameters of the main equation. These estimation procedures are typically two-step. In a first step, the selection equation is estimated by some semi-nonparametric technique. As in case of the model with normally distributed error terms, one augments the main equation with a selectivity correction term (a generalization of the inverse Mills ratio term). Then one either gets rid of the selectivity correction term by differencing out (Powell, 1987; Ahn and Powell, 1993), or approximates the term by, e.g., a series expansion (Das et al., 2003; Newey, 2009). In a second step, estimation of the main equation is carried out by some variant of ordinary or weighted least squares.

Our and the Gallant and Nychka (1987) approach differ from these least-squares based techniques in two important ways. First, our and the Gallant and Nychka (1987) approach are one-step. This facilitates the computation of standard errors and confidence intervals (in case of our estimator) because one does not have to adjust for the uncertainty associated with the first-step estimation. Second, our and the Gallant and Nychka (1987) approach are not based on least-squares but maximum likelihood estimation. This *requires* a specification of the joint distribution of error terms of main and selection equation. Considered conversely, a *virtue* of our and the Gallant and Nychka (1987) approach is that they also provide information on the joint distribution of the error terms.

Information on the joint distribution of error terms is useful for a couple of reasons. First, sample selectivity is a problem only if the error terms are dependent. Distributional

information helps to identify these dependencies, and thus reveals how the sample selection mechanism works. Second, from the joint distribution one can derive the marginal distributions of error terms. For instance, if the main equation is a wage equation, an object of interest might be if wage densities are fat-tailed (Heckman and Sedlacek, 1990). Third, the joint distribution is interesting because treatment parameters depend on the tail behavior of error terms (Heckman et al., 2003).

A drawback of our proposed approach might be that it is necessary to specify a parametric copula for the joint distribution of error terms in advance. However, if one has prior information (e.g., from economic theory or empirical regularities) on the features of the joint distribution (such as fat tails), then the copula framework provides a very flexible environment to include such prior information into the econometric model. Chen et al. (2006) also estimate a copula model with unknown marginal distributions and note that "this class of semiparametric multivariate distributions is able to jointly model any type of dependence with any types of marginal behaviors and has proven useful in diverse fields" (Chen et al., 2006, p. 1228). Hence, our approach exhibits the same flexibility as the semiparametric approach of Gallant and Nychka (1987), but may be preferred due to the reasons given above.

The remainder of the paper is organized as follows. In Section 2 we provide the model and our proposed estimation strategy. In Section 3 we derive the asymptotic properties of our proposed estimator. Section 4 contains remarks and extensions concerning different aspects of estimation, testing, and model specification. Finally, Section 5 concludes the paper.

2 Model Setup and Estimation

We consider an ordinary sample selection model given by

$$y_i^* = x_i'\beta + \varepsilon_i \tag{1}$$

$$d_i^* = w_i'\gamma + u_i \tag{2}$$

$$d_i = 1(d_i^* > 0) \tag{3}$$

$$y_i = \begin{cases} y_i^* \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases}, \tag{4}$$

where $i = 1, \ldots, n$ indexes individuals. The first equation is the main equation, where y^* is a latent outcome variable, x is a vector of (exogenous) explanatory variables with corresponding parameter vector β and ε denotes the error term. The second equation is the selection equation, where d^* is the latent dependent variable, w is a vector of (exogenous) explanatory variables with corresponding parameter vector γ and u denotes the error term. The last two equations comprise the selection mechanism. The latent variable y^* can only be observed if $d^* > 0$, or, equivalently, if the selection indicator d is equal to one.

We make the following assumptions:

Assumption 1: *$\{(x_i, w_i, \varepsilon_i, u_i)\}_{i=1}^n$ are i.i.d. from some underlying distribution.*

Assumption 2: *The joint distribution function of ε and u is given by $H_{\varepsilon,u}(a,b) = C(F_\varepsilon(a), F_u(b); \tau)$, where $C : [0,1]^2 \to [0,1]$ is a known copula with dependence parameter τ, and F_ε and F_u denote the marginal distribution functions of ε and u. Furthermore, the marginal density functions f_ε and f_u are absolutely continuous with respect to Lebesgue measure.*

Assumption 3: *(x, w) and (ε, u) are independent.*

Assumption 4: *(i) x and w do not contain a constant term. (ii) w contains at least one variable (with a nonzero coefficient) which is not included in x. (iii) The first element of γ is equal to one in absolute value.*

Assumptions 1 and 2 imply that our model can be estimated by maximum likelihood.

Assumptions 3 and 4 are basic conditions for identification. Assumption 4 (ii) is a well-known exclusion restriction which is a standard condition for identification of sample selection models. Assumption 4 (iii) is a scale normalization for the parameters of the selection equation, since these are only identified up to scale.

The joint probability density function (p.d.f.) of ε and u is given by

$$h_{\varepsilon,u}(a,b) = c(F_\varepsilon(a), F_u(b); \tau) f_\varepsilon(a) f_u(b), \tag{5}$$

where $c(\cdot,\cdot;\tau)$ denotes the p.d.f. associated with $C(\cdot,\cdot;\tau)$. The log-likelihood function then follows as

$$\begin{aligned}
&\ln L(\beta, \gamma, \tau, f_\varepsilon, f_u; Z) \\
&= \sum_{i=1}^{n} \left\{ (1-d_i) \ln \int_{-\infty}^{\infty} \int_{-\infty}^{-w_i'\gamma} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon + d_i \ln \int_{-w_i'\gamma}^{\infty} h_{\varepsilon,u}(y_i - x_i'\beta, u) du \right\} \\
&= \sum_{i=1}^{n} \left\{ (1-d_i) \ln F_u(-w_i'\gamma) + d_i \ln \left(f_\varepsilon(y_i - x_i'\beta) - \frac{\partial H_{\varepsilon,u}(\varepsilon, -w_i'\gamma)}{\partial \varepsilon}\bigg|_{\varepsilon = y_i - x_i'\beta} \right) \right\},
\end{aligned} \tag{6}$$

where $Z = \{z_i\}_{i=1}^n$ and $z_i = (y_i, x_i, d_i, w_i)$ denotes the observed data. Note that the log-likelihood function is not only maximized over the structural parameters β, γ and τ but over the unknown functions f_ε and f_u as well. Furthermore, note that it suffices that the log-likelihood function depends on f_ε and f_u and not additionally on F_ε and F_u, because we have that $F_\varepsilon(x) = \int_{-\infty}^{x} f_\varepsilon(v)dv$ and $F_u(x) = \int_{-\infty}^{x} f_u(v)dv$. Our interest focuses on estimation of the structural parameters $\theta = (\beta', \gamma', \tau)'$, while the unknown functions f_ε and f_u are considered as nuisance parameters. Remember that the first element of γ is equal to one in absolute value due to identification, hence it need not be estimated. This restriction will be suppressed in the following in order to ease the notation.

Since f_ε and f_u are of infinite dimension, estimation requires that we approximate these functions. We follow Chen et al. (2006) and Chen (2007) and approximate these densities by the method of sieves. That means, we approximate an unknown function (the densities) by a (e.g., linear) combination of known basis functions (such as polynomials or splines) and unknown sieve coefficients. The unknown sieve coefficients are then estimated

along with the structural parameters β, γ and τ. Since we approximate density functions, we have to restrict the approximating functions to satisfy two fundamental properties of densities, i.e., that they are not negative and that they integrate to one. The former property can be satisfied if we approximate not the density function by the method of sieves but the square root of the density function instead. This is the approach taken in Chen et al. (2006), who propose the following sieve space:

$$\mathcal{F}_{n,\eta} = \left\{ f_{n,\eta}(x) = \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta} A_{k,\eta}(x) \right]^2, \int f_{n,\eta,}(x)dx = 1 \right\}, K_{n,\eta} \to \infty, \frac{K_{n,\eta}}{n} \to 0, \quad (7)$$

where $f_{n,\eta}$ is an approximation to $f_\eta, \eta \in \{\varepsilon, u\}$, based on $K_{n,\eta}$ sieve coefficients, $\{A_{k,\eta}(\cdot) : k \geq 0\}$ denote known basis functions and $\{a_{k,\eta}(\cdot) : k \geq 0\}$ are unknown sieve coefficients which must be estimated. Note that $K_{n,\eta}$ depends on the sample size n but grows at a slower rate. For the basis functions Chen et al. (2006) suggest to use Hermite polynomials or splines; for details, see Chen et al. (2006). To ensure that the approximation of the density function integrates to one in applications, one can set

$$f_{n,\eta}(x) = \frac{\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta} A_{k,\eta}(x) \right]^2}{\int \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta} A_{k,\eta}(v) \right]^2 dv}. \quad (8)$$

Let

$$h_{n,\varepsilon,u}(a,b) = c(F_{n,\varepsilon}(a), F_{n,u}(b); \tau) f_{n,\varepsilon}(a) f_{n,u}(b), \quad (9)$$

where $F_{n,\eta}(x) = \int_{-\infty}^{x} f_{n,\eta}(v)dv, \eta \in \{\varepsilon, u\}$. Then, our proposed sieve maximum likelihood

estimator $\hat{\theta}_n$ of θ is obtained by maximizing

$$\begin{aligned}
&\ln L(\beta, \gamma, \tau, a_{0,\varepsilon}, \ldots, a_{K_{n,\varepsilon},\varepsilon}, a_{0,u}, \ldots, a_{K_{n,u},u}; Z) \\
&= \sum_{i=1}^{n} \left\{ (1-d_i) \ln \int_{-\infty}^{\infty} \int_{-\infty}^{-w_i'\gamma} h_{n,\varepsilon,u}(\varepsilon, u) du d\varepsilon + d_i \ln \int_{-w_i'\gamma}^{\infty} h_{n,\varepsilon,u}(y_i - x_i'\beta, u) du \right\} \\
&= \sum_{i=1}^{n} \left\{ (1-d_i) \ln F_{n,u}(-w_i'\gamma) + d_i \ln \left(f_{n,\varepsilon}(y_i - x_i'\beta) - \frac{\partial H_{n,\varepsilon,u}(\varepsilon, -w_i'\gamma)}{\partial \varepsilon}\bigg|_{\varepsilon = y_i - x_i'\beta} \right) \right\}
\end{aligned} \tag{10}$$

over θ and the unknown sieve coefficients $(a_{0,\varepsilon}, \ldots, a_{K_{n,\varepsilon},\varepsilon}, a_{0,u}, \ldots, a_{K_{n,u},u})$.

As an example, we consider the well-known Gaussian copula. In that case, the joint cumulative distribution function of ε and u is given by

$$H_{\varepsilon,u}(a, b) = \Phi_2(\Phi^{-1}(F_\varepsilon(a)), \Phi^{-1}(F_u(b)); \tau), \tag{11}$$

where $\Phi_2(\cdot, \cdot, \tau)$ is the c.d.f. of the bivariate standard normal distribution with correlation coefficient τ, i.e.,

$$\Phi_2(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} \frac{1}{2\pi\sqrt{1-\tau^2}} \exp\left(-\frac{1}{2(1-\tau^2)} (x^2 + y^2 - 2\tau xy) \right) dy dx, \tag{12}$$

and $\Phi^{-1}(\cdot)$ is the inverse of the c.d.f. of the univariate standard normal distribution. This implies that the joint p.d.f. of ε and u is given by

$$\begin{aligned}
h_{\varepsilon,u}(a, b) = &\left| \begin{pmatrix} 1 & \tau \\ \rho & 1 \end{pmatrix} \right|^{-1/2} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(F_\varepsilon(a)) \\ \Phi^{-1}(F_u(b)) \end{pmatrix}' \left(\begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}^{-1} - I_2 \right) \begin{pmatrix} \Phi^{-1}(F_\varepsilon(a)) \\ \Phi^{-1}(F_u(b)) \end{pmatrix} \right) \\
&\times f_\varepsilon(a) f_u(b),
\end{aligned} \tag{13}$$

where I_2 is the 2-by-2 identity matrix. Lee (1983) was the first who applied the Gaussian

copula to sample selection models. He showed that the log-likelihood function is given by

$$\ln L = \sum_{i=1}^{N} \{(1-d_i)\ln(1-F_u(w_i'\gamma)) \\ +d_i \ln f_\varepsilon(y_i - x_i'\beta) + d_i \ln \Phi\left(\frac{\Phi^{-1}(F_u(w_i'\gamma)) + \tau\Phi^{-1}(F_\varepsilon(y_i - x_i'\beta))}{\sqrt{1-\tau^2}}\right)\Bigg\}. \quad (14)$$

Besides the Gaussian copula, there exist many other copulas which can be used to model dependencies among the error terms. Popular examples are copulas from the Farlie-Gumbel-Morgenstern (FGM) family and the Archimedean class of copulas. The Archimedean class encompasses some well-known copulas such as the Clayton copula, the Frank copula and the Gumbel copula. We refer the reader to Smith (2003) for a description of these copulas. Smith (2003) also provides the likelihood functions for sample selection models based on these copulas.

3 Asymptotic Properties

In this section, we derive consistency and asymptotic normality of our proposed sieve maximum likelihood estimator using the results in Chen et al. (2006) and Chen (2007). First, let $\mathcal{A} = \Theta \times \mathcal{F}_\varepsilon \times \mathcal{F}_u$ denote the parameter space. As in the last section, the sieve MLE is defined as

$$\hat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}_n} \ln L(\alpha; Z) = \sum_{i=1}^{n} \ln l(\alpha, z_i), \quad (15)$$

where $\alpha = (\theta', f_\varepsilon, f_u)'$ and $\hat{\alpha}_n = (\hat{\theta}_n', \hat{f}_{n,\varepsilon}, \hat{f}_{n,u})' \in \Theta \times \mathcal{F}_{n,\varepsilon} \times \mathcal{F}_{n,u} = \mathcal{A}_n$. The true value of the parameter vector is denoted as $\alpha_0 = (\theta_0', f_{0,\varepsilon}, f_{0,u})' \in \mathcal{A}$.

Our first goal is to derive consistency of our proposed estimator. Suppose that $d(\cdot,\cdot)$ is a (pseudo) metric on $\mathcal{A}$. We make the following assumptions (in addition to Assumptions 1-4), which are taken from Conditions 3.1', 3.2', 3.3', 3.4 and 3.5 in Chen (2007):

ASSUMPTION 5: *(i)* $E[\ln L(\alpha, Z)]$ *is continuous at* $\alpha_0 \in \mathcal{A}$, $E[\ln L(\alpha_0, Z)] > -\infty$
(ii) for all $\epsilon > 0$, $E[\ln L(\alpha_0, Z)] > \sup_{\{\alpha \in \mathcal{A}:\, d(\alpha,\alpha_0) \geq \epsilon\}} E[\ln L(\alpha, Z)]$.

ASSUMPTION 6: $\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}$ *for all* $k \geq 1$*; and for any* $\alpha \in \mathcal{A}$ *there exists a sequence*

$\pi_k\alpha_0 \in \mathcal{A}_k$ *such that* $d(\alpha_0, \pi_k\alpha_0) \to 0$ *as* $k \to \infty$.

ASSUMPTION 7: *For each* $k \geq 1$,

(i) $\ln L(\alpha, Z)$ *is a measurable function of the data* Z *for all* $\alpha \in \mathcal{A}_k$*; and*

(ii) for any data Z, $\ln L(\alpha, Z)$ *is upper semicontinuous on* $\mathcal{A}_k$ *under the metric* $d(\cdot, \cdot)$.

ASSUMPTION 8: *The sieve spaces,* $\mathcal{A}_k$*, are compact under* $d(\cdot, \cdot)$.

ASSUMPTION 9: *For all* $k \geq 1, \text{plim}_{n\to\infty} \sup_{\alpha\in\mathcal{A}_k} |\ln L(\alpha) - E[\ln L(\alpha)]| = 0$.

Assumption 5 is an identification condition which implies that the true parameter vector α_0 uniquely maximizes the expected value of the log-likelihood function. Assumptions 6 and 8 contain assumptions on the sieve spaces. In particular, it is assumed that asymptotically the difference between an (unknown) function and its sieve approximation tends to zero. Assumption 7 is a continuity condition, while Assumption 9 assumes uniform convergence of the sample log-likelihood to its population counterpart over the sieves.

We establish the following consistency theorem:

THEOREM 1: *Suppose that Assumptions 1-9 hold. Then* $d(\hat{\alpha}_n, \alpha_0) = o_p(1)$.

PROOF: *See Chen (2007), pp. 5589-5591.*

In order to establish asymptotic normality, we show that Conditions 4.1-4.5 in Chen (2007) are fulfilled. We derive asymptotic normality only for the structural parameters of interest contained in θ. Our exposition closely follows Chen (2007, ch. 4).

Let

$$\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[\alpha - \alpha_0] = \lim_{\omega\to 0} \frac{l(\alpha_0 + \omega[\alpha - \alpha_0], z) - l(\alpha_0, z)}{\omega} \tag{16}$$

be the directional derivative of $l(\alpha_0, z)$ in the direction $[\alpha - \alpha_0]$ and suppose that it is well defined for almost all z. Let V be the completion of the space spanned by $\mathcal{A} - \alpha_0$. As in Chen et al. (2006), we define the Fisher norm on this space as

$$||v||^2 = E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[v]\right]^2, \tag{17}$$

which induces the Fisher inner product

$$\langle v, \tilde{v} \rangle = E\left[\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[v]\frac{\partial l(\alpha_0, z)}{\partial \alpha'}[\tilde{v}]\right]. \tag{18}$$

Let $f(\theta_0) = \lambda'\theta_0$, where λ is an arbitrary unit vector with the same dimension as θ. It follows from the Riesz representation theorem that there exists $v^* \in V$ such that, for any $\alpha - \alpha_0 \in V$,

$$\lambda'(\theta - \theta_0) = \langle \alpha - \alpha_0, v^* \rangle \tag{19}$$

with $||v^*|| < \infty$.

To proceed further, it is necessary to compute the Riesz representer v^*. Define

$$D_{w_j}(z) = \frac{\partial l(\alpha_0, z)}{\partial \theta_j} - \frac{\partial l(\alpha_0, z)}{\partial f'}[w_j], \quad j = 1, \ldots, \dim(\theta), \tag{20}$$

where $f = (f_\varepsilon, f_u)'$. Then, the Riesz representer $v^* = ((v_\theta^*)', (v_f^*)')'$ is given by

$$v_f^* = -(w^*)'v_\theta^* \tag{21}$$

$$v_\theta^* = (E[D_{w^*}(z)D_{w^*}(z)'])^{-1}\lambda \tag{22}$$

$$w_j^* = \arg\inf_{w_j} E[(D_{w_j}(z))^2], \tag{23}$$

where $w = (w_1, \ldots, w_{\dim(\theta)})'$ and $D_w(z) = (D_{w_1}(z), \ldots, D_{w_{\dim(\theta)}(z)})'$.

We make the following assumptions:

ASSUMPTION 10: *$\theta \in \text{int}(\Theta), \Theta$ a compact subset of $\mathbb{R}^{\dim(\theta)}$.*

ASSUMPTION 11: *The log-likelihood function $\ln L(\alpha, z)$ is twice continuously pathwise differentiable with respect to $\alpha \in \mathcal{A}$ and $||\alpha - \alpha_0|| = o(1)$, and the derivatives are uniformly bounded with respect to $\alpha \in \mathcal{A}$ and z.*

ASSUMPTION 12: *$E[D_{w^*}(z)D_{w^*}(z)']$ is positive definite.*

ASSUMPTION 13: *There is $\pi_n v^* \in \mathcal{A}_n$ such that $||\pi_n v^* - v^*|| = O(K^{-\psi}) = o(n^{-1/2})$.*

Assumptions 10-12 are standard. Assumption 13 places a smoothness condition on

the Riesz representer v^*, which is similar to Assumption 3 of Newey (1997). We establish the following theorem:

THEOREM 2: *Suppose that Assumptions 1-4 and 10-13 hold, and that* $||\hat{\alpha}_n - \alpha_0|| = o_p(1)$. *Then* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_*(\theta_0)^{-1})$, *where* $I_*(\theta_0) = E[D_{w^*}(z)D_{w^*}(z)']$.

PROOF: *See the Appendix.* Furthermore, $\hat{\theta}_n$ is semiparametrically efficient (see Chen, 2006).

In order to calculate standard errors or confidence intervals for $\hat{\theta}_n$, one needs an estimate of the asymptotic covariance matrix $I_*(\theta_0)^{-1}$. Such an estimate can be obtained in the following way (see Chen, 2007, p. 5616). Let

$$\hat{w}_j^* = \arg \min_{w_j \in (\mathcal{F}_{n,\varepsilon} \times \mathcal{F}_{n,u})} \frac{1}{n} \sum_{i=1}^{n} [(\hat{D}_{w_j}(z_i))^2], \tag{24}$$

with

$$\hat{D}_{w_j}(z) = \frac{\partial l(\hat{\alpha}_0, z)}{\partial \theta_j} - \frac{\partial l(\hat{\alpha}_0, z)}{\partial f'}[w_j], \quad j = 1, \ldots, \dim(\theta). \tag{25}$$

Define $\hat{D}_w(z) = (\hat{D}_{w_1}, \ldots, \hat{D}_{w_{\dim(\theta)}})'$. Then an estimate $\hat{I}_*(\hat{\theta}_n)^{-1}$ of $I_*(\theta_0)^{-1}$ is given by

$$\hat{I}_*(\hat{\theta}_n)^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} [\hat{D}_{\hat{w}^*}(z_i) \hat{D}_{\hat{w}^*}(z_i)'] \right)^{-1}. \tag{26}$$

The following theorem establishes the consistency of $\hat{I}_*(\hat{\theta}_n)^{-1}$ for $I_*(\theta_0)^{-1}$:

THEOREM 3: *Suppose that Assumptions 1-4 and 10-13 hold, and that* $||\hat{\alpha}_n - \alpha_0|| = o_p(1)$. *Then* $\hat{I}_*(\hat{\theta}_n)^{-1} = I_*(\theta_0)^{-1} + o_p(1)$.

PROOF: *Follows from Lemma 2 in Ackerberg et al. (2012), pp. 493-494.*

In fact, Ackerberg et al. (2012) showed that there is a simpler way to obtain $\hat{I}_*(\hat{\theta}_n)^{-1}$. Suppose there is a fictitious practitioner who uses the same sieve space (7) to approximate the unknown densities f_ε and f_u, but she treats the number of sieve terms, $K_{n,\eta}, \eta \in \{\varepsilon, u\}$, as fixed rather than as increasing with the sample size. Consequently, she has a finite dimensional parameter vector and maximum likelihood estimation and inference can be carried out as usual. However, since the number of sieve terms is considered fixed, the

maximum likelihood estimator will not be consistent for the parameters of interest (i.e., θ) and will not have the correct limiting distribution proposed in Theorem 2.

To fix ideas, let $\tilde{\alpha} = (\theta', \kappa')'$ denote the parameter vector to be estimated by our fictitious practitioner, where $\kappa = (a_{0,\varepsilon}, \ldots, a_{K_{n,\varepsilon},\varepsilon}, a_{0,u}, \ldots, a_{K_{n,u},u})'$ contains the sieve coefficients. Note that the practitioner faces the same problem as in our sieve estimation approach, but the difference is that the practitioner treats $K_\eta, \eta \in \{\varepsilon, u\}$ as fixed. The information matrix of the practitioner is given by

$$\tilde{I}(\tilde{\alpha}_0) = \begin{bmatrix} E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta'}\right] & E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa'}\right] \\ E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \theta'}\right] & E\left[\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa}\frac{\partial l(z,\theta_0,\kappa_0)}{\partial \kappa'}\right] \end{bmatrix}, \tag{27}$$

which can be consistently estimated by

$$\hat{\tilde{I}}(\hat{\tilde{\alpha}}_n) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta'} & \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa'} \\ \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \theta'} & \frac{1}{n}\sum_{i=1}^n \frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa}\frac{\partial l(z,\hat{\theta}_n,\hat{\kappa}_n)}{\partial \kappa'} \end{bmatrix}, \tag{28}$$

where $\hat{\theta}_n$ and $\hat{\kappa}_n$ denote the practitioner's estimates of θ_0 and κ_0. An estimate of the asymptotic covariance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is then given by the upper left block of the inverse of $\hat{\tilde{I}}(\hat{\tilde{\alpha}}_n)$. Ackerberg et al. (2012) derived the following result: Despite the fact that the likelihood function is misspecified (since $K_{n,\varepsilon}$ and $K_{n,u}$ are treated as fixed), the practitioner's estimate of the asymptotic covariance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is numerically equivalent to $\hat{I}_*(\hat{\theta}_n)^{-1}$.

The practical implication of this result is simple but powerful. A researcher who wants to carry out sieve maximum likelihood estimation just has to maximize the log-likelihood function over θ and the unknown sieve coefficients, and the (asymptotically) correct standard errors for $\hat{\theta}_n$ can be easily obtained from the inverse of the information matrix, *provided that the information matrix is based on the outer product of gradients* (and not on the Hessian matrix). Hence, any statistical software package which is capable of dealing with user-supplied likelihood functions can be used for sieve maximum likelihood estimation and inference, as long as the researcher is allowed to specify how the information matrix shall be calculated.

4 Remarks and Extensions

4.1 Closed Form Likelihood Function

The log-likelihood function in (10) does not exhibit a closed form expression due to the presence of integral terms. Integrals arise because of the presence of the distribution functions $F_{n,\varepsilon}$ and $F_{n,u}$, which are related to $f_{n,\varepsilon}$ and $f_{n,u}$ via $F_{n,\varepsilon}(x) = \int_{-\infty}^{x} f_{n,\varepsilon(v)}dv$ and $F_{n,u}(x) = \int_{-\infty}^{x} f_{n,u}(v)dv$. Moreover, the copula function may contain integrals as well; the Student t copula would be an example where this is the case (Demarta and McNeil, 2005). Calculating the integrals within an optimization routine is of course possible, but may be computationally demanding if the sample size and/or the number of parameters increases. Put differently, it may take a quite long time until the optimization routine finds the maximum likelihood estimates. In this subsection we describe a method how the integrals in $F_{n,\varepsilon}$ and $F_{n,u}$ can be replaced by closed form expressions, which may facilitate maximum likelihood estimation in practice. The integrals appearing through the copula function are not considered here.[2] Fortunately, many well known copulas (such as the Gaussian copula, Archimedean copulas) indeed have closed form expressions. Sample selection models based on these copulas are analyzed in Smith (2003).

Our method to obtain closed form expressions for $F_{n,\varepsilon}$ and $F_{n,u}$ essentially relies on an expansion of the unknown densities f_ε and f_u by Hermite polynomials. More specifically, we propose to approximate the unknown density functions by

$$f_{n,\eta}(x) = \frac{\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(x/\sigma_\eta)^k\right]^2 \phi(x/\sigma_\eta)/\sigma_\eta}{\int_{-\infty}^{\infty}\left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}, \quad \eta \in \{\varepsilon, u\}, \tag{29}$$

where $\sigma_\eta > 0$ is a scale parameter which must be estimated, and $\phi(\cdot)$ is the standard normal probability density function.

An important advantage of using Hermite polynomials as basis functions in these

[2]To deal with such integrals, Maximum Simulated Likelihood techniques may be employed.

expansions is that $F_{n,\varepsilon}$ and $F_{n,u}$ have closed form expressions.[3] Note that

$$F_{n,\eta}(x) = \frac{\int_{-\infty}^{x} \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}{\int_{-\infty}^{\infty} \left[\sum_{k=0}^{K_{n,\eta}} a_{k,\eta}(v/\sigma_\eta)^k\right]^2 \phi(v/\sigma_\eta)/\sigma_\eta dv}, \quad \eta \in \{\varepsilon, u\}. \tag{30}$$

To see that $F_{n,\varepsilon}$ and $F_{n,u}$ have closed forms, consider the denominator of (30) first. To ease notation, we suppress η in the following formulas. The denominator can be simplified by making a change of variables $z = x/\sigma$, which yields

$$\int_{-\infty}^{\infty} \left[\sum_{k=0}^{K} a_k(x/\sigma)^k\right]^2 \phi(x/\sigma)/\sigma dx \tag{31}$$

$$= \int_{-\infty}^{\infty} [Z'aa'Z]\phi(z)dz \tag{32}$$

$$= \text{tr}\left[aa' \int_{-\infty}^{\infty} ZZ'\phi(z)dz\right], \tag{33}$$

where $a = (a_0, \ldots, a_{K_n})'$ and $Z = (z^0, z^1, z^2, \ldots, z^{K_n})'$. The integral term represents moments of the standard normal distribution. For example, if $K_n = 2$, we have that $Z = (1, z, z^2)'$ and

$$\int_{-\infty}^{\infty} ZZ'\phi(z)dz = \int_{-\infty}^{\infty} \begin{bmatrix} 1 & z & z^2 \\ z & z^2 & z^3 \\ z^2 & z^3 & z^4 \end{bmatrix} \phi(z)dz = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix} \tag{34}$$

with

$$\int_{-\infty}^{\infty} z^k dz = 0, \quad k = 1, 3, 5, \ldots \tag{35}$$

$$\int_{-\infty}^{\infty} z^k dz = (k-1)(k-3) \cdot \ldots \cdot 1, \quad k = 0, 2, 4, \ldots. \tag{36}$$

Hence, the denominator does not involve integrals any more and thus has a closed form.

[3]Using Hermite polynomials to approximate the square root of a density has been suggested by e.g. Gallant and Nychka (1987). Of course, there may exist other basis functions which imply closed form distribution functions. However, such other basis functions will not be considered here.

Next, consider the numerator of (30). We have that

$$\int_{-\infty}^{x} \left[\sum_{k=0}^{K_n} a_k (v/\sigma)^k\right]^2 \phi(x/\sigma)/\sigma dv \tag{37}$$

$$= \int_{-\infty}^{x/\sigma} \left[\sum_{k=0}^{K_n} a_k (z)^k\right]^2 \phi(z) dz \tag{38}$$

$$= \mathrm{tr}\left[aa' \int_{-\infty}^{x/\sigma} ZZ'\phi(z)dz\right] \tag{39}$$

and

$$\int_{-\infty}^{x/\sigma} ZZ'\phi(z)dz = \begin{bmatrix} b'_{0:K_n} \\ b'_{1:(K_n+1)} \\ \vdots \\ b'_{K_n:(2K_n)} \end{bmatrix}_{([K_n+1]\times[K_n+1])}, \tag{40}$$

where $b'_{i:j} = (b_i, \ldots, b_j)$ with

$$b_0 = \Phi(x/\sigma) \tag{41}$$

$$b_1 = -\phi(x/\sigma) \tag{42}$$

$$b_k = -\phi(x/\sigma)(x/\sigma)^{k-1} + (k-1)b_{k-2}, \quad k = 2, \ldots, 2K_n, \tag{43}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function. Hence, by these transformations the integrals in the numerator of (30) vanish as well. Therefore, $F_{n,\varepsilon}$ and $F_{n,u}$ have closed form expressions.

4.2 Initial Values for Maximum Likelihood Estimation

The likelihood function (10) usually contains a lot of parameters to be estimated, since the sieve coefficients must be estimated as well. As in case of integral terms, this may be associated with further computational complexity. However, having good initial values for the maximum likelihood estimation routine may reduce this computational burden. Such initial values can be easily obtained for the parameters β and γ if consistent estimates

are available. For instance, the parameters of the selection equation, γ, may be estimated by a suitable semiparametric estimator for binary choice models. The Klein and Spady (1993) semiparametric estimation procedure can be used in this case. The parameters of the main equation, β, can be estimated by the approaches proposed by Powell (1987) or Newey (2009).

4.3 Testing for the Validity of Parametric Assumptions

As described in the introduction, a great advantage of our estimation approach is that it is easy to test for the validity of parametric assumptions. Testing for the validity of parametric assumptions is important since incorporating (correct) parametric information into a model typically results in efficiency gains.

It is easy to test for the validity of parametric assumptions in our copula framework because one can separately test for the validity of a certain joint distribution (represented by the copula) and for the validity of certain marginal distributions. Suppose we want to test if a certain copula is valid to describe the joint distribution of error terms. We could then estimate the model by the Gallant and Nychka (1987) procedure which does not make any (parametric) assumptions on the joint distribution. Then we would estimate the model by our approach, including the assumed copula whose validity we seek to test. Since the Gallant and Nychka (1987) and our approach are based on maximum likelihood estimation, one can test whether the parametric copula assumption is justified by applying the Vuong (1989) test for nonnested models.

In a similar manner, one can test for the validity of certain parametric marginal distributions. Given a valid parametric copula, we would estimate the model by our approach with unspecified marginal distributions, and then with one or both marginal distributions parametrically specified. Again the Vuong (1989) test may help decide if the parametric assumptions on the marginal distribution(s) are correct. In case of the selection equation only one may also apply the Horowitz and Härdle (1994) testing procedure to test if a certain parametric marginal distribution is valid for the selection

equation's error term.[4].

4.4 Binary Dependent Variable

This subsection focuses on an extension of our semiparametric copula model to the case of a binary dependent variable. Sample selection models with a binary dependent variable have been used by van de Ven and van Praag (1981), Boyes et al. (1989), Greene (1992) and Mohanty (2002). These authors, however, assumed a bivariate normal distribution for the error terms of main and selection equation, as Heckman (1979) did. Thus, the following exposition generalizes these models by allowing for distributions apart from the bivariate normal.

The model is now given by

$$y_i^* = x_i'\beta + \varepsilon_i \tag{44}$$

$$d_i^* = w_i'\gamma + u_i \tag{45}$$

$$d_i = 1(d_i^* > 0) \tag{46}$$

$$y_i = \begin{cases} 1(y_i^* > 0) \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases} . \tag{47}$$

The difference between this model and the benchmark model from Section 2 is that the dependent variable associated with the main equation, y_1, now assumes only the values one or zero.

Under the same assumptions as in Section 2, the log-likelihood function for this model

[4]In the context of sample selection models, the Horowitz and Härdle testing procedure has been applied by e.g. Martins (2001) and Genius and Strazzera (2008).

is given by

$$\ln L(\beta, \gamma, f_\varepsilon, f_u; Z) = \sum_{i=1}^{n} \Bigg\{ (1-d_i) \ln \int_{-\infty}^{\infty} \int_{-\infty}^{-w_i'\gamma} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon$$
$$+ d_i(1-y_i) \ln \int_{-\infty}^{-x_i'\beta} \int_{-w_i'\gamma}^{\infty} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon + d_i y_i \ln \int_{-x_i'\beta}^{\infty} \int_{-w_i'\gamma}^{\infty} h_{\varepsilon,u}(\varepsilon, u) du d\varepsilon \Bigg\}$$
$$= \sum_{i=1}^{n} \{(1-d_i) \ln F_u(-w_i'\gamma) + d_i(1-y_i) \ln[F_\varepsilon(-x_i'\beta) - H_{\varepsilon,u}(-w_i'\gamma, -x_i'\beta)]$$
$$+ d_i y_i \ln H_{\varepsilon,u}(-w_i'\gamma, -x_i'\beta)\} . \tag{48}$$

Estimation and inference can be carried out as described above for the benchmark model. In fact, there is no conceptual difference between the model considered in this section and the benchmark model, apart from the binary nature of the dependent variable.

4.5 Endogeneity

In this subsection we show how our semiparametric copula model can be extended to the case of endogenous covariates. Taking the potential endogeneity of covariates into account is important since parameter estimates will be inconsistent otherwise. To provide an illustration, we consider the classical example for which sample selection models have been used. Suppose a researcher wants to estimate a wage equation for females, and that her interest centers on the female returns to education. If she fitted a wage regression to the observed sample of working females only, she would obtain inconsistent estimates due to sample selectivity. So she would instead fit a sample selection model to the observed data. But is sample selectivity the only source of endogeneity in this example? Indeed, there may be sociological or intelligence-related factors (which we will summarize by the term "ability") which affect not only the wage (main equation) and the probability of labor force participation (selection equation), but education as well. If the researcher does not take the potential endogeneity of education into account, she will obtain an inconsistent estimate of the female returns to education.

To conceptualize these ideas, we consider the following extension of the model from

Section 2:

$$y_{1i}^* = \tilde{x}_i'\tilde{\beta} + \delta_1 y_{2i} + \tilde{\varepsilon}_i \tag{49}$$

$$d_i^* = \tilde{x}_i'\tilde{\gamma} + \delta_2 y_{2i} + \delta_3 \tilde{w}_i + \tilde{u}_i \tag{50}$$

$$d_i = 1(d_i^* > 0) \tag{51}$$

$$y_i = \begin{cases} y_i^* \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases} \tag{52}$$

$$y_{2i} = \tilde{x}_i'\alpha_1 + q_i'\alpha_2 + v_i, \tag{53}$$

where y_2 is the endogenous covariate. The third equation is a reduced form equation for y_2 which includes an *instrumental variable* q which is not contained in x or w (exclusion restriction). Furthermore, v is an error term which is assumed to be independent of $\tilde{x}$, $\tilde{w}$ and q, but correlated with $\tilde{\varepsilon}$ and $\tilde{u}$. For instance, v, $\tilde{\varepsilon}$ and $\tilde{u}$ may be affected by a common variable like ability in the aforementioned example.

To estimate this model, we can insert the reduced form equation for y_2 into the main and selection equation, which gives the following reduced form model:

$$y_{1i}^* = x_i'\beta + \varepsilon_i \tag{54}$$

$$d_i^* = w_i'\gamma + u_i \tag{55}$$

$$d_i = 1(d_i^* > 0) \tag{56}$$

$$y_i = \begin{cases} y_i^* \text{ if } d_i = 1 \\ \text{“missing” otherwise} \end{cases} \tag{57}$$

$$y_{2i} = x_i'\alpha + v_i, \tag{58}$$

where $x = (\tilde{x}', q')', w = (x', \tilde{w}')', \varepsilon = \delta_1 v + \tilde{\varepsilon}, u = \delta_2 v + \tilde{u}, \beta = ((\tilde{\beta} + \delta_1\alpha_1)', \delta_1\alpha_2')', \gamma = ((\tilde{\gamma} + \delta_2\alpha_1)', \delta_2\alpha_2', \delta_3')'$ and $\alpha = (\alpha_1', \alpha_2')'$. Note that this model is conceptually similar to the model from Section 2. If only the reduced form parameters β and γ were of interest, the model could be estimated as in Section 2. However, one usually seeks to estimate the

structural parameters $(\tilde{\beta}', \delta_1, \tilde{\gamma}', \delta_2, \delta_3, \alpha)'$. We propose the following estimation strategy: Obtain the first order conditions associated with the likelihood function resulting from the reduced form equations (54)-(57). The likelihood function is the same as in Section 2, because the reduced form equations contain exogenous covariates only. Then estimate the structural parameters by using the first order conditions *and* the reduced form equation for y_2 in a Generalized Method of Moments or minimum distance framework. This procedure will give consistent estimates of the structural parameters. Asymptotic normality results can be derived as well, but may be different from those in Section 2 (depending on the estimation procedure). However, the results in Ackerberg et al. (2012) can still be applied: The estimation problem may be treated as if it were parametric, and parameter estimates and estimates of standard errors and confidence intervals may be obtained in the usual parametric way. As demonstrated by Ackerberg et al. (2012) for quite general classes of estimators, the standard error estimates are numerically equivalent to those which would be obtained under the correct presumption that the estimation problem was semiparametric.

One word of caution remains, though. The joint distribution implied by the copula and the (unknown) marginal distributions is now the joint distribution of the composite error terms ε and u. This has to be taken into account when interpreting the joint distribution of the error terms associated with the reduced form model.

5 Conclusions

In this paper we proposed a sieve maximum likelihood estimation approach for a copula-based sample selection model. We also provided the asymptotic properties of our proposed estimator and showed that its asymptotic covariance matrix can be easily obtained using statistics software which is capable of dealing with user-supplied likelihood functions. To facilitate estimation, we showed how closed form likelihood functions can be obtained and how appropriate initial values for maximum likelihood estimation may be chosen. We demonstrated that parametric assumptions on the joint or marginal distributions of error terms can be easily tested for in our framework. We also extended our basis model

to the cases of a binary dependent variable and endogeneity of covariates.

The semi-nonparametric maximum likelihood estimation approach of Gallant and Nychka (1987) has not often been used in applied econometrics. One reason may be that no distribution theory is available, which is necessary to compute standard errors and confidence intervals. Another reason may be that the approximation of a two-dimensional density function is rather complex, hence the whole estimation problem is complex as well. The approach derived in this paper reduces the complexity since only one-dimensional densities have to be approximated. Furthermore, standard errors and confidence intervals can be easily obtained in practice by treating the estimation problem *as if* it were parametric. We thus hope that our exposition fosters the application of semi-nonparametric maximum likelihood estimators to sample selection models, especially if the distribution of the error terms of main and selection equation is of interest.

References

Ackerberg D, Chen X, Hahn J. 2012. A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics* **94**: 481-498.

Ahn H, Powell JL. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* **58**: 3-29.

Boyes WJ, Hoffman DL, Low SA. 1989. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics* **40**: 3-14.

Chen X. 2007. Large sample sieve estimation of semi-nonparametric models. In Heckman J, Leamer E (eds.) *Handbook of Econometrics*, volume 6 of *Handbook of Econometrics*, chapter 76. Elsevier.

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

Chen X, Fan Y, Tsyrennikov V. 2006. Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association* **101**: 1228-1240.

Das M, Newey WK, Vella F. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* **70**: 33-58.

Demarta S, McNeil AJ. 2005. The t copula and related copulas. *International Statistical Review* **73**: 111-129.

Gallant AR, Nychka DW. 1987. Semi-nonparametric maximum likelihood estimation. *Econometrica* **55**: 363-390.

Genius M, Strazzera E. 2008. Applying the copula approach to sample selection modelling. *Applied Economics* **40**: 1443-1455.

Greene WH. 1992. A statistical model for credit scoring. Working Paper No. EC-95-6, Department of Economics, Stern School of Business, New York University.

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* **47**: 153-161.

Heckman JJ, Sedlacek GL. 1990. Self-selection and the distribution of hourly wages. *Journal of Labor Economics* **8**: S329-S363.

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

Horowitz JL, Härdle W. 1994. Testing a parametric model against a semiparametric alternative. *Econometric Theory* **10**: 821-848.

Joe H. 1997. *Multivariate Models and Dependence Concepts*. Chapman & Hall: London.

Klein RW, Spady RH. 1993. An efficient semiparametric estimator for binary response models. *Econometrica* **61**: 387-421.

Lee LF. 1983. Generalized econometric models with selectivity. *Econometrica* **51**: 507-12.

Martins MFO. 2001. Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in portugal. *Journal of Applied Econometrics* **16**: 23-39.

Mohanty, MS. 2002. A bivariate probit approach to the determination of employment: a study of teem employment differentials in Los Angles county. *Applied Economics* **34**: 143-156.

Mulligan CB, Rubinstein Y. 2008. Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics* **123**: 1061-1110.

Nelsen RB. 2006. *An Introduction to Copulas*. Springer: New York, NY.

Newey WK. 2009. Two-step series estimation of sample selection models. *Econometrics Journal* **12**: S217-S229.

Powell JL. 1987. Semiparametric estimation of bivariate limited dependent variable models. Manuscript, University of California, Berkeley.

Prieger JE. 2002. A flexible parametric selection model for non-normal data with application to health care usage. *Journal of Applied Econometrics* **17**: 367-392.

Smith MD. 2003. Modelling sample selection using archimedean copulas. *Econometrics Journal* **6**: 99-123.

Trivedi PK, Zimmer DM. 2007. *Copula Modeling: An Introduction for Practitioners*. Foundations and Trends in Econometrics. Vol. 1, No. 1. Now Publishers.

Van den Ven WPMM, van Praag BMS. 1981. The demand for deductibles in private health insurance: a probit model with sample selection. *Journal of Econometrics* **17**: 229-252.

Vella F. 1998. Estimating models with sample selection bias: A survey. *Journal of Human Resources* **33**: 127-169.

Vuong QH. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57**: 307-333.

Appendix

Proof of Theorem 2:

We prove Theorem 2 by verifying that the Conditions 4.1-4.5 in Chen (2007) are fulfilled. For convenience, we restate these conditions here. In the following, $\mu_n(g(z)) = \frac{1}{n}\sum_{i=1}^{n}(g(z_i) - E[g(z_i)])$ denotes the empirical process indexed by the function g.

Condition 4.1:

(i) There is $\omega > 0$ *such that* $|f(\theta) - f(\theta_0) - \frac{\partial f(\theta_0}{\theta}[\theta - \theta_0]| = O(||\theta - \theta_0||^{\omega})$ *uniformly in* $\theta \in \Theta$ *with* $||\theta - \theta_0|| = o(1)$.

(ii) $||\frac{\partial f(\theta_0)}{\partial \theta}|| < \infty$.

(iii) There is $\pi_n v^* \in \mathcal{A}_n$ such that $||\pi_n v^* - v^*|| \times ||\hat{\alpha}_n - \alpha_0|| = o_p(n^{-1/2})$.

Condition 4.2':

$$\sup_{\{\bar{\alpha}\in\mathcal{A}_n:||\bar{\alpha}-\alpha_0||<\delta_n\}} \mu_n\left(\frac{\partial l(\bar{\alpha}, z)}{\partial \alpha}[\pi_n v^*] - \frac{\partial l(\alpha_0, z)}{\partial \alpha}[\pi_n v^*]\right) = o_p(n^{-1/2}). \tag{59}$$

Condition 4.3':

$$E\left[\frac{\partial l(\hat{\alpha}_n, z)}{\partial \alpha}[\pi_n v^*]\right] = \langle \hat{\alpha}_n - \alpha_0, \pi_n v^* \rangle + o(n^{-1/2}). \tag{60}$$

Condition 4.4:

(i) $\mu_n(\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]) = o_p(n^{-1/2})$.

(ii) $E[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^*]]$.

Condition 4.5: $n^{1/2}\mu_n(\frac{\partial l(\alpha_0,z)}{\partial \alpha}[v^*]) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{v^*})$, with $\sigma^2_{v^*} > 0$.

Condition 4.1 (i) is fulfilled with $\omega = \infty$. Condition 4.1 (ii) is fulfilled by Assumption 12. Condition 4.1 (iii) is satisfied by Assumption 13 and the consistency of $\hat{\alpha}_n$.

Condition 4.2 can be verified using Theorem 3 from Chen et al. (2003). Theorem 3 requires continuity conditions on $m(z,\alpha) = \frac{\partial l(\alpha,z)}{\partial \alpha}[\pi_n v^*] - E[\frac{\partial l(\alpha,z)}{\partial \alpha}[\pi_n v^*]]$, which are satisfied in our case because of Assumption 11.

Condition 4.3 is trivially satisfied because we have used the Fisher norm (Chen 2007, p. 5617). Condition 4.4 (i) is fulfilled because we have i.i.d. observations and

$$E\left[\mu_n(\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*])\right]^2 = n^{-1}E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]\right]^2 \tag{61}$$

$$= n^{-1}||\pi_n v^* - v^*||^2 = o(n^{-1}). \tag{62}$$

Hence, by the Markov inequality we have that $\mu_n(\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]) = o_p(n^{-1/2})$. Condition 4.4 (ii) is satisfied since

$$\begin{aligned} E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^*]\right] &= E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^*]\right] - E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[v^*]\right] + E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[v^*]\right] \\ &= E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]\right], \end{aligned} \tag{63}$$

and by Jensen's inequality,

$$\begin{aligned} \left(E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]\right]\right)^2 &\leq E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]\right]^2 \\ &= ||\pi_n v^* - v^*||^2 = O(K^{-2\psi}) = o(n^{-1}) \end{aligned} \tag{64}$$

by Assumption 13, hence $E\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[\pi_n v^* - v^*]\right] = o(n^{-1/2})$. Condition 4.5 is fulfilled because we have i.i.d. observations and

$$\sigma^2_{v^*} = Var\left[\frac{\partial l(\alpha_0,z)}{\partial \alpha}[v^*]\right] \tag{65}$$

$$= Var\left[\frac{\partial l(\alpha_0,z)}{\partial \theta} - \frac{\partial l(\alpha_0,z)}{\partial f'}[w](v^*_\theta)\right] \tag{66}$$

$$= (v^*_\theta)E[D_{w^*}(z)D_{w^*}(z)'](v^*_\theta)' \tag{67}$$

$$= \lambda'(E[D_{w^*}(z)D_{w^*}(z)'])^{-1}E[D_{w^*}(z)D_{w^*}(z)'](E[D_{w^*}(z)D_{w^*}(z)'])^{-1}\lambda \tag{68}$$

$$= \lambda'(E[D_{w^*}(z)D_{w^*}(z)'])^{-1}\lambda > 0 \tag{69}$$

by Assumption 12. By Theorem 4.3 in Chen (2007) it follows that $n^{1/2}(f(\hat{\theta}_n) - f(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{v^*})$, hence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, (E[D_{w^*}(z)D_{w^*}(z)'])^{-1}) = \mathcal{N}(0, I_*(\theta_0)^{-1}). \tag{70}$$