

# A semi-parametric empirical likelihood approach for conditional estimating equations under endogenous selection

Yves G. Berger<sup>a,\*</sup>, Valentin Patilea<sup>b</sup>

<sup>a</sup>Economic, Social and Political Sciences, University of Southampton, SO17 1BJ, United Kingdom

<sup>b</sup>Center for Research in Economics and Statistics, ENSAI, France

---

## ARTICLE INFO

### Keywords:

Conditional estimating equations  
Endogenous covariates  
Endogenous stratification  
Transformation model  
Two-stage least-squares

## ABSTRACT

The estimation and inference for conditional estimating equations models with endogenous selection, are considered. The approach takes into account possible endogenous selection which may lead to a selection bias. It can be used for a wide range of statistical models not covered by the model-based sampling theory. Endogeneity can be either part of the selection or within the covariates. It is particularly well suited for models with unknown heteroscedasticity, uncontrolled confounders and measurement errors. It will not be necessary to model the relationship between the endogenous covariates and the instrumental variables, which offer major advantages over two-stage least-squares. The approach proposed has the advantage of being based on a fixed number of constraints determined by the size of the parameter.

---

## 1. Introduction

We consider a class of models defined by conditional estimating equations; that is, when the conditional expectation of an estimation function is zero when evaluated at the value of the parameter of interest. There are many situations where moment and conditional estimating equations models are used in statistical modelling and econometrics. For example, general regression and transformation models, non-linear quantile regression models, or non-linear (in the parameter) simultaneous equation models. The econometrics literature provides more examples such as non-linear simultaneous equations models of consumption, econometrics models of optimising agents (Hansen and Singleton, 1982), regression models with endogenous covariates or instrumental variables and non-separable regression models. Other examples can be found in Amemiya (1977), Newey (1993), Ai and Chen (2003), Smith (2007), Chen and Pouzo (2009) and in Sections E.4 and E.3 of the Supplement.

There are situations when the parameters of non-linear models cannot be estimated with least squares or maximum likelihood, because it gives insoluble estimating equations, even with simple models. However, we may be able to estimate those parameters semi-parametrically with conditional estimating equations. The survey sampling literature focuses on unconditional estimating equations or likelihood approaches, with exogenous covariates. We considered a wider class of models defined by conditional moment conditions, such as models with unknown heteroscedasticity, uncontrolled confounders, measurement errors and endogenous covariates.

Let  $\mathbf{Y} \in \mathbb{R}^{d_Y}$  and  $\mathbf{W} \in \mathbb{R}^{d_W}$  denote two different random vectors. Let  $\mathbf{Y}$  contains the response variables and some covariates  $\mathbf{X}$  which could be endogenous. The vector  $\mathbf{W}$  is a set of exogenous or instrumental variables. Consider a parameter  $\theta_0$  defined by a ‘conditional estimating equation’ (e.g. Chamberlain, 1987; Newey, 1993) given by


$$\mathbb{E}\{\mathbf{g}(\mathbf{Y}, \mathbf{W}, \theta) \mid \mathbf{W}\} = \mathbf{0}_{d_g} \quad a.s., \quad \text{if and only if } \theta = \theta_0, \quad (1)$$


where  $\mathbf{g}(\cdot) \in \mathbb{R}^{d_g}$  and  $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ . We assume  $d_\theta < \infty$ . Here,  $\mathbf{g}(\cdot)$  is some given differentiable function and  $\mathbf{0}$ , denotes an  $r \times 1$  vector of zeros. Equation (1) specifies a distribution free model, which does not rely on assumptions about the distribution over an error term. It allows for heteroscedastic errors and endogenous covariates and non-separable models.

We assume a random process which selects  $n$  realisations with varying probabilities (e.g. Cosslett, 1993, §3.2). We also assume that  $n$  is a constant, which leads to non-independent and identically distributed (i.i.d.) observations.

---

\*Corresponding author

 [y.g.berger@soton.ac.uk](mailto:y.g.berger@soton.ac.uk) (Yves G. Berger)

 <http://yvesberger.co.uk> (Yves G. Berger)

ORCID(s): 0000-0002-9128-5384 (Yves G. Berger)

This defines, what we will call hereafter, the “*random selection process*”. This process mirrors the structure of most survey data, where a sample of  $n$  units is selected with unequal probabilities (e.g. Pfeiffermann and Rao, 2009a, Part 1). Ignoring the selection process may lead to a selection bias. In practice, these probabilities may depend on the endogenous vector  $\mathbf{Y}$ , for example, with “*endogenous stratification*” (e.g. Cosslett, 1993). This situation hereafter will be called “*endogenous selection process*”. Most importantly, in this paper, endogeneity could be present in two distinct forms: either within the selection process or as part of the covariates.

Informative sampling and endogenous selection are similar, but different concepts. The literature on informative sampling focuses on the effect of sampling on parametric likelihoods or on unconditional estimating equations based on least squares (e.g. Pfeiffermann and Rao, 2009b, Part 6). For example, we have an informative sampling when the selection process distorts the distribution of the error term. Under endogenous selection, the selection probabilities may be independent from the error term, but still depend on endogenous variables (response or covariates). Here, we have used a semi-parametric approach which does not need distributional assumptions about the error term.

Semi-parametric methods such as pseudo-likelihood, as in Binder (1983), can be used for regression parameters. However, they cannot be used with models with endogenous covariates or non-separable regression models, such as Box-Cox transformation models (see Section E.3 in the Supplement) or models with unknown heterogeneity. Furthermore, pseudo-likelihood confidence intervals are based on Wald’s statistics, involving linearisation. There is no Wilks’s type confidence interval for pseudo-likelihood.

The customary approach consists in solving (1) by using an arbitrary “*instrument*” matrix  $\mathbf{I}(\mathbf{W}, \theta) \in \mathbb{R}^{q \times d_g}$ , with  $q \geq d_\theta$  (e.g. Donald, Imbens and Newey, 2009) which is such that

$$\mathbb{E}\{\mathbf{I}(\mathbf{W}, \theta) \mathbf{g}(\mathbf{Y}, \mathbf{W}, \theta)\} = \mathbf{0}_q, \quad \text{if and only if } \theta = \theta_0. \quad (2)$$

A sample analogue to (2) can be used to estimate  $\theta_0$ . Linear models are often based on (2), or equivalently based on ordinary least-squares estimators with the endogenous covariates replaced by fitted values of the linear relationship between endogenous covariates and instrumental variables  $\mathbf{W}$  (the so-called “*two-stage least-squares*” approach). If this relationship is non-linear, these estimators may not be consistent (see Section E.4 in the Supplement). The approach proposed is not based upon (2) and accommodates the non-linearity between the endogenous and instrumental variables.

The approach based on (2) has several issues. The instrument is arbitrary and does not guarantee that the solution to (2) is unique, even when (1) has a single solution. In this case,  $\theta_0$  cannot be identified by (2), and yield an inconsistent estimation. Consistency and identification depend on the chosen instrument (Domínguez and Lobato, 2004; Newey, 1993). With the approach proposed, we obtain consistent and identifiable estimators. It is not necessary to select an instrument which identifies the parameter unconditionally.

Empirical likelihood is a semi-parametric technique that can be used for point estimation, testing hypotheses and constructing confidence regions (e.g. Hall, 1990; Chaudhuri, Handcock and Rendall, 2008; Chen and Van Keilegom, 2009), without the need of variance estimates, which can be cumbersome. We develop a semi-parametric empirical likelihood approach which can be used for conditional estimating equations. We will see that it differs from Owen’s (1988) mainstream empirical likelihood approach, because it deals with conditional estimating equations and it takes the selection process into account. It is also different from Berger and Torres’s (2016) empirical likelihood approach, because the constraint we use is a double sum, which leads to additional theoretical challenges. The Chaudhuri and Handcock (2018) approach takes the selection process into account, when the model is defined by customary unconditional moment restrictions (2), as in Qin and Lawless (1994). We deal with a more general class of models defined by (1).

Kitamura, Tripathi and Ahn (2004) proposed an empirical likelihood approach for conditional estimating equations by using a kernel function and a trimming function within the objective function. We propose using the standard empirical likelihood objective function and to include a kernel function within the constraints. Trimming as in Kitamura et al. (2004), will not be necessary. There is also the polynomial based approach of Donald, Imbens and Newey (2003), which was reconsidered by Chang, Chen and Chen (2015). The point estimator proposed is based on a finite number  $d_\theta$  of constraints. On the other hand, with the approach of Kitamura et al. (2004), the number of constraints ( $nd_g$ ) increases with the sample size  $n$ , as with the approach of Donald et al. (2003). The fact that our estimator is determined by a fixed number  $d_\theta$  of constraints, is a major advantage over its competitors.

In Section 2, we define the selection process and the sample data. In Section 4, we show that the parameter of interest minimises a smooth distance function, by imposing a condition on the selection process, described in Section 3. We show that this condition holds under endogenous selection. The smooth distance function and point estimator

are defined in Section 4. In Section 5, we describe the empirical likelihood approach proposed. In Section 6, we present regularity conditions under which the empirical likelihood estimator is  $\sqrt{n}$ -consistent. The pivotal property of the empirical likelihood ratio function, the asymptotic normality and variance estimator of the empirical likelihood estimator are given in Section 7. Simulation studies can be found in Section 8. The proofs can be found in the Supplement. An extension to multi-stage selection and examples can be found in Section E.1 of the Supplement.

## 2. Selection process

Suppose we have  $N$  independent realisations  $(\mathbf{Y}_i^\top, \mathbf{W}_i^\top)^\top$  of the vector  $(\mathbf{Y}^\top, \mathbf{W}^\top)^\top$ , with  $i \in \mathcal{P} = \{1, \dots, N\}$ . Here and in the following,  $\mathbf{A}^\top$  denotes the transpose of a matrix  $\mathbf{A}$ . We assume that we have a selection process; that is, some but not all the  $(\mathbf{Y}_i^\top, \mathbf{W}_i^\top)^\top$  are actually observed. Let  $\pi(\mathbf{W}_i, \mathbf{Z}_i) > 0$  be the probability of observing  $(\mathbf{Y}_i^\top, \mathbf{W}_i^\top)^\top$ . Here,  $\mathbf{Z}_i$  is the realisation of a vector  $\mathbf{Z} \in \mathbb{R}^{d_Z}$ , which denotes some additional variables which characterise the selection process. In practical situations, the vector  $\mathbf{Z}$  contains stratification variables, defined by (5) below. It may also include size measurements and cluster indicators. We have an endogenous selection when the vector  $\mathbf{Z}$  includes components from  $\mathbf{Y}$ . In other words, the response variables and/or the covariates may contain variables that are part of  $\mathbf{Z}$ . This situation is considered in Section 3.

In what follows,  $\mathbf{V}$  denotes the vector containing all the variables and  $\mathbf{V}_i$  its realisation for a unit  $i \in \mathcal{P}$ ; that is,

$$\mathbf{V} := (\mathbf{Y}^\top, \mathbf{W}^\top, \mathbf{Z}^\top)^\top \quad \text{and} \quad \mathbf{V}_i := (\mathbf{Y}_i^\top, \mathbf{W}_i^\top, \mathbf{Z}_i^\top)^\top. \quad (3)$$

We assume a “with-replacement selection process”; that is,  $n$  units are selected with replacement from  $\mathcal{P}$  with probabilities  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$ , as in Durbin (1953) and Gabler (1984). When  $n/N$  is negligible, as is often in practice, inference based on, with-and without-replacement selection are asymptotically equivalent (e.g. Hájek, 1981, p.112). It is common practice to assume with-replacement sampling even under without-replacement selection processes (e.g. Lohr, 2010, p.221), as long as  $n/N$  is negligible. Let  $\tau_i$  be the random variable which specifies the number of times an observation  $\mathbf{V}_i$  is selected from  $\mathcal{P}$ . We assume that the sample size given by

$$n := \sum_{i \in \mathcal{P}} \tau_i,$$

is constant. Thus, the  $\tau_i$  are not independent. We consider that  $(\tau_i, \mathbf{V}_i^\top)^\top$  is observed if  $\tau_i \geq 1$  and not observed if  $\tau_i = 0$ . Since we assume that  $n/N$  is negligible, the  $\tau_i$  are mostly equal 0 or 1, because very few units selected more than once.

Let  $\mu_i$  denote the conditional expectation:

$$\mu_i := \mathbb{E}(\tau_i \mid \mathbf{W}_i, \mathbf{Z}_i), \quad (4)$$

where the expectation is with respect to the random selection of the sample. We assume that the random variables  $\mu_i$  are observed for all sampled units and that the units  $i$  which are such that  $\tau_i \geq 1$ . The quantities  $\mu_i$  are assumed known and shall be used for inference. In practice,  $\mu_i^{-1}$  are the sampling weights which are often provided with survey data (Lohr, 2010, p.103). When  $\mu_i^{-1}$  are not available, proxies such as calibrated weights, can be used instead of  $\mu_i^{-1}$ .

In practice, data are often stratified; that is,  $\mathcal{P}$  is divided into  $H$  non-overlapping sets called strata, denoted by  $\mathcal{P}_1, \dots, \mathcal{P}_h, \dots, \mathcal{P}_H$ ; where  $\cup_{h=1}^H \mathcal{P}_h = \mathcal{P}$ . Within  $\mathcal{P}_h$ ,  $n_h$  units are selected with-replacement, with the unequal probabilities  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  introduced above in this Section. Hence,  $\sum_{h=1}^H n_h = n$ , where  $n_h$  is assumed constant. Let  $\delta_i$  be the stratification variable defined by

$$\delta_i := (\delta_{i1}, \dots, \delta_{ih}, \dots, \delta_{iH})^\top, \quad (5)$$

where  $\delta_{ih} = 1$  when  $i \in \mathcal{P}_h$  and  $\delta_{ih} = 0$  otherwise. We have an endogenous stratification when  $\delta_i$  is associated with  $\mathbf{Y}$  (e.g. Cosslett, 1993).

When  $i \in \mathcal{P}_h$ , we have  $n_h$  trials within  $\mathcal{P}_h$  and  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  is the probability of selecting  $i$ . Thus, the count  $\tau_i$  follows a binomial distribution with expectation  $\pi(\mathbf{W}_i, \mathbf{Z}_i)n_h$ ; i.e.

$$\mu_i = \pi(\mathbf{W}_i, \mathbf{Z}_i) \sum_{h=1}^H \delta_{ih} n_h, \quad (6)$$

and  $\sum_{i \in \mathcal{P}_h} \pi(\mathbf{W}_i, \mathbf{Z}_i) = 1$ . Since  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  depend on stratification, we consider that  $\mathbf{Z}_i$  contains the stratification variable  $\delta_i$ .

The observed data are

$$\mathcal{D} := \left\{ (\tau_i, \mu_i, \mathbf{V}_i^\top)^\top : \tau_i \geq 1 \text{ and } i \in \mathcal{P} \right\},$$

where the observed value of  $\mu_i$  given by (6), is known for all  $\tau_i \geq 1$ . Note that since we know  $\mu_i$ , the  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  are also known, despite the fact that the functional relationship between  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  and  $(\mathbf{W}_i^\top, \mathbf{Z}_i^\top)^\top$  remains unknown. The variable  $\mu_i$  is included within  $\mathcal{D}$ , because  $\mu_i$  will be used hereafter to adjust for the selection process. Note that  $\mathbf{V}_i$  is known if  $\tau_i \geq 1$  and unknown otherwise. The  $\delta_i$  are only available when  $\tau_i \geq 1$ , because  $\delta_i \in \mathbf{V}_i$ .

We adopt the convention that we have  $n_1$  draws from  $\mathcal{P}_1$ , followed by  $n_2$  draws from  $\mathcal{P}_2$ , etc. Hence, the  $k$ -th draw is a unit selected from  $\mathcal{P}_h$ , if  $k \in \mathcal{U}_h$ , where  $k = 1, \dots, n$ ,

$$\mathcal{U}_h := \{k : n_{h-1}^c < k \leq n_h^c\}$$

and

$$n_\ell^c := \sum_{h=1}^{\ell} n_h, \quad \text{with } n_0^c = 0 \text{ and } \ell \in \{0, \dots, H\},$$

where  $\mathcal{U}_h$  and  $n_\ell^c$  are not random.

### 3. Ignorable and endogenous selection process

We consider that the selection process is ignorable conditionally on  $\mathbf{W}$  and  $\mathbf{Z}$ ; that is,

$$S_{ki} \perp\!\!\!\perp Y_i \mid (\mathbf{W}_i, \mathbf{Z}_i), \quad \forall k \in \{1, \dots, n\} \text{ and } \forall i \in \mathcal{P}; \quad (7)$$

where

$$S_{ki} := \begin{cases} 1 & \text{if the unit } i \text{ is selected at the } k\text{-th draw,} \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Condition (7) is needed to establish Theorem 1 in Section 4.

Under the particular case of  $\mathbf{Z}_i$  containing  $\mathbf{Y}_i$ , there is no variation of  $\mathbf{Y}_i$  conditionally on  $\mathbf{Z}_i$ . Hence,  $S_{ki}$  is still conditionally independent of  $\mathbf{Y}_i$ , because any random variable is independent of any degenerated variable.

At the  $k$ -th draw, we select one unit. Hence, we have,

$$S_{ki} S_{kj} = 0 \quad a.s. \quad \text{for } i \neq j. \quad (9)$$

Moreover,

$$S_{ki} \perp\!\!\!\perp S_{\ell j} \mid \mathbf{V}_i, \mathbf{V}_j \quad \text{for all } i, j \text{ and } k \neq \ell. \quad (10)$$

By definition of  $\tau_i$  and  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$ , we have  $\tau_i = \sum_{k=1}^n S_{ki}$  and

$$\mathbb{E}(S_{ki} \mid \mathbf{V}_1, \dots, \mathbf{V}_N) = \mathbb{E}(S_{ki} \mid \mathbf{V}_i) = \mathbb{E}(S_{ki} \mid \mathbf{W}_i, \mathbf{Z}_i) = \pi(\mathbf{W}_i, \mathbf{Z}_i) \delta_{ih}, \quad \text{with } k \in \mathcal{U}_h. \quad (11)$$

See (A.1) in Appendix A, for more details about (11).

A crucial situation is when  $\mathbf{Z}_i$  contains variables associated with components from  $\mathbf{Y}_i$ ; that is, when we have an “endogenous selection process”. Consider that (7) holds with

$$\mathbf{Z} = (\mathbf{Z}^{\dagger\top}, \mathbf{Z}^{*\top})^\top \quad \text{and} \quad \mathbf{Y} = (\mathbf{Y}^{\dagger\top}, \mathbf{Y}^{*\top})^\top, \quad \text{with } \mathbf{Y}^* \perp\!\!\!\perp \mathbf{Z}^* \mid \mathbf{W},$$

i.e.  $\mathbf{Z}$  contains  $\mathbf{Z}^\dagger$  which may or may not be dependent on  $\mathbf{Y}^\dagger$ . When  $\mathbf{Y}^\dagger \perp\!\!\!\perp \mathbf{Z}^\dagger \mid \mathbf{W}$ , if for instance  $(\mathbf{Y}^\dagger, \mathbf{Z}^\dagger) \perp\!\!\!\perp (\mathbf{Y}^*, \mathbf{Z}^*) \mid \mathbf{W}$ , we have  $(\mathbf{W}, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}$ , which always implies that

$$\pi(\mathbf{W}, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}. \quad (12)$$

On the other hand, if  $\mathbf{Z}^\dagger = \mathbf{Y}^\dagger$  or  $\mathbf{Z}^\dagger \perp\!\!\!\perp \mathbf{Y}^\dagger \mid \mathbf{W}$ , we have  $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}$  and  $(\mathbf{W}, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}$ , which usually implies

$$\pi(\mathbf{W}, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}, \quad (13)$$

even if  $(\mathbf{Y}^\dagger, \mathbf{Z}^\dagger) \perp\!\!\!\perp (\mathbf{Y}^*, \mathbf{Z}^*) \mid \mathbf{W}$ . We have an “*endogenous selection process*” when (13) holds. Our approach does not rely on (12) and allows for (13). The advantage of (7) is that (13) still hold under (7), by incorporating within  $\mathbf{Z}$ , the variables correlated with a sub-vector of  $\mathbf{Y}$ , or by having  $\mathbf{Z}^\dagger = \mathbf{Y}^\dagger$ . Hence, the approach proposed accounts for endogenous selection.

If (13) holds, the selection probability  $\pi(\mathbf{W}, \mathbf{Z})$  depends on  $\mathbf{Y}$ , after conditioning on  $\mathbf{W}$ . This situation often occurs in practice. For example, the stratification variable (5), included within  $\mathbf{Z}_i$ , may depend on (or may be a sub-vector of)  $\mathbf{Y}_i$ . The  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  may also be proportional to (or correlated with) a sub-vector of  $\mathbf{Y}_i$ . This can be the case for retrospective studies. It is not uncommon to have  $\mathbf{Z}^\dagger = \mathbf{Y}^\dagger$ , when  $\pi(\mathbf{W}, \mathbf{Z})$  is proportional to a component of  $\mathbf{Y}^\dagger$ . These situations may lead to an “*informative selection*” (Pfeffermann, Krieger and Rinott, 1998) extensively studied in the model-based sampling literature.

The parametric approach of Pfeffermann et al. (1998) requires modelling the relationship between  $\pi(\mathbf{W}, \mathbf{Z})$  and  $\mathbf{Y}$ , when the covariates are exogenous. With our semi-parametric approach, it will not be necessary to know or to model this relationship. It also accommodates endogenous covariates. Our only requirement is the knowledge of  $\mu_i$  and  $\delta_i$  respectively given by (5) and (6) (with  $\tau_i \geq 1$ ), which are also needed for the approach of Pfeffermann et al. (1998). In other words,  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  can be any function of (or correlated with) a sub-vector of  $\mathbf{Y}_i$ . The functional relationship between  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  and  $(\mathbf{W}_i^\top, \mathbf{Z}_i^\top)^\top$  does not need to be known, but the observed values of  $\pi(\mathbf{W}_i, \mathbf{Z}_i)$  are known, for the sampled units ( $\tau_i \geq 1$ ).

The assumption (7) is similar to the concept of ignorability introduced by Rubin (1976), Little (1982) and Sugden and Smith (1984). This implies that standard parametric approaches can be used, as long as the inference is based on the conditional distribution of  $\mathbf{Y} \mid \mathbf{W}, \mathbf{Z}$ . However, this approach has several problems. We need to know  $\mathbf{Z}_i$  for all  $i \in \mathcal{S}$  (Pfeffermann et al., 1998). In practice, the  $\mathbf{Z}_i$  are only known for the units sampled ( $\tau_i \geq 1$ ). Furthermore, this approach does not allow  $\mathbf{Z}^\dagger = \mathbf{Y}^\dagger$ , because the distribution of  $\mathbf{Y} \mid \mathbf{W}, \mathbf{Z}$  would have degenerate components. Our framework is semi-parametric and allows for the components of  $\mathbf{Z}$  to be a function of  $\mathbf{Y}$ .

Weighted unconditional estimating equations can be derived from (2) under (13), as long as  $\mathbf{I}(\mathbf{W}, \boldsymbol{\theta})$  is a valid instrument. This approach is called pseudo-likelihood (Binder, 1983). However, as mentioned in Section 1,  $\mathbf{I}(\mathbf{W}, \boldsymbol{\theta})$  may not be a valid instrument.

Chaudhuri and Handcock (2018) proposed a “*conditional empirical likelihood*” technique, similar in spirit to the procedure of Pfeffermann et al. (1998). It relies on assumptions about the design such conditional independence and a model for the inclusion probabilities, as in Pfeffermann et al. (1998). The distribution of the conditional empirical likelihood ratio statistics is not available. Our main contribution is to provide the asymptotic distribution of the empirical likelihood ratio statistics. Chaudhuri and Handcock’s (2018) assume that  $\pi(\mathbf{W}, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ , which is closely related to (7), and that  $\mathbf{I}(\mathbf{W}, \boldsymbol{\theta})$  is a valid instrument identifying the parameter, which is not always true in general. Our approach deals with a wider class of models with parameters being properly identified.

#### 4. Smooth distance function

The aim of this Section is to derive an unconditional moment condition, which gives a solution to (1). Let  $\mathbf{V}_i$  and  $\mathbf{V}_j$  ( $i \neq j$ ) be two independent copies of  $\mathbf{V}$ . Consider a “*smooth distance function*” defined by

$$Q(\boldsymbol{\theta}, h, f) := \mathbb{E} \left\{ \frac{\tau_i \tau_j}{\mu_i \mu_j} \mathbf{g}_i(\boldsymbol{\theta})^\top \mathbf{g}_j(\boldsymbol{\theta}) \mathcal{K}(\mathbf{W}_i - \mathbf{W}_j) \right\}, \text{ with } \mathcal{P}_h \ni i \neq j \in \mathcal{P}_f, \quad (14)$$

where  $h, f \in \{1, \dots, H\}$ ,

$$\mathbf{g}_i(\boldsymbol{\theta}) := \mathbf{g}(\mathbf{Y}_i, \mathbf{W}_i, \boldsymbol{\theta})$$

and  $\mu_i$  is defined by (4). Here,  $\mathcal{K}(\cdot) \in \mathbb{R}$  denotes any symmetric kernel function defined on  $\mathbb{R}^{d_{\mathbf{W}}}$  with a strictly positive integrable Fourier transform. We assume that  $\mathbf{W}_i$  and  $\mathbf{W}_j$  are standardised within  $\mathcal{K}(\cdot)$ .

**Theorem 1.** *Under (7), we have that  $Q(\boldsymbol{\theta}, h, f) \geq 0$  for all  $\boldsymbol{\theta} \in \Theta$  and for all  $h$  and  $f$ , with  $Q(\boldsymbol{\theta}, h, f) = 0$ , if and only if  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ .*

The proof of Theorem 1 can be found in the Appendix A.

An example of a kernel function  $\mathcal{K}(\cdot)$  is given by (49). The kernel  $\mathcal{K}(\cdot)$  within  $Q(\theta, h, f)$  ensures that Theorem 1 holds, and is not intended for local smoothing. Therefore, it has the advantage of not requiring a bandwidth selection. For the asymptotic derivation, we assume a fixed bandwidth. In the simulation study, we shall consider bandwidths that depend on  $n$ , and show that they have little effect on the precision.

Under mild usual conditions which ensure differentiation under the integral sign and that the map  $\theta \mapsto Q(\theta, h, f)$  is convex in the neighbourhood of  $\theta_0$ , Theorem 1 implies that for  $\mathcal{P}_h \ni i \neq j \in \mathcal{P}_f$ ,

$$\frac{\partial Q(\theta, h, f)}{\partial \theta} = \mathbb{E} \left\{ \frac{\tau_i \tau_j}{\mu_i \mu_j} \mathbf{g}_{ji}(\theta) \right\} + \mathbb{E} \left\{ \frac{\tau_j \tau_i}{\mu_j \mu_i} \mathbf{g}_{ij}(\theta) \right\} = \mathbf{0}_{d_\theta}, \quad \text{iff } \theta = \theta_0, \quad (15)$$

where

$$\mathbf{g}_{ji}(\theta) := \frac{\partial \mathbf{g}_j(\theta)}{\partial \theta} \mathbf{g}_i(\theta) K_{ij} \in \mathbb{R}^{d_\theta}$$

and

$$K_{ij} := \delta_{\{i \neq j\}} \mathcal{K}(\mathbf{W}_i - \mathbf{W}_j).$$

Here,  $\delta_{\{i \neq j\}} = 1$  if  $i \neq j$  and  $\delta_{\{i \neq j\}} = 0$  otherwise. The notation  $\partial/\partial \theta$  stands for the column vector of partial derivatives with respect to  $\theta$ . Since the two expectations on the right-hand side of (15) are both equal, we have that (15) implies

$$\mathbb{E} \left\{ \frac{\tau_i \tau_j}{\mu_i \mu_j} \mathbf{g}_{ji}(\theta) \right\} = \mathbf{0}_{d_\theta}, \quad \text{iff } \theta = \theta_0, \text{ for } \mathcal{P}_h \ni i \neq j \in \mathcal{P}_f. \quad (16)$$

## 5. Adjusted empirical likelihood

Consider the “adjusted empirical likelihood function”

$$\ell(\theta) = \max_{p_i: i \in \mathcal{S}} \left\{ \sum_{i \in \mathcal{P}} \tau_i \log(np_i) : p_i > 0; \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{P}} \frac{\tau_i \tau_j}{\mu_i \mu_j} p_i \mathbf{g}_{ji}(\theta) = \mathbf{0}_{d_\theta}; \sum_{i \in \mathcal{P}} \tau_i p_i \delta_i = \frac{\mathbf{n}_{st}}{n} \right\}, \quad (17)$$

where

$$\mathbf{n}_{st} := (n_1, \dots, n_H)^\top. \quad (18)$$

The vectors  $\delta_i$  are defined by (5). We assume that the constraints within (17) satisfy the usual convex hull assumption (Qin and Lawless, 1994), so that a unique value of  $\ell(\theta)$  exists for a given  $\theta$ . Within (17), it is important to use  $p_i \mathbf{g}_{ji}(\theta)$  and not  $p_i \mathbf{g}_{ij}(\theta)$ , in order for condition (29) to hold. The reason is explained in Appendix E of the Supplement.

The function (17) is different from the empirical likelihood function of Berger and Torres (2016) and Berger (2020), despite having some common features. In (17), the constraints involving  $\theta$  is a double sum, as opposed to a single sum in Berger and Torres (2016) and with customary empirical likelihood approaches (Owen, 1988). The double sum is weighted by  $(\mu_i \mu_j)^{-1}$ . In Berger and Torres (2016), we have single sum with weights  $\mu_i^{-1}$ . Furthermore, a design-based approach was adopted by Berger and Torres (2016). Here, we have a semi-parametric model-based approach. Most importantly, the double sum brings challenges in proving the self-normalising property (see Section 7 for more details).

Consistency cannot be straightforwardly derived from the asymptotic results of Berger and Torres (2016) and Qin and Lawless (1994). Empirical likelihood relies on a law-of-large-numbers result for the estimating function when  $\theta = \theta_0$ . This is naturally achieved with single sums, but requires more specific arguments with double sums. In Section 6, we establish point-wise and  $\sqrt{n}$ -consistency. Note that point-wise consistency is implicitly assumed in Berger and Torres (2016) and in Berger (2020).

The second constraint within (17) involving  $\delta_i$ , can be found in Berger and Torres (2016). This constraint is necessary to achieve consistency. It implicitly includes the leading constraint  $\sum_{i \in \mathcal{P}} \tau_i p_i = 1$ , as in Owen (1988), because it implies  $n \sum_{i \in \mathcal{P}} \tau_i \delta_{ih} p_i = n_h$  which gives  $\sum_{i \in \mathcal{P}} \tau_i p_i = 1$ , because  $\sum_{h=1}^H n_h = n$  and  $\sum_{h=1}^H \delta_{ih} = 1$ .

Kitamura et al. (2004) proposed multiplying  $\log(p_i)$  by  $\mathcal{K}(\mathbf{W}_i - \mathbf{W}_j)$  and trimming factors. On the other hand, with (17),  $\mathcal{K}(\mathbf{W}_i - \mathbf{W}_j)$  appears within  $\mathbf{g}_{ij}(\boldsymbol{\theta})$ . Unlike Kitamura et al. (2004), we use the traditional empirical likelihood function. Hence, usual empirical likelihood packages can be used to implement the approach proposed.

Using Lagrangian multipliers, we have

$$\ell(\boldsymbol{\theta}) = \sum_{i \in \mathcal{P}} \tau_i \log \{ n \hat{p}_i^*(\boldsymbol{\theta}) \}, \quad (19)$$

where

$$\hat{p}_i^*(\boldsymbol{\theta}) := n^{-1} \left\{ 1 + \mu_i^{-1} \mathbf{c}_i^*(\boldsymbol{\theta})^\top \hat{\boldsymbol{\eta}}^*(\boldsymbol{\theta}) \right\}^{-1}, \quad (20)$$

$$\mathbf{c}_i^*(\boldsymbol{\theta}) := \{ \mathbf{c}_i^\top, \hat{\mathbf{g}}_i(\boldsymbol{\theta})^\top \}^\top, \quad (21)$$

$$\mathbf{c}_i := N n^{-1} \mu_i \boldsymbol{\delta}_i \quad (22)$$

and

$$\hat{\mathbf{g}}_i(\boldsymbol{\theta}) := \frac{1}{N} \sum_{j \in \mathcal{P}} \tau_j \frac{\mathbf{g}_{ji}(\boldsymbol{\theta})}{\mu_j}. \quad (23)$$

The quantity  $\hat{\boldsymbol{\eta}}^*(\boldsymbol{\theta})$  is such that the constraint

$$n \sum_{i \in \mathcal{P}} \tau_i \frac{p_i}{\mu_i} \mathbf{c}_i^*(\boldsymbol{\theta}) = \mathbf{C}^*,$$

holds, where

$$\mathbf{C}^* := (\mathbf{C}^\top, \mathbf{0}^\top)^\top \quad (24)$$

and

$$\mathbf{C} := N n^{-1} \mathbf{n}_{st}. \quad (25)$$

The quantity  $\hat{\boldsymbol{\eta}}^*(\boldsymbol{\theta})$  can be computed using a modified Newton-Raphson approach. When  $N$  is unknown, it can be replaced by its unbiased estimate  $\hat{N}_\mu := \sum_{i \in \mathcal{P}} \tau_i \mu_i^{-1}$  within (22) and (25). This does not affect (19).

The ‘maximum empirical likelihood point estimator’ of  $\boldsymbol{\theta}_0$  is defined as the vector  $\hat{\boldsymbol{\theta}}$  which maximises the function (17); that is,

$$\hat{\boldsymbol{\theta}} := \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}). \quad (26)$$

The following Theorem shows that  $\hat{\boldsymbol{\theta}}$  is the solution to the  $d_\theta$  estimating equations, unlike the approaches of Kitamura et al. (2004) and Donald et al. (2003).

**Theorem 2.** *We have*

$$\hat{\mathbf{G}}_\mu(\boldsymbol{\theta}) = \mathbf{0}_{d_\theta} \quad \text{if and only if } \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \quad (27)$$

where  $\hat{\boldsymbol{\theta}}$  is defined by (26) and

$$\hat{\mathbf{G}}_\mu(\boldsymbol{\theta}) := \frac{1}{N^2} \sum_{i \in \mathcal{P}} \tau_i \sum_{j \in \mathcal{P}} \tau_j \frac{\mathbf{g}_{ji}(\boldsymbol{\theta})}{\mu_i \mu_j}. \quad (28)$$

The proof of Theorem 2 can be found in the Appendix B of the Supplement. Note that (28) is a sample analogue of (16). Thus, we have that  $\hat{\boldsymbol{\theta}}$  is a natural estimator of  $\boldsymbol{\theta}_0$ , because (16) implies  $\mathbb{E}\{\hat{\mathbf{G}}_\mu(\boldsymbol{\theta})\} = \mathbf{0}_{d_\theta}$ , if and only if  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . The weights  $(\mu_i \mu_j)^{-1}$  ensure consistency (see Section 6) under the selection process considered. The empirical likelihood point estimator can be computed directly from (28) without invoking (17). The main advantage of empirical likelihood is the fact that (17) can be used for testing (see Section 7) and for showing point-wise consistency (see Section 6).

## 6. Regularity conditions and $\sqrt{n}$ -consistency

The quantity  $n$  is a constant which tends to infinity; that is, we consider a sequence of constants  $\{n_{[t]} : n_{[t]} < n_{[t+1]}, n_{[t]} < N_{[t]}, \forall t > 0\}$ ; where  $\{N_{[t]} : N_{[t]} < N_{[t+1]}\}$  denotes a sequence of population sizes. Thus  $t \rightarrow \infty$  implies  $n_{[t]} \rightarrow \infty$  and  $N_{[t]} \rightarrow \infty$ . Hereafter, the index  $t$  is removed, for simplicity. We consider that the number of strata  $H$  is bounded. We assume  $N_h/N$  and  $n_h/n$  bounded and bounded away from zero sequences, where  $n_h$  is the number of unit selected from  $\mathcal{P}_h$  and  $N_h := \sum_{i \in \mathcal{P}} \delta_{ih}$  (see Section 2). The quantity  $\mathcal{O}_p(a)$  and  $\mathcal{o}_p(a)$  denote random vectors such that  $\|\mathcal{O}_p(a)\| = O_p(a)$  and  $\|\mathcal{o}_p(a)\| = o_p(a)$ , where  $\|\cdot\|$  is the Frobenius norm. The  $\sqrt{n}$ -consistency of  $\hat{\boldsymbol{\theta}}$  can be established under the regularity conditions outlined in this Section.

We assume that there exists constants  $\lambda^\Sigma$  and  $\lambda^\nabla$  such that for  $n$  large enough,

$$\mathbb{P}\left[\inf_{\boldsymbol{\theta} \in \mathcal{B}_n} \lambda_{\min}\left\{\frac{n}{N^2} \hat{\boldsymbol{\Sigma}}_{cc}^*(\boldsymbol{\theta})\right\} \geq \lambda^\Sigma > 0\right] \rightarrow 1, \text{ where } \mathcal{B}_n := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq b_n\} \quad (29)$$

and

$$\mathbb{P}\left[\lambda_{\min}\{\hat{\nabla}(\boldsymbol{\theta}_0)^\top \hat{\nabla}(\boldsymbol{\theta}_0)\} \geq \lambda^\nabla > 0\right] \rightarrow 1, \quad (30)$$

where  $\lambda_{\min}\{\mathbf{A}\}$  denotes the smallest eigenvalue of a symmetric matrix  $\mathbf{A}$ . The quantity  $b_n$  denotes an arbitrary sequence tending to zero and such that  $nb_n^2 \rightarrow \infty$ . Here,

$$\hat{\boldsymbol{\Sigma}}_{cc}^*(\boldsymbol{\theta}) := \sum_{i \in \mathcal{P}} \tau_i \mu_i^{-2} \mathbf{c}_i^*(\boldsymbol{\theta}) \mathbf{c}_i^*(\boldsymbol{\theta})^\top \in \mathbb{R}^{r \times r} \quad (31)$$

and

$$\hat{\nabla}(\boldsymbol{\theta}) := \frac{\partial \hat{\mathbf{G}}_\mu(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{d_\theta \times d_\theta}. \quad (32)$$

The maximum eigenvalue of  $nN^{-2} \hat{\boldsymbol{\Sigma}}_{cc}^*(\boldsymbol{\theta})$  is also assumed finite  $\forall \boldsymbol{\theta} \in \mathcal{B}_n$ .

We assume that there exists  $a_1$  and  $a_2$  such that

$$\mathbb{P}(0 < a_1 \leq \|\hat{\nabla}(\boldsymbol{\theta}_0)\| \leq a_2 < \infty) \rightarrow 1, \quad (33)$$

$$\hat{\nabla}(\boldsymbol{\theta}) \text{ is continuous within } \boldsymbol{\Theta}, \quad (34)$$

$$\frac{\partial \hat{\nabla}(\boldsymbol{\theta})_k}{\partial \boldsymbol{\theta}} = \mathcal{O}_p(1) \text{ uniformly for } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \text{ where } \hat{\nabla}(\boldsymbol{\theta})_k \text{ denotes the } k\text{-th row of } \hat{\nabla}(\boldsymbol{\theta}), \quad (35)$$

$$\mathbb{E}(\|\mathbf{g}_{ji}(\boldsymbol{\theta})\|^4) < \infty \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and } \forall i, j, \quad (36)$$

$$\mathbb{P}\left\{N \pi(\mathbf{W}, \mathbf{Z}) \geq \zeta_1 > 0\right\} = 1 \quad (37)$$

and

$$\mathbb{P}\left\{N \pi(\mathbf{W}, \mathbf{Z}) \leq \zeta_2\right\} = 1, \quad (38)$$

where  $\zeta_1$  and  $\zeta_2$  denote constants that do not vary as  $n \rightarrow \infty$ .

For condition (29) to hold, we need to use  $p_i \mathbf{g}_{ji}(\boldsymbol{\theta})$  within (17). More details can be found in Appendix E of the Supplement. Inequality (30) is a mild condition on the Hessian, which appears under a slightly different form, but closely related to that usually found in the standard frameworks. Condition (37) ensures that  $\pi(\mathbf{W}, \mathbf{Z})$  is not disproportionately small compared to  $N^{-1}$ . Conditions (37) and (38) can be found in Isaki and Fuller (1982). See Appendix E for more details.

Lemma B.6 in the supplement, establishes the point-wise consistency of the estimator; that is  $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \mathcal{o}_p(1)$ . Theorem 3 shows the  $\sqrt{n}$ -consistency.

**Theorem 3.** *Under (29)–(38), we have that  $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \mathcal{O}_p(n^{-\frac{1}{2}})$ .*

The proof can be found in Appendix B of the supplement. Note that  $n/N \rightarrow 0$  is not required in Theorem 3; i.e.  $\sqrt{n}$ -consistency still holds even when  $n/N$  does not tend to zero.



## 7. Hypotheses tests and confidence intervals based on the empirical likelihood ratio statistics

Since, we have a double sum, we cannot apply the results of Berger and Torres (2016) directly. Furthermore, the double sums is challenging with unequal probability sampling, because the observations are not independent, even under with-replacement sampling (see (9)). The “*empirical likelihood ratio function*” is defined by the following function.

$$\hat{r}(\boldsymbol{\theta}) := -2 \ell(\boldsymbol{\theta}). \quad (39)$$

Here, the challenge is to derive the asymptotic properties which guarantee that (39) follows a  $\chi^2$ -distribution asymptotically, when  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . The key idea is to approximate the empirical likelihood ratio function, by a quadratic function involving only single sums (see Theorem C.1 in the Supplement). In Appendix C of the Supplement, we show that

$$\hat{r}(\boldsymbol{\theta}_0) \xrightarrow{d} \chi_{d_\theta}^2, \quad (40)$$

as  $n/N \rightarrow 0$ , where  $\chi_{d_\theta}^2$  denotes the  $\chi^2$ -distribution with  $d_\theta$  degrees of freedom. Thus,  $\hat{r}(\boldsymbol{\theta}_0)$  is an ancillary pivotal statistics that can be used for statistical tests on  $\boldsymbol{\theta}_0$  and for constructing a confidence region  $\{\boldsymbol{\theta} : \hat{r}(\boldsymbol{\theta}) \leq \chi_{d_\theta}^2(\alpha)\}$ , where  $\chi_{d_\theta}^2(\alpha)$  denotes the upper  $\alpha$ -quantile of  $\chi_{d_\theta}^2$ . This may be computationally intensive when  $d_\theta$  is large.

It is often necessary to construct confidence intervals for scalar components of  $\boldsymbol{\theta}_0$  or to test a sub-vector of  $\boldsymbol{\theta}_0$ . This can be achieved by profiling  $\hat{r}(\boldsymbol{\theta})$ . Let  $\boldsymbol{\theta}^\dagger$  denote a sub-vector of  $\boldsymbol{\theta}$ . Let  $\boldsymbol{\vartheta} \in \mathbb{R}^{d_\theta}$  be the components of  $\boldsymbol{\theta}$  which are not part of  $\boldsymbol{\theta}^\dagger$ ; say  $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\dagger\top}, \boldsymbol{\vartheta}^\top)^\top$ . The ‘*profile empirical likelihood ratio function*’ is defined by

$$\hat{r}(\boldsymbol{\theta}^\dagger)_{\min} := \min_{\boldsymbol{\vartheta} \in \Xi} \hat{r}(\boldsymbol{\theta}), \quad (41)$$

where  $\Xi$  denotes the parameter space of  $\boldsymbol{\vartheta}$ .

In Appendix C of the Supplement, we show that

$$\hat{r}(\boldsymbol{\theta}_0^\dagger)_{\min} \xrightarrow{d} \chi_{d_{\theta^\dagger}}^2. \quad (42)$$

When  $\boldsymbol{\theta}_0^\dagger$  is scalar ( $d_{\theta^\dagger} = 1$ ), the  $\alpha$  confidence interval for  $\boldsymbol{\theta}_0^\dagger$  is  $\{\boldsymbol{\theta}^\dagger : \hat{r}(\boldsymbol{\theta}^\dagger)_{\min} \leq \chi_1^2(\alpha)\}$ . The bounds can be computed with a root-finding algorithm, such as the method of Brent (1973) and Dekker (1969).

The asymptotic normality of  $\hat{\boldsymbol{\theta}}$  can be justified by assuming

$$n^{\frac{1}{2}} \hat{\mathbf{G}}_\mu^\circ(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_{d_\theta}(\mathbf{0}, \boldsymbol{\Omega}), \quad (43)$$

in distribution as  $n \rightarrow \infty$ , where  $\mathcal{N}_{d_\theta}(\cdot, \cdot)$  denotes the multivariate normal distribution and  $\boldsymbol{\Omega}$  is the limit of the variance of  $n^{\frac{1}{2}} \hat{\mathbf{G}}_\mu^\circ(\boldsymbol{\theta}_0)$ , where

$$\hat{\mathbf{G}}_\mu^\circ(\boldsymbol{\theta}_0) := \frac{1}{N} \sum_{i \in \mathcal{P}} \tau_i \mu_i^{-1} \mathbf{g}_i^\circ(\boldsymbol{\theta}_0), \quad \text{with} \quad \mathbf{g}_i^\circ(\boldsymbol{\theta}_0) := \frac{1}{N} \sum_{j \in \mathcal{P}} \mathbf{g}_{ji}(\boldsymbol{\theta}_0). \quad (44)$$

The asymptotic results of Prášková and Sen (2009) and Feller (1971) can be used to justify (43).

**Theorem 4.** Under (43) and assuming that  $\hat{\mathbf{V}}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{\nabla}$ , where  $\mathbf{\nabla}$  is some positive definite matrix, we have

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_{d_\theta}(\mathbf{0}, \mathbf{\nabla}^{-1} \boldsymbol{\Omega} \mathbf{\nabla}^{-1}). \quad (45)$$

The proof can be found in Appendix C. Wald test statistics and their confidence intervals can be derived from the variance estimator of  $\hat{\boldsymbol{\theta}}$ :

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) := \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})^{-1} \hat{\mathbf{V}}\{\hat{\mathbf{G}}_\mu^\circ(\boldsymbol{\theta}_0)\} \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})^{-1}, \quad (46)$$

where

$$\widehat{\mathbf{V}}\{\widehat{\mathbf{G}}_{\mu}^{\circ}(\theta_0)\} := \frac{1}{N^2} \sum_{i \in \mathcal{P}} \tau_i \frac{1}{\mu_i^2} \widehat{\boldsymbol{\varepsilon}}_i \widehat{\boldsymbol{\varepsilon}}_i^{\top}, \quad (47)$$

is a variance estimator of  $\widehat{\mathbf{G}}_{\mu}^{\circ}(\theta_0)$ . Here,  $\widehat{\boldsymbol{\varepsilon}}_i := \widehat{\mathbf{g}}_i(\widehat{\boldsymbol{\theta}}) - \widehat{\boldsymbol{\Sigma}}_{cg}^{\top} \boldsymbol{\Sigma}_{cc}^{-1} \mathbf{c}_i$ ,

$$\widehat{\boldsymbol{\Sigma}}_{cg} := \sum_{i \in \mathcal{P}} \tau_i \frac{1}{\mu_i^2} \mathbf{c}_i \widehat{\mathbf{g}}_i(\widehat{\boldsymbol{\theta}})^{\top} \quad \text{and} \quad \boldsymbol{\Sigma}_{cc} := \sum_{i \in \mathcal{P}} \tau_i \frac{1}{\mu_i^2} \mathbf{c}_i \mathbf{c}_i^{\top}. \quad (48)$$

The estimator (47) can be justified by using Theorem C.2 in the Supplement, after replacing  $\mathbf{g}_i^{\circ}(\theta_0)$  by  $\widehat{\mathbf{g}}_i(\widehat{\boldsymbol{\theta}})$  within  $\widehat{\boldsymbol{\Sigma}}_{ee}^{\circ}(\theta_0)$ . Proving the consistency of (46) is beyond the scope of this paper.

## 8. Simulation studies

The simulation studies focus on three situations: (i) endogenous selection in Section 8.1 (ii) transformation models in Section 8.2, and (iii) linear models with endogenous covariates in Section 8.3. We consider  $N = 50\,000$  and  $n = 500$  or 200. We consider  $\mathbf{W} \in \mathbb{R}$  to be exogenous or instrumental. We use the exponential kernel function for  $\mathcal{K}(\cdot)$ :

$$\mathcal{K}(W_i - W_j) = \exp\left\{-\frac{(W_i - W_j)^2}{2\hbar^2 \sigma_W^2}\right\}, \quad (49)$$

where  $\sigma_W$  is a proxy for the standard deviation of  $W$ . Here,  $\hbar$  denotes a bandwidth. We consider  $\hbar = 1$  and  $\hbar = n^{-1/5}$ .

### 8.1. Linear models with endogenous selection

In this Section, we show that endogenous selection can produce biased estimates, whereas the proposed estimator is less biased. Consider a model with heteroscedastic errors,

$$Y = \alpha_0 + \beta_0 W + W^{\frac{1}{2}} \varepsilon, \quad W \sim \Gamma(\text{shape} = 3, \text{scale} = 2), \quad (50)$$

and  $\varepsilon \sim \mathcal{N}(0, \sigma^2 = 16)$ . The parameter is  $\boldsymbol{\theta}_0 = (\alpha_0, \beta_0)^{\top} = (1, 3)^{\top}$ . The correlation between  $Y$  and  $W$  is approximately 0.73. The endogenous selection is achieved by using the standard, single-stratum, randomised systematic technique (e.g. Tillé, 2006, §7.2), with  $\mu_i \propto \xi_i$ , where  $\xi_i \geq 1$  and given by

$$\xi_i := 1 + \xi_i^* - \min_{i \in \mathcal{P}} \{\xi_i^*\}, \quad \text{with} \quad \xi_i^* := 1 + \rho Y_i + u_i, \quad (51)$$

and  $u_i \sim \mathcal{N}\{0, \sigma^2 = (1 - \rho^2)\sigma_Y^2\}$ , where  $\sigma_Y^2$  denotes the variance of  $Y_i$ . The usual method for computing  $\mu_i$  proportional to  $\xi_i$  can be found in Tillé (2006, §2.10). The quantity  $\rho$  denotes the correlation between  $\xi_i$  and  $Y$ . The value of  $\rho$  specifies the level of endogenous selection, since a large (small) value of  $\rho$  implies that  $\pi(\mathbf{W}, \mathbf{Z})$  and  $Y$  are highly (slightly) correlated, because  $\mu_i \propto \xi_i$ . Different values of  $\rho$  will be considered.

In Table 1, we have the relative efficiencies, the relatives biases and the observed coverages based on 1000 replicates. The empirical likelihood (EL) approach proposed is compared with the customary “ordinary least squares” (OLS) approach. The relative efficiencies are defined as the root-mean square errors of EL divided by those of OLS estimates. The EL is more efficient in all cases because it accounts for endogenous selection and heteroscedasticity. The OLS approach gives less accurate estimates and low coverages rates for  $\beta_0$ . The relative efficiency decreases with the level of endogenous selection given by  $\rho$ . Not surprisingly, when the selection is less endogenous ( $\rho = 0.1$ ), both approaches have almost the same efficiency, but with a very low coverage rate for  $\beta_0$  with the OLS approach. Most of the coverage rates of the EL approach are not significantly different from the nominal level (95%). The coverages of the Wald-type confidence interval based on (46) (column “Wald”) are similar to those observed for EL.

### 8.2. Transformation models

We consider the Box-Cox model with heteroscedastic or homoscedastic errors,

$$T(Y, \lambda_0) = \alpha_0 + \beta_0 W + W^a \varepsilon, \quad \text{with} \quad W \sim \mathcal{N}(1, \sigma^2 = 0.04), \quad (52)$$

**Table 1**

Model (50). Relative efficiency (Rel. Eff.), relative biases and coverages rate based on (41). Bandwidth  $\hat{h} = 1$ . Relative efficiency =  $\sqrt{MSE(EL)/MSE(OLS)}$ .  $n = 500$ . 1000 replicates.

| Parameter                | $\rho$ | Rel. Eff. | Relative bias (%) |      | Coverage rate (%) |       |       |
|--------------------------|--------|-----------|-------------------|------|-------------------|-------|-------|
|                          |        |           | OLS               | EL   | OLS               | Wald  | EL    |
| Intercept ( $\alpha_0$ ) | 0.1    | 0.97      | -1.2              | -0.3 | 95.0              | 96.1  | 96.3  |
|                          | 0.5    | 0.94      | 10.8              | -3.8 | 94.6              | 95.6  | 95.6  |
|                          | 0.7    | 0.90      | 32.6              | -0.4 | 94.2              | 95.1  | 95.6  |
|                          | 0.9    | 0.78      | 66.7              | -2.5 | 88.3†             | 94.5  | 94.2  |
| Slope ( $\beta_0$ )      | 0.1    | 0.95      | 1.0               | 0.0  | 88.1†             | 96.6† | 96.5† |
|                          | 0.5    | 0.79      | 3.9               | 0.3  | 78.6†             | 95.2  | 95.2  |
|                          | 0.7    | 0.73      | 4.8               | 0.1  | 72.5†             | 95.5  | 95.8  |
|                          | 0.9    | 0.67      | 6.0               | 0.1  | 63.6†             | 94.3  | 94.4  |

† Coverages rates significantly different from 95%: p-value  $\leq 0.05$ .

**Table 2**

Model (52). Rejection rates (%) of the empirical likelihood test  $H_0: \theta_0 = (\lambda_0, -1, 1)^\top$ , based upon (40). The nominal level is 5%.  $\hat{h} = 1$ .  $n = 500$ . 1000 replicates.

| $\rho$ | $\lambda_0 = 0.5$ |           | $\lambda_0 = 0.8$ |           |
|--------|-------------------|-----------|-------------------|-----------|
|        | Homosc.           | Heterosc. | Homosc.           | Heterosc. |
|        | $a = 0$           | $a = 1$   | $a = 0$           | $a = 1$   |
| 0.1    | 4.6               | 6.3       | 5.6               | 6.0       |
| 0.5    | 5.2               | 6.6†      | 5.7               | 5.0       |
| 0.7    | 5.3               | 5.2       | 4.8               | 5.1       |

† Rejection rate significantly different from 5%: p-value  $\leq 0.05$

where, for any  $y > 0$ ,  $T(y, \lambda) := (y^\lambda - 1)\lambda^{-1}$  if  $\lambda \neq 0$  and  $T(y, 0) := \log(y)$  otherwise. Here,  $a = 1$  or 0. Here,  $\epsilon \sim \text{Beta}(2, 5)\sqrt{1.568} - \sqrt{0.128}$ , such that  $\mathbb{E}(\epsilon) = 0$  and the correlation between  $Y$  and  $W$  is approximately 0.7. The parameter is  $\theta_0 = (\lambda_0, \alpha_0, \beta_0)^\top = (\lambda_0, -1, 1)^\top$ , with  $\lambda_0 = 0.5$  or 0.8. The quantity  $a$  is not a parameter. We have heteroscedastic or homoscedastic errors by setting  $a = 0$  or  $a = 1$ . We use the same endogenous selection as in Section 8.1, with  $\mu_i \propto \xi_i$  defined by (51).

In Table 2, we have the rejection rates (%) of the empirical likelihood test  $H_0: \theta_0 = (\lambda_0, -1, 1)^\top$ , based upon (40). These rates are not significantly different from the nominal level 5%, except in one case, we observe a slightly higher rate of 6.6% which is significantly different from 5%. The relative efficiency and relative biases can be found in Appendix E of the Supplement.

### 8.3. Linear models with endogenous covariates

The following simulation study illustrates the superiority of our approach over the “two-stage least squares” approach. Let  $\mathbf{Y} = (\mathcal{Y}, X)^\top$ , where  $\mathcal{Y}, X \in \mathbb{R}$ . Consider the linear model

$$\mathcal{Y} = \alpha_0 + \beta_0 X + \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad (53)$$

where the endogenous covariate  $X$  is given by

$$X = (1 - \beta_0)^{-1} \{ \alpha_0 + \Psi(W) + \epsilon \}, \quad \text{with } W \sim \mathcal{N}(0, 1). \quad (54)$$

Here,  $W \in \mathbb{R}$  is the instrumental variable. The function  $\Psi(W)$  and the values of  $\sigma_\epsilon$  considered are

**Table 3**

Model (53). Relative efficiency (Rel. Eff.), relative biases (RB) and coverages rate based on (41). Bandwidth  $\hat{h} = 1$  and  $\hat{h} = n^{-1/5}$ . Relative efficiency =  $\sqrt{MSE(EL)/MSE(2SLS)}$ , with  $\hat{h} = 1$ .  $n = 500$ .

| $\Psi(W)$ | $\rho$ | $\alpha_0$ |           |             |                    |           | $\beta_0$ |           |             |                    |           |
|-----------|--------|------------|-----------|-------------|--------------------|-----------|-----------|-----------|-------------|--------------------|-----------|
|           |        | Cov. 2SLS  | Cov. Wald | Cov. EL     |                    | Rel. Eff. | Cov. 2SLS | Cov. Wald | Cov. EL     |                    | Rel. Eff. |
|           |        |            |           | $\hat{h}=1$ | $\hat{h}=n^{-1/5}$ |           |           |           | $\hat{h}=1$ | $\hat{h}=n^{-1/5}$ |           |
| Linear    | 0.1    | 94.6       | 93.9      | 93.8        | 93.7               | 1.15      | 94.0      | 94.2      | 93.2†       | 93.3†              | 1.16      |
|           | 0.5    | 95.3       | 94.8      | 94.4        | 94.7               | 1.13      | 95.1      | 94.5      | 93.8        | 94.3               | 1.15      |
|           | 0.7    | 94.3       | 94.1      | 93.6†       | 93.4†              | 1.14      | 94.2      | 94.1      | 93.3†       | 94.0               | 1.17      |
| Quad.     | 0.1    | 97.4†      | 94.0      | 93.7        | 93.9               | 0.01      | 97.6†     | 93.9      | 93.3†       | 94.0               | 0.01      |
|           | 0.5    | 97.4†      | 94.2      | 93.7        | 94.6               | 0.01      | 97.5†     | 93.8      | 93.4†       | 94.4               | 0.01      |
|           | 0.7    | 97.3†      | 94.6      | 94.2        | 95.1               | 0.01      | 97.3†     | 94.7      | 94.7        | 95.1               | 0.01      |
| Cyclic    | 0.1    | 95.5       | 94.6      | 94.2        | 94.5               | 0.56      | 94.8      | 94.1      | 93.8        | 93.8               | 0.53      |
|           | 0.5    | 94.9       | 95.5      | 95.0        | 94.5               | 0.58      | 94.3      | 95.4      | 95.0        | 94.8               | 0.56      |
|           | 0.7    | 94.8       | 94.4      | 93.9        | 94.4               | 0.58      | 94.6      | 94.7      | 94.1        | 95.0               | 0.56      |

† Coverages rates significantly different from 95%: p-value  $\leq 0.05$ .

(i) Linear case:  $\Psi(W) = W + 5$ , with  $\sigma_\varepsilon^2 = 1$ ,

(ii) Quadratic case:  $\Psi(W) = 0.25(W + 4)(W - 4)$ , with  $\sigma_\varepsilon^2 = 0.04$ ,

(iii) Cyclic case:  $\Psi(W) = \sin(6 + 0.6W^2)$ , with  $\sigma_\varepsilon^2 = 0.09$ .

The parameter is  $\theta_0 = (\alpha_0, \beta_0)^\top = (1, 0.5)^\top$ . Equations (53) and (54) can be viewed as basic simultaneous equations model (e.g. Johnston, 1963, §9), with (54) being its reduced form. However, we consider that (54) is unknown and only used to generate  $X$  with different associations with  $W$ . The conditional estimating equation is

$$\mathbb{E}\{\mathcal{Y} - (\alpha_0 + \beta_0 X) \mid W\} = 0, \quad a.s.$$

We have that  $\tilde{X} := \mathbb{E}(X \mid W) = (1 - \beta_0)^{-1} \{\alpha_0 + \Psi(W)\}$ , which is a non-linear function of  $W$ , when  $\Psi(W)$  is non-linear. We assume that  $\Psi(W)$  is unknown from inference, since in practice the relationship between  $\tilde{X}$  and  $W$  is usually unknown. The correlation between  $\mathcal{Y}$  and  $X$  is approximately 0.95 in all cases. The correlation between  $X$  and  $W$  is strong (0.70) with the linear case, negligible (0.01) with the quadratic case or moderate (0.45) with the cyclic case.

We use 3 endogenous strata defined by 3 quantiles classes based upon the variable  $\xi_i$  generated from (51), with  $Y = \mathcal{Y}$ . Within each strata, we use the same endogenous selection as in Section 8.1, with  $\mu_i \propto \xi_i$  defined by (51). Here,  $n_h$  is the nearest integer from  $nN_hN^{-1}$ , with  $N_h := \sum_{i \in \mathcal{P}} \delta_{ih}$ . Note that the strata are endogenous because they are based on  $\xi_i$  which is associated with  $\mathcal{Y}$ . The correlation  $\rho$  is also a measure of endogenous selection, as in Section 8.1.

“Two-stage least-squares” (2SLS) is based on fitted values  $\hat{X}_i$  of the linear regression

$$X = \vartheta_1 + \vartheta_W W + \varepsilon. \quad (55)$$

For 2SLS, we use the R function “`ts1s()`” from the library “`sem`”. We voluntarily use the linear model (55), despite the fact that the relationship between  $\tilde{X}$  and  $W$  may not be linear. The reason is to show that the 2SLS can be inconsistent, when we fail to capture the non-linear relationship. Our approach does not rely on modelling the relationship between  $X$  and  $W$ , and is consistent whatever the relationship.

The results based on 1000 replicates are given in Table 3. The “linear case” corresponds to the ideal situation tailored for 2SLS, since we have a strong correlation between  $X$  and  $W$  and the model (55) holds. The observed coverages for EL can be slightly less than the nominal level. EL is also slightly less efficient than 2SLS. The “quadratic case” is a situation where the correlation between  $X$  and  $W$  is weak, although there is a non-linear relationship between

$X$  and  $W$ . In this case, the coverage of 2SLS is significantly larger than 95%. EL is more efficient than 2SLS. With the “cyclic case”, we observe a loss of efficiency for 2SLS and the observed coverages are not significantly different from 95%. The coverages of the Wald-type confidence interval based on (46) (column “Wald”), are similar to those observed for EL. We also compared the effect of the bandwidth  $\hat{h}$  with the proposed approach. It seems that  $\hat{h} = n^{-1/5}$  provides better coverage rates. A detailed discussion about the efficiency and biases can be found in Appendix E of the Supplement.

## 9. Concluding remarks

The estimator is a solution of  $d_\theta$  estimating equations involving double sums (see (17) and (28)). This increases the complexity of asymptotic developments needed to prove the consistency and to derive the asymptotic distribution of the empirical likelihood ratio statistics. Furthermore, sampling theory mostly deals with single sums. Double sums are challenging for different reasons. With several strata, we have variances within and between strata. They involve high order selection probabilities (see proof of Lemma B.1 in Appendix D). The law-of-large-numbers cannot be straightforwardly assumed when deriving the asymptotic order of the Lagrangian parameter (see Lemmas B.1 and B.2 in Appendix B). The self-normalising property cannot be simply derived from an empirical likelihood approach involving single sums. The double sum within (17) makes it fundamentally different from the the empirical likelihood function of Berger and Torres (2016), Berger (2020) and Qin and Lawless (1994).

The parametric likelihood-based approach of Pfeiffermann et al. (1998) requires using the distribution of an error term and specification of the heteroscedasticity, which are usually unknown. The covariates need to be exogenous, rather than endogenous. Furthermore, the resulting maximum likelihood estimator can be based on cumbersome, or even insoluble, estimating equations, even for linear models. Confidence intervals are often based on ad-hoc bootstrap techniques, often only justified by simulation studies. Our approach is a model-based, semi-parametric approach which naturally takes endogeneity into account within the selection process and the covariates. Our approach is computationally simpler to implement and provides pivotal empirical likelihood ratio statistics for confidence interval, tests and model building.

Our results rely on the with-replacement assumption and on  $n/N \rightarrow 0$ , which are needed to ensure that the empirical likelihood ratio statistics is pivotal (see (40) and (42)). In a recent paper, Berger (2021) shows how the assumption  $n/N \rightarrow 0$  can be relaxed, to allow for large sampling fractions. When  $n/N \rightarrow 0$ , with and without-replacement sampling are asymptotically equivalent (e.g. Hájek, 1981, p.112). These assumptions should not be viewed as a drawback, because they are often implicitly assumed with inference based on survey data, without being properly stated. For example, the rescaled bootstrap technique relies on the same assumptions. Furthermore, the parametric approach of Pfeiffermann et al. (1998) also relies on  $n/N \rightarrow 0$ . Model-based approaches often require negligible  $n/N$  (e.g. Binder and Roberts, 2009; Chambers, Steel, Wang and Welsh, 2012). Most surveys are based on small  $n/N$ . The main advantage of the proposed approach is that it can deal with a class of models which are not covered by the traditional parametric approaches for survey data, as in Chambers et al. (2012). An extension to multi-stage selection can be found in the Appendix E of the Supplement.

## Acknowledgement

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P4-ID-PCE-2020-1112, within PNCDI III.

## Supplementary material

Appendix A contains the proof of the core result given by Theorems 1. Appendix B focuses on point estimation and consistency (Theorems 2 and 3). The Proof of (40), (42) and Theorem 4 are in Appendix C. Appendix D contains additional proofs of secondary Lemmas which can be found in previous Appendices. An extension to multi-stage designs, remarks, examples and additional simulation results, can be found in Appendix E.

## References

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

- Amemiya, T., 1977. The maximum likelihood and the nonlinear three stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica* 45, 955–968.
- Berger, Y.G., 2020. An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics* 72, 91–121. doi:10.1007/s10463-018-0681-x.
- Berger, Y.G., 2021. Unconditional empirical likelihood approach for analytic use of public survey data. Submitted manuscript .
- Berger, Y.G., Torres, O.D.L.R., 2016. An empirical likelihood approach for inference under complex sampling design. *Journal of the Royal Statistical Society Series B* 78, 319–341. doi:10.1111/rssb.12115.
- Binder, D.A., 1983. On the variance of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* 51, 279–292.
- Binder, D.A., Roberts, G., 2009. Design- and model-based inference for model parameters, in: Pfeiffermann, D., Rao, C.R. (Eds.), *Sample Surveys: Design, Methods and Applications*. Elsevier, Amsterdam. volume 29B of *Handbook of Statistics*, pp. 33–54.
- Brent, R.P., 1973. *Algorithms for Minimization without Derivatives*. Prentice-Hall ISBN 0-13-022335-2, New-Jersey.
- Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34, 305–334.
- Chambers, R.L., Steel, D.G., Wang, S., Welsh, A., 2012. *Maximum Likelihood Estimation for Sample Surveys*. Chapman and Hall/CRC, New York.
- Chang, J., Chen, S.X., Chen, X., 2015. High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics* 185, 283–304.
- Chaudhuri, S., Handcock, M.S., 2018. A conditional empirical likelihood based method for model parameter estimation from complex survey datasets. *Statistics and Applications* 16, 245–268.
- Chaudhuri, S., Handcock, M.S., Rendall, M.S., 2008. Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *Journal of the Royal Statistical Society Series B* 70, 311–328.
- Chen, S., Van Keilegom, I., 2009. A review on empirical likelihood methods for regression. *Test* 18, 415–447.
- Chen, X., Pouzo, D., 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics* 152, 46–60.
- Cosslett, S.R., 1993. Estimation from endogenously stratified samples, in: Rao, A.S.R., Rao, C.R. (Eds.), *Econometrics*. Elsevier, Amsterdam. volume 11 of *Handbook of Statistics*, pp. 1–43.
- Dekker, T.J., 1969. Finding a zero by means of successive linear interpolation, in: Dejon, B., Henrici, P. (Eds.), *Constructive Aspects of the Fundamental Theorem of Algebra*. Wiley-Interscience, London. *Handbook of Statistics*, pp. 37–489.
- Domínguez, M.A., Lobato, I.N., 2004. Consistent estimation of models defined by conditional moment restrictions. *Econometrica* 72, 1601–1615.
- Donald, S., Imbens, G., Newey, W., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics* 117, 55–93.
- Donald, S., Imbens, G., Newey, W., 2009. Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics* 152, 28–36.
- Durbin, J., 1953. Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society Series B* 15, 262–269.
- Feller, W., 1971. *An introduction to probability theory and its application* (2nd ed.). John Wiley, New York.
- Gabler, S., 1984. On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement. *Biometrika* 71, 171–175.
- Hájek, J., 1981. *Sampling from a Finite Population*. Marcel Dekker, New York.
- Hall, P., 1990. Pseudo-likelihood theory for empirical likelihood. *The Annals of Statistics* 18, 121–140.
- Hansen, L.P., Singleton, K.J., 1982. Generalized instrumental variable estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–1286.
- Isaki, C.T., Fuller, W.A., 1982. Survey design under the regression super-population model. *Journal of the American Statistical Association* 77, 89–96.
- Johnston, J., 1963. *Econometric Methods*. McGraw-Hill Book Company, New York.
- Kitamura, Y., Tripathi, G., Ahn, H., 2004. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica* 72, 1667–1714.
- Little, R., 1982. Models for nonresponse in sample surveys. *Journal of the American Statistical Association* 77, 237–249.
- Lohr, S.L., 2010. *Sampling: Design and Analysis*. Brooks/Cole, Boston.
- Newey, W.K., 1993. Efficient estimation of models with conditional moment restrictions, in: Maddala, G., Rao, C., Vinod, H. (Eds.), *Econometrics*. Elsevier, Amsterdam. volume 11 of *Handbook of Statistics*, pp. 2111–2245.
- Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–249.
- Pfeiffermann, D., Krieger, A., Rinott, Y., 1998. Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* 8, 1087–1114.
- Pfeiffermann, D., Rao, C., 2009a. *Handbook of Statistics* 29A. Elsevier, Amsterdam.
- Pfeiffermann, D., Rao, C., 2009b. *Handbook of Statistics* 29B. Elsevier, Amsterdam.
- Prášková, Z., Sen, P.K., 2009. Asymptotic in finite population sampling, in: Pfeiffermann, D., Rao, C. (Eds.), *Sample Surveys: Design, Methods and Applications*. Elsevier, Amsterdam. *Handbook of Statistics*, pp. 489–522.
- Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. *Annals of Statistics* 22, pp. 300–325.
- Rubin, D.B., 1976. Inference and missing data. *Biometrika* 63, 581–592.
- Smith, R.J., 2007. Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics* 138, 430–460.
- Sugden, R., Smith, T., 1984. Ignorable and informative designs in survey sampling inference. *Biometrika* 71, 495–506.
- Tillé, Y., 2006. *Sampling Algorithms*. Springer Series in Statistics, Springer, New York.