Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16, No. 1, 2018 (New Series), pp 245-268

A Conditional Empirical Likelihood Based Method for Model Parameter Estimation from Complex survey Datasets

Sanjay Chaudhuri[1] **and Mark S. Handcock**[2]
[1]*Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546*
[2]*Department of Statistics, University of California, Los Angeles, CA 90095–1554*

Received: January 22, 2018; Revised: March 06, 2018; Accepted: March 06, 2018

Abstract

We consider an empirical likelihood framework for inference for a statistical model based on an informative sampling design. Covariate information is incorporated both through the weights and the estimating equations. The estimator is based on conditional weights. We show that under usual conditions, with population size increasing unbounded, the estimates are strongly consistent, asymptotically unbiased and normally distributed. Our framework provides additional justification for inverse probability weighted score estimators in terms of conditional empirical likelihood. In doing so, it bridges the gap between design-based and model-based modes of inference in survey sampling settings. We illustrate these ideas with an application to an electoral survey.

Key words: Complex survey data; Design weights; Model parameter estimation; Conditional likelihood; Inverse probability weighted estimation; Design-based survey inference; Generalised linear models.

1 Introduction

Because of their easy interpretability, parametric models are popular in statistics for explaining natural phenomenon. These model parameters are usually estimated by maximising the so called likelihood function computed from preferably independent sample observations identically distributed according to the model in the population. In practice however, such data-sets are often unavailable. More often than not, practitioners are forced to work with data obtained from various surveys. In real world, surveys are complex. Observations are drawn according to informative designs and are accompanied by unequal sampling weights. Because of such complex sampling the observed data distribution varies from the distribution specified in the model. These sampling weights thus contain important information about the distribution in the population and cannot be

Corresponding Author: Sanjay Chaudhuri
E-mail: sanjay@stat.nus.edu.sg

ignored. In most situations, ignoring the weights leads to sevrely biased and/or inefficient estimators (Skinner et al., 1989).

There is a large literature dealing with the analysis of complex survey data. However, the incorporation of design weights in a parametric likelihood framework is difficult and many common approaches are not fully likelihood based. In most cases design unbiased estimators for large but finite population parameters are used (Narain, 1951; Horvitz and Thompson, 1952) and the model parameters are assumed to be close to this estimator.

Empirical likelihood (Owen, 2001) provides alternative to such design unbiased estimators. For data sampled with equal probabilities, empirical likelihood based procedures re-weighs the data points with unknown weights. These weights are then estimated from the data by maximising their product under various constraints. These constraints can involve unknown parameters which can also be estimated simultaneously (Qin and Lawless, 1994). The constraints can also represent known characteristics of the population obtained from census, registration data etc. (Chaudhuri et al., 2008). Historically, the term was coined and made popular by Owen (1988). However, it has been argued that precursor methods were available (see, for example, Hartley and Rao, 1968; Thomas and Grunkemeier, 1975).

Empirical likelihood based methods which take into account the design weights in the sample have been studied by several researchers. Chen and Sitter (1999) were motivated by the Horvitz-Thompson estimator in survey sampling and proposed a *pseudo-empirical likelihood*. In brief, their procedure estimates the total of the logarithm of unknown weights in the population using a design unbiased Horvitz-Thompson estimator. The weights are then determined by maximising this sum under various constraints. This procedure is based on an estimated likelihood not an observed one. Rao, Wu and colleagues (notably Chen et al. (2002), Wu and Rao (2006), Rao and Wu (2008), among others) study this method extensively and apply it to several design based surveys. The pseudo empirical likelihood can be re-interpreted as a "backward" Kullback-Leibler divergence of the unknown weights from the sampling weights. The distribution is specified by the choosing the weights that minimise this divergence. Wu (2004) discuss a similar minimised weighted entropy estimator.

In this article we develop a framework that results in a procedure which fundamentally differs from the pseudo-empirical likelihood. We consider an observed likelihood based on the conditional distribution of the observations given that the individuals were selected in the sample and estimate their distribution in the population using empirical likelihood. We are motivated by Pfeffermann and colleagues (e.g. Pfeffermann et al. (1998), Pfeffermann and Sverchkov (1999), Krieger and Pfeffermann (1992)) who used a similar but fully parametric procedure in modelling survey data. A previous instance of similar use in a more restrictive parametric set-up occurs in Patil and Rao (1978), where it has been implemented on size-biased sampling. We propose a empirical likelihood based semi-parametric approach here. The resulting conditional empirical likelihood is similar to Vardi (1985). However, he is motivated solely by the non-parametric estimation of the distribution from multiple samples obtained through different designs. The asymptotic properties of this non-parametric estimator have been studied by Gill et al. (1988). For complex survey data, a fully non-parametric estimation procedure has been studied by Chambers et al. (2003).

Use of empirical likelihood in complex data goes back to Qin (1993), who employed it in a two-sample testing problem, where only one sample was biased by the design. He showed that under certain conditions the empirical log-likelihood ratio has an asymptotic Chi-squared limit. A similar approach has been taken by Qin et al. (2002) to analyse data with non-ignorable non-response. Qin and Zhang (2007) use an empirical likelihood based method in observational studies where part of the response is missing. Calibration estimation using a similar empirical likelihood in Poisson sampling has been considered by Kim (2009).

We start with basic assumptions on the model, design variables and the sampling probabilities, which justifies the analytic form of the likelihood seen in Pfeffermann et al. (1998). In this endeavour, the sampling weights are interpreted as random variables depending on all observations of all design variables in the population. Our framework does not require one to make all the design variables available in the sample. Neither does it assume all observations of the design variables available in the sample are known. This general framework results in a composite likelihood of the data, which is then converted to an conditional empirical likelihood. The parameter estimates are obtained by maximising this likelihood under the constraints imposed by the model. It is seen that such estimates are strongly consistent and asymptotically normally distributed for the population distribution under usual regular conditions. We also provide estimators of the variances. We end by applying our method to estimating electoral success in the 2004 presidential election in United States of America.

2 Model and Design Specification

2.1 Basic Assumptions and Notations

We consider a "superpopulation" model with response Y, a set of auxiliary variables $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(p)}\}$ and a set of design variables $D = \{D^{(1)}, D^{(2)}, \ldots, D^{(q)}\}$. The superpopulation is same as the probability space under which the model and the random variables are defined. The set of events also includes all subsets of the set of integers, that is any funnction defined on the all subsets of integers is well defined. These functions are used in defining the sample. The population is comprised of N independent and identically distributed draws from the super-population model. We label the elements of the population by $\mathcal{P} = \{1, 2, \ldots, N\}$.

A random sample $\mathcal{S}$ of n observations is drawn from $\mathcal{P}$ according to a design depending on D and possibly on some unknown parameters (specified in Section 2.4). The available data does not contain all variables in D, only a subset $Z = \{Z^{(1)}, Z^{(2)}, \ldots, Z^{(m)}\}$ is supplied. Let $Z^c = D \setminus Z$. Variables in X and Y are not directly involved in the sampling design. We denote $V = Y \cup X \cup Z$ to be the $m + p + 1$ dimensional random vector observed in the data set. Further, we collect all the explanatory variables in the model in a set $A \subseteq V$. Suppose, $D_{\mathcal{P}}$, $X_{\mathcal{P}}$, $Y_{\mathcal{P}}$, $Z_{\mathcal{P}}$, $Z^c_{\mathcal{P}}$ denote the vectors and matrices of all observations of the corresponding variables on the population $\mathcal{P}$. For $S \subseteq \mathcal{P}$, V_S is the matrix of observations in S. $V_{\bar{S}}$ are the observations not in S, where $\bar{S} = \mathcal{P} \setminus S$.

Primary scientific interest focuses on the relationship between a response Y and the set of explanatory variables A. Examples of such models are generalised linear models (GLM) (McCullagh and Nelder, 1989). As an important special case, we consider joint models for Y and A, $P_\theta(Y, A)$, parametrised by θ. For example, for GLM $\mu(\theta) = A\theta$. We specify the broader class of applicable models in Section 2.2 below.

2.2 Model specification

Suppose F^0 is the distribution of V_1 in the population with density dF^0 w.r.t. a suitable measure. The relationship between the response Y and the set of auxiliary variables A is assumed to be specified by:

$$E_{F^0}\left[\psi_\theta\left(Y_1, A_1\right)\right] = 0. \tag{2.1}$$

Here $\psi = (\psi_1, \ldots, \psi_r)$ is a pre-specified function depending only on Y_1 and A_1 and some unknown parameter θ. we assume r is at least as big as the dimension of the parameter vector. There may be several choices for ψ (Qin and Lawless, 1994). For parametric models, such as the GLM considered in the introduction, the corresponding *score functions* $S_\theta\left(Y, A\right)$ are natural choices.

For simplicity we would assume that the function ψ is well behaved and that the Hessian matrix is continuous near the true value of the parameter (see A.6 in Section 5.1). This condition of course can be relaxed. Medians, quantiles etc. can also be estimated similarly.

2.3 Design Specification

The sample $\mathcal{S}$ is a random subset or random multiset (for sampling with replacement) of size n of $\mathcal{P}$. If the sample units are drawn according to a design, the sampling mechanism may not be *ignorable*. The observed distribution of V in the sample $\mathcal{S}$ may be different from its distribution in the population (or the superpopulation) and may depend on the particular sample selected. The likelihood of the parameter θ, based on the sample observations of V (ie. $V_\mathcal{S}$) differs from the likelihood based on $V_\mathcal{P}$ (ie. its population observations). In reality, it is almost impossible to specify the likelihood based on the observations of V from a non-ignorable unequal probability sample. The sample at hand would rarely contain all design variables, thus the actual design procedure cannot be determined. Even if the design mechanism can be determined, incorporating design information in the likelihood is anything but straightforward.

Pfeffermann et al. (1998) have introduced a sample likelihood of the parameter based on the population density of V conditional on the event of selection in the sample. This likelihood demands essentially no design information, other than the sampling probabilities of the sampled observations. It is not however immediately clear that the proposed sample likelihood is meaningful under the population distribution. Below, we first show the conditions under which the sample likelihood indeed is the true population conditional likelihood, using which interpretable inference about the model parameter can be drawn from the available sample. For simplicity, we consider only the subsets of $\mathcal{P}$ (ie. sampling without replacement) here. Description for multisets (ie. sampling with replacement) is similar.

For $S \subseteq \mathcal{P}$ suppose I_S is the random indicator function for $S \subseteq \mathcal{S}$. The sampled units are drawn according to a design depending on all observations of the set of design variables in the population (ie. $D_{\mathcal{P}}$). With $Pr_{\mathcal{P}}[\cdot]$ denoting the probability under the population (and by definition the superpopulation), this implies that, for any $S \subseteq \mathcal{P}$, the probability of selecting S in the sample is given by:

$$\pi_S = Pr_{\mathcal{P}}\left[I_S = 1 \mid D_{\mathcal{P}}\right]. \tag{2.2}$$

That is, π_S represents the (joint) conditional probability for inclusion of subset S in the sample given all observations of all design variables in the population. Being conditional probability, for each S, π_S is a function of $D_{\mathcal{P}}$, S, n, N and possibly some other parameters. That is, π_S is a random variable because of $D_{\mathcal{P}}$. Since I_S is a binary variable, its distribution is completely specified through π_S.

We first argue that our definition of π_S as a conditional expectation under the population distribution does not conflict with the notion of sampling from a finite population. To that end, we first assume that:

Assumption 1 (Conditional independence given the design.)**.** For all $S \subseteq \mathcal{P}$, under the population distribution, π_S is conditionally independent of $Y_{\mathcal{P}}$ and $X_{\mathcal{P}}$ given $D_{\mathcal{P}}$. That is

$$\pi_S \perp\!\!\!\perp (Y_{\mathcal{P}}, X_{\mathcal{P}}) \mid D_{\mathcal{P}}. \qquad \text{for all } S \subseteq \mathcal{P}. \tag{2.3}$$

The assumption ensures that π_S depends on the $Y_{\mathcal{P}}$ and $X_{\mathcal{P}}$ only through the design variables $D_{\mathcal{P}}$. Under Assumption 1 for all S, $E_{\mathcal{P}}\left[\pi_S \mid Y_{\mathcal{P}}, X_{\mathcal{P}}, D_{\mathcal{P}}\right] = E_{\mathcal{P}}\left[\pi_S \mid D_{\mathcal{P}}\right] = \pi\left(S, D_{\mathcal{P}}\right)$, for some function π. This function may depend on the sample and the population size, but does not depend directly on the distributions of $X_{\mathcal{P}}$ and $Y_{\mathcal{P}}$.

By construction, I_S is a well-defined random variable under the superpopulation structure. Its distribution is completely determined by π, which is specified by the practitioner or the sampling procedure used to obtain $\mathcal{S}$ from $\mathcal{P}$ and not automatically determined by the super-population model. A sampling design can be viewed as a list of π_S assigned ideally to every subset S of $\mathcal{P}$. That is, once the function π is specified, the first-order probabilities of selection for $\{i\}$, $i \in \mathcal{P}$ are given by $\pi_i = \pi\left(\{i\}, D_{\mathcal{P}}\right), i = 1, \ldots, n$. The second-order probabilities are similarly determined by $\pi_{ij} = \pi\left(\{i,j\}, D_{\mathcal{P}}\right)$. Higher order probabilities can be specified exactly the same way.

In some cases, π_S for each $S \subseteq \mathcal{P}$ is defined in advance and the design is constructed to ensure that the subset S is selected with probability π_S. As for example, by definition simple random sampling without replacement procedure ensures that each subset of size $\mid S \mid$ is chosen with pre-specified probability of $\binom{n}{|S|}/\binom{N}{|S|}$, for $\mid S \mid \leq n$ and zero otherwise. On the other hand, in some cases a partial list of π_S may be available. Various sampling designs are constructed to match these specified probabilities. The sampling probabilities of rest of the subsets are then design specified. As for example, for probability proportional to size (PPS) sampling the function π can be chosen to yield the target first-order probabilities of selection $\pi_{\{i\}} = Z_i^{(1)} / \sum_{i=1}^{n} Z_i^{(1)}$, for some positive random variable $Z^{(1)}$. Several procedures for PPS sampling to sample unit $\{i\}$ with a specified probability $\pi_{\{i\}}$ are known (see Brewer and Hanif (1983), Chaudhuri and Voss (1988), Tille (2006), among others). Each of these procedures have different higher order selection probabilities.

Once a list of π_S is specified for all $S \subseteq \mathcal{P}$, the physical act of sampling observations from the population $\mathcal{P}$ to the sample $\mathcal{S}$ contributes no more to the statistical inference. That is, once we know the specification of π_S, the second randomisation and sampling from $\mathcal{P}$ to $\mathcal{S}$ can be subsumed under the superpopulation probability structure. The concept is akin to the *resampling* procedures popularly used in statistics. The difference as well as the difficulty for non-ignorable sampling is that the population resampled from is unobserved and we can draw only one *weighted* resample.

The concept that the actual sampling procedure can be subsumed in the probability structure of the superpopulation is valid even if Assumption 1 does not hold. From Assumption 1 and the definition of π_S as $E_{\mathcal{P}}\left[I_S \mid D_{\mathcal{P}}\right]$ we obtain the following result.

Lemma 2.1. Assumption 1 holds iff $\pi_S = \pi\left(S, D_{\mathcal{P}}\right)$.

Lemma 2.1 follows from the definition of conditional independence (Lauritzen, 1996). It further shows that, under Assumption 1, conditioning on $D_{\mathcal{P}}$ and the pair $(D_{\mathcal{P}}, \pi_S)$ is same. The following relationships can also be obtained from Assumption 1 and Lemma 2.1.

Lemma 2.2. Suppose $E_{\mathcal{P}}\left[\cdot\right]$ denotes the expectation under the population distribution. Under Assumption 1, for all $S \subseteq \mathcal{P}$, the following holds:

1. $E_{\mathcal{P}}\left[I_S \mid \pi_S\right] = \pi_S$.
2. $I_S \perp\!\!\!\perp D_{\mathcal{P}} \mid \pi_S$.

Pfeffermann et al. (1998) use the relationship in Conclusion 1. to justify their parametric likelihood (see below). The conditional independence relation in 2. is exactly the “Condition 1” in Sugden and Smith (1984), which implies that under Assumption 1 the set of joint probabilities $\{\pi_S \,:\, S \subseteq \mathcal{P}\}$ contains all design information and given the design the selection procedure (i.e., the actual dependence of π_S on $D_{\mathcal{P}}$) can be ignored for inference.

The assumption that the set of joint selection probabilities contains all information about the sampling mechanism is natural and facilitates analysis. In sample surveys, the probability of selecting an observation becomes unequal due to clustering, stratification, post-stratification, attrition, purposive “oversampling” and other non-response adjustments. In most cases, the published data does not contain all the design variables, thus the actual design procedure cannot be determined. Further, in many cases large data sets are constructed by merging several available data sets obtained from different surveys (e.g. Rendall et al. (2008); Tighe et al. (2010)). Typically, each survey is based on different designs dependent on different variables. A common design for the merged data set is mostly unavailable or may not be easy to specify, but weights from individual surveys can be used to provide information about the underlying designs.

Assumption 2 (Conditional independence given the sampling probabilities)**.** For all $S \subseteq \mathcal{P}$, under the population distribution, I_S is conditionally independent of $X_{\mathcal{P}}$, $Y_{\mathcal{P}}$ and $D_{\mathcal{P}}$ given π_S. That is,

$$I_S \perp\!\!\!\perp (X_{\mathcal{P}}, Y_{\mathcal{P}}, D_{\mathcal{P}}) \mid \pi_S \qquad \text{for all } S \subseteq \mathcal{P}. \tag{2.4}$$

Assumption 2 implies that inclusion depends on the $X_{\mathcal{P}}, Y_{\mathcal{P}}$ and $D_{\mathcal{P}}$ only through the joint inclusion probabilities. In particular, it implies

$$Pr_{\mathcal{P}}\left[I_S = 1 \mid X_{\mathcal{P}}, Y_{\mathcal{P}}, D_{\mathcal{P}}, \pi_S\right] = Pr_{\mathcal{P}}\left[I_S = 1 \mid \pi_S\right] = E_{\mathcal{P}}\left[I_S \mid \pi_S\right] = \pi_S.$$

Pfeffermann et al. (1998) make this assumption without stating it explicitly.

Lemma 2.3. For all $S \subseteq \mathcal{P}$, under Assumptions 1, Assumption 2 is equivalent to $I_S \perp\!\!\!\perp D_{\mathcal{P}} \mid \pi_S$ and $I_S \perp\!\!\!\perp (X_{\mathcal{P}}, Y_{\mathcal{P}}) \mid D_{\mathcal{P}}$.

The condition $I_S \perp\!\!\!\perp (X_{\mathcal{P}}, Y_{\mathcal{P}}) \mid D_{\mathcal{P}}$ is the basic design assumption of Scott (1977). According to Sugden and Smith (1984), any design which only depends on $D_{\mathcal{P}}$ should satisfy this condition.

Lemma 2.4. For all $S \subseteq \mathcal{P}$, Assumptions 1 and 2 imply the following conditional independence relationships.

1. $I_S \perp\!\!\!\perp V_S \mid \pi_S$,
2. $(I_S, \pi_S) \perp\!\!\!\perp (X_{\mathcal{P}}, Y_{\mathcal{P}}) \mid D_{\mathcal{P}}$ and
3. $I_S \perp\!\!\!\perp (X_S, Y_S) \mid D_{\mathcal{P}}$.

The statement 1. of Lemma 2.4 implies that $E_{\mathcal{P}}\left[I_S \mid V_S, \pi_S\right] = E_{\mathcal{P}}\left[I_S \mid \pi_S\right] = \pi_S$ for all $S \subseteq \mathcal{P}$. From this, following Pfeffermann and Sverchkov (2003) we obtain

$$\begin{aligned} Pr_{\mathcal{P}}\left[I_S = 1 \mid V_S\right] &= E_{\mathcal{P}}\left[I_S \mid V_S\right] = E_{\mathcal{P}}\left[E_{\mathcal{P}}\left[I_S \mid V_S, \pi_S\right] \mid V_S\right] = E_{\mathcal{P}}\left[\pi_S \mid V_S\right], \\ Pr_{\mathcal{P}}\left[I_S = 1, V_S\right] &= Pr_{\mathcal{P}}\left[I_S = 1 \mid V_S\right] Pr_{\mathcal{P}}\left[V_S\right] = E_{\mathcal{P}}\left[\pi_S \mid V_S\right] Pr_{\mathcal{P}}\left[V_S\right]. \end{aligned} \tag{2.5}$$

Equation (2.5) shows that under our assumptions the joint probability of selection of a subset S and the measurement of the variable V on S i.e. V_S can be expressed without modelling the design procedure explicitly. Expectation of the selection probability conditional on the observation is required. However, this conditional expectation is a function of the data observed in the sample. The joint probability does not depend on rest of the unobserved elements in the population.

A graphical representation of the conditional independencies in Assumptions 1 and 2 for $S = \mathcal{S}$ can be found in Figure 1. In the graph, $\pi_{\mathcal{S}}$ and $I_{\mathcal{S}}$ may depend on the whole of $Y_{\mathcal{P}}$ and $X_{\mathcal{P}}$ via $Z_{\mathcal{P}}^c$ which are not available in $\mathcal{S}$. Thus even though $\pi_{\mathcal{S}}$ does not depend on $Y_{\mathcal{S}}$ and $X_{\mathcal{S}}$ directly, in (2.5), $E_{\mathcal{P}}\left[\pi_{\mathcal{S}} \mid V_{\mathcal{S}}\right] \neq E_{\mathcal{P}}\left[\pi_{\mathcal{S}} \mid Z_{\mathcal{S}}\right]$ in general. Furthermore, the relation $Y_{\mathcal{S}} \perp\!\!\!\perp I_{\mathcal{S}} \mid A_{\mathcal{S}}$ does not hold. So the design ignorability condition of Pfeffermann et al. (1998) and Pfeffermann and Sverchkov (1999) is clearly not satisfied.

Note that our assumptions do not require $I_{\mathcal{S}}$ to be conditionally independent of $V_{\bar{\mathcal{S}}}$ given $V_{\mathcal{S}}$ (see Figure 1). Thus the observations are not missing at random (in the sense of Little and Rubin (2002)). However missing at random is an important special case, because the relationship, $I_{\mathcal{S}} \perp\!\!\!\perp (V_{\bar{\mathcal{S}}}, \pi_{\bar{\mathcal{S}}}) \mid (V_{\mathcal{S}}, \pi_{\mathcal{S}})$ holds.

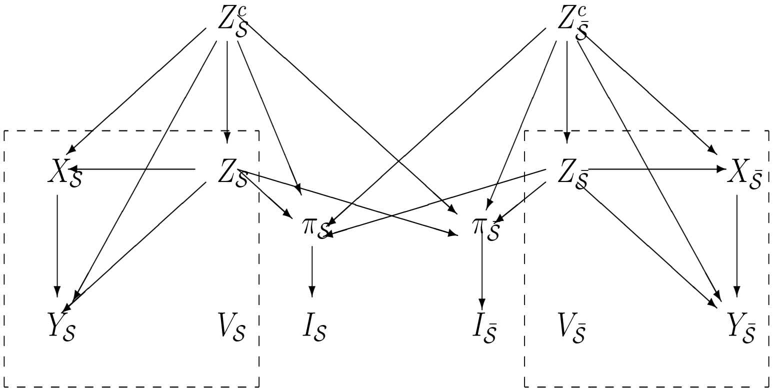


Figure 1: A graphical representation of the assumptions. The directed edges do not necessarily indicate a causal relationship.

Finally, we note that, Assumption 2 is sufficient but not necessary for (2.5) to hold. One of the conditions $I_{\mathcal{S}} \perp\!\!\!\perp V_{\mathcal{S}} \mid \pi_{\mathcal{S}}$ or $I_{\mathcal{S}} \perp\!\!\!\perp (X_{\mathcal{S}}, Y_{\mathcal{S}}) \mid D_{\mathcal{P}}$ would suffice. We could have alternatively assumed:

Assumption $2'$. For all $S \subseteq \mathcal{P}$, under the population distribution, I_S is conditionally independent of X_S and Y_S given $D_{\mathcal{P}}$. That is,

$$I_S \perp\!\!\!\perp (X_S, Y_S) \mid D_{\mathcal{P}} \qquad \text{for all } S \subseteq \mathcal{P}. \tag{2.6}$$

Unlike Assumption 2, however, Assumption $2'$ still allows $I_{\mathcal{S}}$ to be conditionally dependent on $(X_{\bar{\mathcal{S}}}, Y_{\bar{\mathcal{S}}})$ given $\pi_{\mathcal{S}}$ without violating Lemma 2.2. This will happen in very special situations where typically the information about the design available from $\pi_{\mathcal{S}}$ is incomplete and the design is potentially mis-specified. In most cases though Assumption 2 would be satisfied.

2.4 A composite likelihood for unequal probability sampling

Let the ith element in $\mathcal{S}$ be drawn with probability π_i (i.e. $\pi_{\{i\}}$), $i = 1, 2, \ldots, n,$. Suppose that π_i is positive for $i = 1, 2, \ldots, N$.

We consider the implication of (2.5) on each V_i (i.e. $V_{\{i\}}$), $i = 1, 2, \ldots, n$ selected in the sample. Let $F^{(i)}_{\mathcal{S}}$ be the conditional distribution of V_i given $\{i\} \subseteq \mathcal{S}$, with density $dF^{(i)}_{\mathcal{S}}$. Using Bayes' rule (Pfeffermann et al., 1998), (A.1) and (2.5) it follows that:

$$dF^{(i)}_{\mathcal{S}} = \frac{Pr_{\mathcal{P}}(I_{\{i\}} = 1, V_i)}{Pr_{\mathcal{P}}(I_{\{i\}} = 1)} = \frac{E_{\mathcal{P}}\left[\pi_i \mid V_i\right] dF^0(V_i)}{Pr_{\mathcal{P}}(I_{\{i\}} = 1)}, \tag{2.7}$$

where

$$Pr_{\mathcal{P}}(I_{\{i\}} = 1) = \int Pr_{\mathcal{P}}(I_{\{i\}} = 1, V_i) dV_i = \int E_{\mathcal{P}}\left[\pi_i \mid V_i\right] dF^0(V_i) dV_i. \tag{2.8}$$

We call the conditional inclusion probability $\nu_i \equiv E_{\mathcal{P}}\left[\pi_i \mid V_i\right]$ the *conditional visibility* for the ith element in the population and $\Upsilon_i \equiv \int \nu_i dF^0(V_i)dV_i = E_{\mathcal{P}}[\pi_i] = E_{F^0(V_i)}[\nu_i]$ the *visibility factor* for the ith element in the population (Patil and Rao, 1978). By substituting these expressions into (2.7) we obtain:

$$dF_{\mathcal{S}}^{(i)} = \frac{\nu_i dF^0(V_i)}{\Upsilon_i}. \tag{2.9}$$

To specify $dF_{\mathcal{S}}^{(i)}$ in (2.9) it is typically necessary to model the conditional visibility ($E_{\mathcal{P}}\left[\pi_i \mid V_i\right]$) and the distribution of V_i in the population ($dF^0(V_i)$). Both of these models may depend on unknown parameters. We denote the parameter for $dF^0(V_i)$ by θ and that for the model for $E_{\mathcal{P}}\left[\pi_i \mid V_i\right]$ by α.

The composite likelihood for α and θ, using all V_i, $i = 1, 2, \ldots, s$ can now be constructed as:

$$L\left(V, \alpha, \theta\right) = \prod_{i=1}^{n} dF_{\mathcal{S}}^{(i)}. \tag{2.10}$$

It is similar to the sample likelihood of Pfeffermann and Sverchkov (2003). Note that (2.10) does not capture the dependence structure of the $F_{\mathcal{S}}^{(i)}$. It is a conditional likelihood if the units are drawn independently of each other, for example, via Poisson sampling. However, Pfeffermann, Krieger, and Rinott (1998) show that for several designs, and under fairly general conditions, the sampled observations in the conditional distribution are asymptotically independent as the population size $N \to \infty$. These results suggest that the (2.10) may be a useful surrogate for the conditional likelihood in these settings. This is also seen in the illustrative example presented in Section 6 below.

Notice that, (2.10) is invariant to the scale of π and ν, which can be specified up to an arbitrary positive scaling constant. We can estimate ν_i directly from the data if the conditional distribution of π_i given V_i in $\mathcal{S}$ is equal to that in $\mathcal{P}$. Otherwise, from Pfeffermann and Sverchkov (1999), we obtain $Pr_{\mathcal{S}}\left(\pi_i^{-1} \mid V_i\right) = \{\pi_i Pr_{\mathcal{P}}\left(\pi_i^{-1} \mid V_i\right)\}/E_{\mathcal{P}}\left[\pi_i \mid V_i\right]$. This implies:

$$E_{\mathcal{P}}\left[\pi_i \mid V_i\right] = \left[E_{\mathcal{S}}\left(\pi_i^{-1} \mid V_i\right)\right]^{-1}. \tag{2.11}$$

Thus when the sample and population distribution differ a model for π_i^{-1} in terms of V_i is sought. The required population expectation can be estimated from the reciprocal of fitted values of π_i^{-1} obtained from the model. It is however not clear how to check if the conditional distribution of π given V_i in the population and the sample are different. We reckon that in most cases use of (2.11) would be appropriate.

Pfeffermann et al. (1998) discuss a class of conjugate parametric models for the distribution of V_i and conditional distribution with π_i given V_i such that $dF_{\mathcal{S}}^{(i)}$ is in the same class as $dF^0(V_i)$. This avoids a complicated computation of Υ_i. However, estimation of θ is typically complex. The parameters in $dF_{\mathcal{S}}^{(i)}$ usually depends on both θ and α. This is uneconomical since estimates of α usually are not of primary interest. Furthermore, (2.10) may have multiple modes in θ and α (Pfeffermann and Sverchkov, 1999, 2003). Thus $\hat{\alpha}$ and $\hat{\theta}$ are estimated separately.

Typically, ν_i would only depend on a subset of variables in V which may be quite different from A. In particular, if the sample $\mathcal{S}$ was obtained by merging several sub-samples drawn from different designs, ν_i depend on the particular sample the ith observation belongs to. Such sample indicator variables usually would not be useful in modelling the response.

Parametric estimation of F^0 by maximising (2.10) has been discussed in Patil and Rao (1978). Vardi (1985); Gill et al. (1988) consider the corresponding non-parametric likelihood when $\nu = \pi$ and study the empirical distribution for biased sampling models in one dimension.

3 Empirical likelihood to incorporate sampling weights in parameter estimation

3.1 A Conditional Empirical likelihood based formulation

If F_0 is specified by a parametric family $\mathcal{F}_\theta$, a natural way to estimate θ is to maximise (2.10) over Θ. Direct maximisation of (2.10) however poses several problems. First of all, analytic expression of $dF_{\mathcal{S}}^{(i)}$ are available only if one restricts to conjugate families of distributions for V_i and ν_i. Outside this class Υ_i has to be computed numerically (see Pfeffermann and Sverchkov (2003)) which may be time consuming. Furthermore, correct specification of the parametric joint distribution of V_i is difficult in many situations, specially when V_i contains design variables.

An alternative is to use empirical likelihood (Owen, 2001) and estimate F^0 non-parametrically from the observed weighted sample and include all the available parametric information in the analysis.

For the empirical likelihood based formulation the following assumption on Υ is made.

Assumption 3 (Label-independence of the visibility factors)**.** Visibility factors do not depend on the population labels but only on the design variables, i.e. $\Upsilon_i = \Upsilon\left(D_{\mathcal{P}}\right) \equiv \Upsilon$.

Whether the visibility factors should depend on the individual labels have been discussed in survey sampling literature in the past. See e.g. Godambe (1975); Hartley (1975). Under Assumption 3 each element in the population and sample will have equal visibility factor. In view of Assumption 1, $\Upsilon_i = E_{\mathcal{P}}[\pi_i]$. Thus Assumption 3 implies $E_{\mathcal{P}}[\pi_i]$ are all equal in the population.

We introduce our conditional empirical likelihood. To that goal, suppose that, for each $F \in \mathcal{F}$, $w_i = F(\{V_i\})$ is the weight F assigns on V_i ($w_i = 0$ for all F continuous at V_i). Let Δ_{n-1} denote the n dimensional simplex and for each $\theta \in \Theta$ we define,

$$\mathcal{W}_\theta = \left\{ w \in \Delta_{n-1} \ : \ \sum_{i=1}^{n} w_i \psi_\theta\left(Y_i, A_i\right) = 0 \right\} \text{ and } \mathcal{W} = \bigcup_{\theta \in \Theta} \mathcal{W}_\theta. \qquad (3.1)$$

Under Assumptions 1, 2 and 3, a natural conditional empirical composite likelihood function corresponding to (2.10) is obtained by substituting $dF_i^0 = F^0(\{V_i\})$ by w_i and Υ by $\hat{\Upsilon} =$

$\sum_{i=1}^{n} \nu_i w_i$. It takes the form:

$$L\left(V, \alpha, \theta\right) = n^n \frac{\prod_{i=1}^{n} \nu_i w_i}{\left(\sum_{i=1}^{n} \nu_i w_i\right)^n}. \tag{3.2}$$

The factor n^n is for normalisation. The log-likelihood is given by:

$$L_{CE}(\theta, w, \nu) = n\log(n) + \sum_{i=1}^{n} \log(\nu_i w_i) - n\log\left(\sum_{i=1}^{n} \nu_i w_i\right). \tag{3.3}$$

In presence of parametric information we estimate the weights $\hat{w}_{CE}$ as $\arg\ \max_{w\in\mathcal{W}} L_{CE}(w, \nu)$. A constrained estimator $\hat{\theta}_{CE} \in \Theta$ can be obtained as (Qin and Lawless, 1994; Chaudhuri et al., 2008)

$$\hat{\theta}_{CE} = \arg\max_{\theta\in\Theta}\left\{\max_{w\in\mathcal{W}_\theta}\left(L_{CE}(w, \nu)\right)\right\}. \tag{3.4}$$

Kim (2009) considers estimation of population mean under Poisson sampling and uses expression (3.3) with ν_i replaced by π_i. In the context of two sample testing, Qin (1993) maximises (3.3) w.r.t. w_i and Υ with the additional constraint $\sum_{i=1}^{n} w_i\pi_i = \Upsilon$. Similar approaches have been taken by Qin et al. (2002); Qin and Zhang (2007) to include auxiliary information in the presence of non-ignorable missing observations.

Choice of $\hat{\Upsilon}$ in the second term of (3.3) is crucial. Our choice involves both ν_i and w_i. Use of the sample mean of π or ν would lead to unweighted estimator of the parameters.

We follow Pfeffermann and Sverchkov (1999, 2003) and estimate α separately from w. In particular, $\hat{\alpha}$, the maximum likelihood estimator for α obtained under the model for $E_{\mathcal{P}}[\pi|V]$ is used to obtain ν. In most cases, our main interest is in finding ν, not $\hat{\alpha}$.

4 A characterisation of the maximum empirical likelihood estimator

4.1 Connection to inverse conditional visibility weighted pseudo-likelihood estimator

When the data is collected through a complex design, in order to estimate θ, it is perhaps natural to consider the solutions of the following estimating equations:

$$\sum_{i=1}^{n} \frac{1}{\nu_i}\psi_\theta\left(y_i, a_i\right) = 0 \tag{4.1}$$

Estimators based on *inverse probability weighted* score functions, as in (4.1), but with ν replaced by π, have been studied in details in the statistics literature. They occur very often in connection with missing data, two-phase designs, etc. Even if the conditional visibilities are replaced by fixed sampling probabilities π, the left hand side of equation (4.1) is based on a pseudo-likelihood (Pfeffermann and Sverchkov, 2003) and can be seen as a design unbiased Horvitz-Thompson estimator of the total of $\psi_\theta\left(y_i, a_i\right)$ in the finite population. Beaumont (2008) uses smoothed version of fixed weights, in practice similar to conditional visibilities in what he regards as a smoothed Horvitz-Thompson estimation. However, in our case since both π and ν are assumed random, the justification (4.1) as an usual Horvitz-Thompson type estimator is not entirely appropriate. In this section we show that an alternative explanation using our proposed empirical likelihood based estimator is available.

Suppose $\mathcal{V}$ is the set of solutions of (4.1) and $\hat{\Theta}_{CE}$ is the collection of all $\hat{\theta}_{CE}$ in (3.4). Then we have the following result.

Theorem 4.1. If $\mathcal{V}$ is non-empty, then $\mathcal{V} = \hat{\Theta}_{CE}$.

Our procedure works even when $\mathcal{V}$ is empty. Thus the framework introduced above is more general. Furthermore, it avoids invoking Horvitz-Thompson estimator and provides a better explanation of inverse probability weighted score function based estimators in terms of conditional empirical likelihood. It is plain to see that our derivation follows naturally from a likelihood framework. The resulting log-likelihood is also different from a typical weighted log-likelihood found in the literature. This can be exploited in Bayesian formulations of related problems specially in small-area estimation and in multi-phase sampling where the design in the later phases depend on the observed variables in the earlier phase (e.g. Breslow and Wellner (2006)).

If (4.1) has a unique solution, $\hat{\theta}_{CE}$ is unique. Thus if ψ is obtained from a score function corresponding to a generalised linear model, $\hat{\theta}_{CE}$ would be unique.

The following corollary is an easy consequence of Theorem 4.1 which provides an estimate of F^0.

Corollary 4.1. When $\mathcal{V}$ is non-empty, the estimate of F_0 obtained by maximising (3.3) over $\mathcal{W}$ is given by:

$$\hat{F}_{CE}(C) = \sum_{i=1}^{n} \frac{(1/\nu_i)}{\sum_{i=1}^{n}(1/\nu_i)} \mathbf{1}_{\{V_i \in C \subseteq \mathbb{R}^{m+p+1}\}}. \tag{4.2}$$

4.2 General Result

For a given θ, in order to get a constrained estimator of w the objective function is given by:

$$L\left(w, \lambda_1, \lambda_2\right) = \sum_{i=1}^{n} \log(w_i) - n \log\left(\sum_{i=1}^{n} \nu_i w_i\right) - \lambda_1 \left(\sum_{i=1}^{n} w_i - 1\right) - n\lambda_2^T \sum_{i=1}^{n} w_i \psi_i, \tag{4.3}$$

where $\psi_i = \psi_\theta\left(y_i, a_i\right)$. λ_1 and λ_2 are unknown Lagrange multipliers which depend on θ as well.

By differentiating (4.3) with respect to w_i and following Owen (2001) mutatis mutandis, it follows that $\lambda_1 = 0$ and with $\kappa = \lambda_2 \sum_{i=1}^{n} \nu_i w_i$ we get:

$$w_i = \frac{\sum_{i=1}^{n} \nu_i w_i}{n} \frac{1}{\nu_i + \kappa^T \psi_i}. \tag{4.4}$$

Under our assumptions about ψ, (see Qin and Lawless (1994)) κ is a continuous differentiable function of θ. It is easily seen that κ satisfies the following equation,

$$\sum_{i=1}^{n} w_i \psi_i = \sum_{i=1}^{n} \frac{\psi_i}{\nu_i + \kappa^T \psi_i} = 0. \tag{4.5}$$

The value of κ for a given θ can be determined from (4.5). Since for each i, $w_i \leq 1$, from (4.4) for each i, κ would satisfy the constraint $n\left(\nu_i + \kappa^T \psi_i\right) \geq \sum_{i=1}^{n} \nu_i w_i$. Even though our motivation and formulations are completely different, the expression for w_i in (4.4) is similar to those obtained by, Kim (2009); Berger and De La Riva Torres (2016); Oguz-Alper and Berger (2016) in their respective settings. In their formulation however, expressions with ν replaced by π are obtained.

Now by substituting the value of w_i in the expression of log-likelihood in (3.3) we get,

$$L_{CE}(\theta, w, \nu) = -\sum_{i=1}^{n} \log(\nu_i + \kappa^T \psi_i) + \sum_{i=1}^{n} \log(\nu_i). \tag{4.6}$$

It is evident that (4.6) is satisfied by any θ such that $\mathcal{W}_\theta \neq \emptyset$. So the derivative, $\partial L_{CE}/\partial\theta$ evaluated at each $\hat{\theta}_{CE}$ will be equal to zero.

The discussion outlined above leads to the following general sets of estimating equations for $\hat{\theta}_{CE}$.

Theorem 4.2. Under our assumptions, each $\hat{\theta}_{CE} \in \hat{\Theta}_{CE}$ satisfy the following sets of equations:

$$\sum_{i=1}^{n} \frac{\psi_{\hat{\theta}_{CE}}(y_i, a_i)}{\nu_i + \kappa^T(\hat{\theta}_{CE})\psi_{\hat{\theta}_{CE}}(y_i, a_i)} = 0, \tag{4.7}$$

$$\sum_{i=1}^{n} \left.\frac{\kappa^T(\theta)\psi'_\theta(y_i, a_i)}{\nu_i + \kappa^T(\theta)\psi_\theta(y_i, a_i)}\right|_{\theta=\hat{\theta}_{CE}} = 0 \tag{4.8}$$

where $\psi'_{\theta'} = \partial\psi_\theta/\partial\theta|_{\theta=\theta'}$.

Equations (4.7) and (4.8) closely resemble those used by Qin and Lawless (1994) to inspect the asymptotic properties of empirical likelihood based estimators using independent and identically distributed data sets. These equations are also used in Section 5.1 to derive the asymptotic properties of proposed $\hat{\theta}_{CE}$.

5 Asymptotic properties and estimators of standard errors

5.1 Asymptotic properties

We consider the asymptotic properties of $\hat{\theta}_{CE}$ under the true population distribution F^0 as the population size N grows to infinity. We shall show that these properties of $\hat{\theta}_{CE}$ closely resembles those for ordinary empirical likelihood based estimator as described in Qin and Lawless (1994). For a formal framework for asymptotic analysis of growing population size we refer to Fuller (2009).

Let us denote, $f\left(y, a, \nu, \theta\right) = \psi_\theta\left(y, a\right)/\nu$. Suppose θ_0 be the true value of θ. We make the following assumptions on f (Qin and Lawless, 1994; Serfling, 1980):

A.1. For all $1 \leq i \leq N$, $f\left(y_i, a_i, \nu_i, \theta_0\right)$ are independent and identically distributed random vectors for any θ.

A.2. For all ν, $E_{\mathcal{P}}\left[f(y, a, \nu, \theta_0)\right] = 0$.

A.3. $E_{\mathcal{P}}\left[f\left(y_1, a_1, \nu_1, \theta_0\right) f\left(y_1, a_1, \nu_1, \theta_0\right)^T\right]$ is positive definite.

A.4. The Jacobian $\partial f\left(y_1, a_1, \nu_1, \theta\right)/\partial\theta$ is continuous in a neighbourhood of θ_0. Furthermore, in this neighbourhood $||\partial f\left(y_1, a_1, \nu_1, \theta\right)/\partial\theta||$ and $||f\left(y_1, a_1, \nu_1, \theta\right)||^3$ are both bounded by an integrable function $\mathcal{G}(y, a, \nu)$.

A.5. $E_{\mathcal{P}}\left[\partial f\left(y_1, a_1, \nu_1, \theta\right)/\partial\theta|_{\theta=\theta_0}\right]$ has rank p.

A.6. The Hessian matrix $\partial^2 f\left(y_1, a_1, \nu_1, \theta\right)/\partial\theta\partial\theta^T$ is continuous in θ in a neighbourhood of the true value θ_0 and in this neighbourhood, $||\partial^2 f\left(y_1, a_1, \nu_1, \theta\right)/\partial\theta\partial\theta^T||$ is bounded by some integrable function $\mathcal{H}(y, a, \nu)$.

Lemma 5.1. Suppose the assumptions A.1.- A.5. hold. Then, as $N \to \infty$, underF_0, with probability 1, L_{CE} attains its maximum value at some point $\hat{\theta}$ in the interior of the ball $||\theta - \theta_0|| \leq N^{-1/3}$. $\hat{\theta}$ and $\hat{\kappa} = \kappa(\hat{\theta})$ satisfy equations (4.7) and (4.8).

To prove asymptotic unbiasedness and normality, we first define, $\mathbb{G} = E_{\mathcal{P}}\left[\nu_1^{-1}\,\partial\psi_\theta\left(y_1, a_1\right)/\partial\theta|_{\theta_0}\right]$ and $\mathbb{G}^\star = E_{\mathcal{P}}\left[\nu_1^{-2}\psi_{\theta_0}\left(y_1, a_1\right)\psi_{\theta_0}\left(y_1, a_1\right)^T\right]$.

Theorem 5.1. Suppose the assumptions A.1. - A.6. hold. Then equations (4.7) and (4.8), with n replaced by N, admits a sequence of solutions $\left(\hat{\theta}^{(N)}_{CE}, \hat{\kappa}^{(N)}\right)$ such that,

1. $\left(\hat{\theta}^{(N)}_{CE}, \hat{\kappa}^{(N)}\right) \longrightarrow (\theta_0, 0)$ almost surely,

2. $\sqrt{N}\left(\hat{\theta}_{CE}^{(N)} - \theta_0\right) \Longrightarrow N\left(0, V_{CE}\right)$ in distribution, where
$$V_{CE} = \left\{\mathbb{G}^T(\mathbb{G}^\star)^{-1}\mathbb{G}\right\}^{-1}, \tag{5.1}$$
3. $\sqrt{N}\left(\hat{\kappa}^{(N)} - 0\right) \Longrightarrow N\left(0, V_\kappa\right)$ in distribution, where
$$V_\kappa = (\mathbb{G}^\star)^{-1}\left\{I - \mathbb{G}V_{CE}\mathbb{G}^T(\mathbb{G}^\star)^{-1}\right\}, \tag{5.2}$$
4. $\hat{\theta}_{CE}^{(N)}$ and $\hat{\kappa}^{(N)}$ are asymptotically independent.

Corollary 5.1. If $\mathbb{G}$ is invertible, $V_{CE} = \mathbb{G}^{-1}\mathbb{G}^\star(\mathbb{G}^T)^{-1}$. Furthermore, $\hat{\kappa}$ is asymptotically degenerate at 0.

5.2 Estimators for finite sample sizes

For finite sample size n, the standard error of $\hat{\theta}_{CE}$ can be estimated from a sandwich estimator based on the expression of V_{CE} in (5.1). To that effect we define $\hat{\mathbb{G}} = \sum_{i=1}^n \hat{w}_i \partial\psi_\theta\left(y_i, a_i\right)/\partial\theta|_{\theta=\hat{\theta}_{CE}}$ and $\hat{\mathbb{G}}^\star = \sum_{i=1}^n \hat{w}_i^2 \psi_{\hat{\theta}_{CE}}\left(y_i, a_i\right)\psi_{\hat{\theta}_{CE}}\left(y_i, a_i\right)^T$. The estimated variance of $\hat{\theta}_{CE}$ is then given by $\hat{V}_{CE} = \left\{\hat{\mathbb{G}}^T(\hat{\mathbb{G}}^\star)^{-1}\hat{\mathbb{G}}\right\}^{-1}$.

5.3 Predictors for finite population

Suppose that the parameter θ can be interpreted as a valid summary of the finite population of size N which is the solution of the set of equations $\sum_{i=1}^N \psi_\theta\left(y_i, a_i\right) = 0$. Let $\hat{\theta}_{\mathcal{P}}$ is the predicted value of θ in the population. Clearly, one choice of $\hat{\theta}_{\mathcal{P}}$ is $\hat{\theta}_{CE}$.

When the sample size is N and each $\pi_i = \nu_i = 1$, by Theorem 4.1 $\hat{w}_i = N^{-1}$ for each $i = 1$, 2, . . ., N. That is $\hat{\theta}_{CE} = \hat{\theta}_{\mathcal{P}}$. Likewise, under the same conditions $\hat{V}_{CE}$ becomes

$$\frac{1}{N}\cdot\left[\frac{1}{N}\sum_{i=1}^N \left.\frac{\partial\psi_\theta\left(y_i, a_i\right)}{\partial\theta}\right|_{\hat{\theta}_{\mathcal{P}}}\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N \psi_{\hat{\theta}_{\mathcal{P}}}\left(y_i, a_i\right)\psi_{\hat{\theta}_{\mathcal{P}}}\left(y_i, a_i\right)^T\right]\left[\frac{1}{N}\sum_{i=1}^N \left.\frac{\partial\psi_\theta\left(y_i, a_i\right)}{\partial\theta}\right|_{\hat{\theta}_{\mathcal{P}}}\right]^{-1}.$$

This is exactly the square of the standard error of $\hat{\theta}_{CE}$ obtained from N i.i.d. observations. This shows that our estimates comply with the assumptions of the superpopulation model.

The prediction variance $V_{\mathcal{P}}$ for this choice of $\hat{\theta}_{\mathcal{P}}$, is more difficult to estimate. Such variances should converge to zero as the sample size increases to the population size. The simple finite population correction leading to $\hat{V}_{\mathcal{P}}\left(\hat{\theta}_{\mathcal{P}}\right) = \left(1 - n/N\right)\hat{V}_{CE}$ does not seem to be very accurate. It should be also noted that, the proposed variance estimate does not include the dependence among the sampled observations or the association between the observations and the inclusion probabilities. These terms will play a significant role in estimating the prediction variance.

6 Illustration in the 2004 U.S. presidential election

In this section, we illustrate the inference resulting from this framework and compare it to a standard estimator. We consider the county-wise vote counts from the presidential election in United States of America in 2004. The data contains the total votes cast in favour of John Kerry, George W. Bush and Ralph Nader in each of the $N = 4600$ counties in the country. Let p be the proportion of counties where John Kerry had the majority of the cast votes in the election. From the data we find that Mr. Kerry won 1507 counties, which means the true value of the proportion is approximately 0.3276. Suppose we want to estimate p from a sample of size $n = 40$. Samples were drawn with probability proportional to the total votes cast in the county for the three candidates. We use the Tillé, Midzuno and PPS-systematic schemes (Cochran, 1963; Tille, 2006) to draw our samples. The observations from the Tillé scheme are nearly independent, those from the Midzuno scheme observations are slightly more dependent, and those from the systematic sample are highly dependent. All samples were drawn using the `sampling` package for `R` (Lumley, 2013; R Development Core Team, 2013).

Suppose I_i is the indicator that Kerry won the ith county in the sample. Thus the constraint imposed by the model on the unknown weights w is given by

$$\sum_{i=1}^{n} w_i \left(I_i - p\right) = 0. \tag{6.1}$$

Suppose c_i denote the total votes cast in the ith county for the three candidates. The probability of its selection, π_i, is proportional to c_i. Since c_i is available in the sample, we take $V_i = (I_i, c_i)$ and since π is proportional to c_i, we get $\nu_i = \pi_i$ for all i.

The unique solution of the equation $\sum_{i=1}^{n} \left(I_i - p\right) \pi_i^{-1} = 0$ is $\hat{p}_{CE} = \sum_{i=1}^{n} I_i \pi_i^{-1} / \sum_{i=1}^{n} \pi_i^{-1}$. This coincides with the Hajek estimator, usually motivated by design-based considerations. Note that, although the point estimators coincide, the CE approach provides a different inferential framework. From Theorem 4.1, $\hat{p}_{CE}$ is the unique empirical likelihood based estimate and $\hat{w} \propto \pi^{-1}$. Furthermore, from Theorem 5.1 and Corollary 5.1, an estimate of the variance is given by

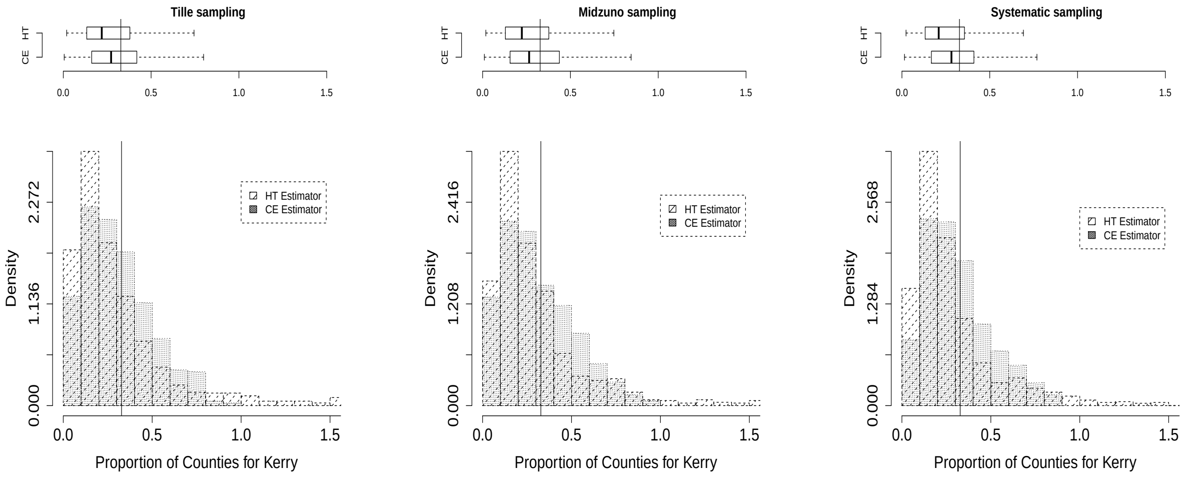
$$\hat{V}_{CE} = \frac{\sum_{i=1}^{n} \left(I_i - \hat{p}_{CE}\right)^2 \pi_i^{-2}}{\left(\sum_{i=1}^{n} \pi_i^{-1}\right)^2}. \tag{6.2}$$

We estimate $\hat{p}_{\mathcal{P}}$ and $\hat{V}_{\mathcal{P}}$ as described in Section 5.3.

We compare the CE estimator to the Horvitz-Thompson (HT) estimator for a mean, given by $\hat{p}_{HT} = (4600)^{-1} \sum_{i=1}^{n} I_i / \pi_i$.

We summarise the results of our study in Figure 2, Table 1 and Table 2. They are based on an average of 1000 draws. Standard errors and the coverage of the two-sided nominally 95% confidence intervals, obtained using Gaussian limits, are also presented.

From the histograms in Figure 2 it is seen that the Horvitz-Thompson estimator is not always bounded above by one. So in some cases the estimates of the proportion is hard to interpret. This is



(a) (b) (c)

Figure 2: Histograms of Horvitz-Thompson and Constrained Empirical Likelihood estimator of the proportion of counties won by John Kerry in the 2004 United States Presidential election. The plots are based on 1000 replications each of sample size 40 collected using 2(a) Tille, 2(b) Midzuno and 2(c) PPS-systematic schemes, respectively. The vertical black line indicates the true value of this proportion in the population, which is 0.3276.

specially true for PPS-systematic sampling where the highest value turns out to be 11.34 (truncated in Figure 2(c)). This estimator turns out to be larger than the one in other two types of sampling as well. In contrast, the Constrained Empirical Likelihood estimator always varies between zero and one. The mean of the Horvitz-Thompson estimator of proportion turns out to be quite close to the correct value for all sampling schemes (see Table 1). This is similar to the proposed $\hat{p}_{\mathcal{P}}$. The histograms show that it has a higher modal value than the proposed estimator. However, the skewness of the former is larger.

In Table 2, we intend, first, to compare the variation in the Horvitz-Thompson estimator with the Constrained Empirical Likelihood estimator, and second, to compare the performance of the Constrained Empirical Likelihood asymptotic variance estimator in (6.2) with some competing ones. For comparison we calculate (using the `survey` package in `R` (R Development Core Team, 2013)) the Hartley-Rao and Yates-Grundy-Sen estimators of the variance of the Hajek estimator. Note that, like $\hat{V}_{CE}$, the Hartley-Rao estimator only uses the first-order inclusion probabilities. The pairwise inclusion probabilities are used in the Yates-Grundy-Sen estimator only. All three estimators are compared with the observed root mean squared error of $\hat{p}_{\mathcal{P}}$ from 1000 draws. This root mean squared error is used as a benchmark in our comparison. As expected, it has coverage close to the nominal value.

The root mean squared errors of the Constrained Empirical Likelihood estimator are insensitive to the choice of sampling procedures. An exception is the Yates-Grundy-Sen estimator under systematic sampling, where negative estimates of variance are obtained. We see that all three estimators, that is the proposed $\hat{V}_{CE}$, Hartley-Rao and Yates-Grundy-Sen estimators underestimates

Table 1: Point estimates of the Constrained Empirical Likelihood and the Horvitz-Thompson estimator under three different sampling schemes. We estimate the proportion of counties won by John Kerry in the 2004 presidential election. The true value is 0.3276.

Sampling Scheme	Estimator	Mean Estimate
Tillé	Horvitz-Thompson	0.3376
	Conditional EL	0.3086
Midzuno	Horvitz-Thompson	0.3205
	Conditional EL	0.3089
Systematic	Horvitz-Thompson	0.3277
	Conditional EL	0.3111

Table 2: Standard errors and coverages for two-sided nominally 95% confidence intervals of the Horvitz-Thompson and the Constrained Empirical Likelihood estimator under three different sampling schemes.

Sampling scheme	Estimator	Variation							
		Observed		Conditional EL		Hartley-Rao		Yates-Grundy-Sen	
		SE	CI%	SE	CI%	SE	CI%	SE	CI%
Tillé	HT	0.383	95.0%	-	-	0.386	63.1%	0.385	63.1%
	CE	0.188	95.1%	0.139	72.3%	0.136	69.4%	0.136	69.4%
Midzuno	HT	0.348	95.6%	-	-	0.366	63.1%	0.342	62.0%
	CE	0.189	95.8%	0.136	69.4%	0.139	71.0%	0.144	69.5%
Systematic	HT	0.511	97.8%	-	-	0.524	62.7%	NA	NA
	CE	0.187	95.5%	0.136	74.0%	0.137	74.5%	NA	NA

the observed root mean squared error (when it exists). This underestimation is expected, since both estimators only approximates the true variance of the Hajek estimator (Särndal et al., 2003). Their performances seems to be comparable.

The Horvitz-Thompson estimator, in general, has higher variance and lower coverage than the proposed estimator. It is also seen that the average values of Hartley-Rao and Yates-Grundy-Sen estimators are very close to its observed variance (except for systematic sampling). However, the histograms of these two variance estimators are skewed, which explains the low coverage.

We intend to make the data and R code available for these procedures on CRAN (R Development Core Team, 2013). The core routines will be compatible with the `survey` package (Lumley, 2013).

A Proofs

In this section we present the proofs of the theorems.

Proof of Lemma 2.1

Proof. Recall that $E_{\mathcal{P}}\left[I_S \mid D_{\mathcal{P}}\right] = \pi_S$. Now Assumption 1 implies

$$\pi\left(S, D_{\mathcal{P}}\right) = E_{\mathcal{P}}\left[\pi_S \mid D_{\mathcal{P}}\right] = E_{\mathcal{P}}\left[E_{\mathcal{P}}\left[I_S \mid D_{\mathcal{P}}\right] \mid D_{\mathcal{P}}\right] = E_{\mathcal{P}}\left[I_S \mid D_{\mathcal{P}}\right] = \pi_S.$$

The other side is immediate. □

Proof of Lemma 2.2

Proof.

1. Using $\pi_S = \pi\left(S, D_{\mathcal{P}}\right)$, $E_{\mathcal{P}}\left[I_S \mid \pi_S, D_{\mathcal{P}}\right] = E_{\mathcal{P}}\left[I_S \mid D_{\mathcal{P}}\right] = \pi_S$. This means

$$E_{\mathcal{P}}\left[I_S \mid \pi_S\right] = E_{\mathcal{P}}\left[E_{\mathcal{P}}\left[I_S \mid \pi_S, D_{\mathcal{P}}\right] \mid \pi_S\right] = E_{\mathcal{P}}\left[E_{\mathcal{P}}\left[I_S \mid D_{\mathcal{P}}\right] \mid \pi_S\right] = \pi_S. \qquad \text{(A.1)}$$

2. Clearly $Pr_{\mathcal{P}}\left[I_S = 1 \mid \pi_S, D_{\mathcal{P}}\right] = \pi_S$. Since I_S is binary, its conditional distribution given π_S and $D_{\mathcal{P}}$ is a function of π_S only. So from the definition of conditional independence (Lauritzen, 1996) the result follows.

□

Proof of Lemma 2.3

Proof. From Lauritzen (1996, page 29) it can be shown that, $I_S \perp\!\!\!\perp \left(X_{\mathcal{P}}, X_{\mathcal{P}}, D_{\mathcal{P}}\right) \mid \pi_S$ is equivalent to $I_S \perp\!\!\!\perp D_{\mathcal{P}} \mid \pi_S$ and $I_S \perp\!\!\!\perp \left(X_{\mathcal{P}}, X_{\mathcal{P}}\right) \mid \left(\pi_S, D_{\mathcal{P}}\right)$. From Lemma 2.1, under Assumption 1, the second conditional independence relationship is equivalent to $I_S \perp\!\!\!\perp \left(X_{\mathcal{P}}, X_{\mathcal{P}}\right) \mid D_{\mathcal{P}}$. □

Proof of Lemma 2.4

Proof. The proofs follow from Lauritzen (1996, page 29). We only present a sketches.

1. Follows from Assumption 2.

2. From Assumption 2, it follows that $I_S \perp\!\!\!\perp \left(X_{\mathcal{P}}, X_{\mathcal{P}}\right) \mid \left(\pi_S, D_{\mathcal{P}}\right)$ holds. This together with Assumption 1 completes the proof.

3. This statement follows from 2. above.

□

Proof of Theorem 4.1

Proof. Since the geometric mean is bounded by the arithmetic mean. Clearly the relationship $\left(\prod_{i=1}^{n} n\nu_i w_i\right) / \left(\sum_{i=1}^{n} \nu_i w_i\right)^n \leq 1$ holds. The equality holds iff $w_i \propto \nu_i^{-1}$ for each $i = 1, 2, \ldots, n$.

Now for any $\hat{\theta} \in \mathcal{V}$, $\sum_{i=1}^{n} \psi_{\hat{\theta}}\left(y_i, a_i\right)/\nu_i = 0$. Thus $\hat{w}_i(\hat{\theta}) = \nu_i^{-1}/\sum_{i=1}^{n} \nu_i^{-1}$, $i = 1, 2, \ldots, n$ satisfy the constraints. Furthermore, $\left\{\prod_{i=1}^{n} n\nu_i \hat{w}_i(\hat{\theta})\right\} / \left\{\sum_{i=1}^{n} \nu_i \hat{w}_i(\hat{\theta})\right\}^n = 1$. Thus $\hat{\theta} \in \hat{\Theta}$ and $\mathcal{V} \subseteq \hat{\Theta}$.

Now let $\theta \in \hat{\Theta} \setminus \mathcal{V}$. For each fixed θ, $\mathcal{W}_\theta$ is a convex set and the log-likelihood (3.3) is a concave function. Thus the latter has a unique maxima $\hat{w}$. Since $\sum_{i=1}^{n} \psi_\theta\left(y_i, a_i\right)/\nu_i \neq 0$, $\hat{w}_i(\theta) \neq \nu_i^{-1}/\sum_{i=1}^{n} \nu_i^{-1}$, for at least one i. Thus $\left\{\prod_{i=1}^{n} n\nu_i \hat{w}_i(\theta)\right\} / \left\{\sum_{i=1}^{n} \nu_i \hat{w}_i(\theta)\right\}^n < 1$. However, since $\mathcal{V} \neq \emptyset$, there is a $\theta^\star \in \mathcal{V}$ such that $\left\{\prod_{i=1}^{n} n\nu_i \hat{w}_i(\theta^\star)\right\} / \left\{\sum_{i=1}^{n} \nu_i \hat{w}_i(\theta^\star)\right\}^n > \left\{\prod_{i=1}^{n} n\nu_i \hat{w}_i(\theta)\right\} / \left\{\sum_{i=1}^{n} \nu_i \hat{w}_i(\theta)\right\}^n$. This implies $\theta \notin \hat{\Theta}$ and $\hat{\Theta} \setminus \mathcal{V} = \emptyset$. □

Proof of Corollary 4.1

Proof. Evident from the definition of $\hat{F}_{CE}$ and Theorem 4.1 above. □

Proof of Lemma 5.1

Proof. Note that, for any θ, equation (4.7) can be re-expressed as,

$$\sum_{i=1}^{n} \frac{\psi_\theta(y_i, a_i)/\nu_i}{1 + \kappa(\theta)\psi_\theta(y_i, a_i)/\nu_i} = 0. \tag{A.2}$$

Now the rest of the proof follows from Owen (2001) and Qin and Lawless (1994, Lemma 1.), mutatis mutandis. □

Proof of Theorem 5.1

Proof. The proof is very close to that of Qin and Lawless (1994, Theorem 1). We expand (4.7) and (4.8) around $(\theta_0, 0)$. After some algebra this yields

$$\begin{pmatrix} \hat{\kappa}^{(N)} \\ \hat{\theta}_{CE}^{(N)} - \theta_0 \end{pmatrix} = J_N^{-1} \begin{pmatrix} -\frac{1}{N}\sum_{i=1}^{N} f(y_i, a_i, \nu_i, \theta_0) + o_p(\delta_N) \\ o_p(\delta_N) \end{pmatrix},$$

where

$$J_N = \begin{pmatrix} -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\nu_i^2}\psi_{\theta_0}(y_i,a_i)\psi_{\theta_0}(y_i,a_i)^T & \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\nu_i}\left.\frac{\partial\psi_\theta(y_i,a_i)}{\partial\theta}\right|_{\theta=\theta_0} \\ \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\nu_i}\left.\left(\frac{\partial\psi_\theta(y_i,a_i)}{\partial\theta}\right)^T\right|_{\theta=\theta_0} & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} -\mathbb{G}^\star & \mathbb{G} \\ \mathbb{G}^T & 0 \end{pmatrix} \text{ a.s.}$$

and $\delta_N = ||\hat{\theta}_{CE}^{(N)} - \theta_0|| + ||\hat{\kappa}^{(N)}|| = O_p\left(N^{-1/2}\right)$. The proof follows since $\sum_{i=1}^{N} f(y_i, a_i, \nu_i, \theta_0)/N = O_p(N^{-1/2})$. □

Proof of Corollary 5.1

Proof. The first part is trivial. For the second part note that if $\mathbb{G}$ is invertible, $V_\kappa = 0$. □

References

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, **95(3)**, 539–553.

Berger, Y. G. and O. De La Riva Torres (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78(2)**, 319–341.

Breslow, N. and J. Wellner (2006). Weighted likelihood for semiparametric models and two - phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, **34** 86–102.

Brewer, K. R. W. and M. Hanif (1983). *Sampling with unequal probabilities*, Volume 15 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Chambers, R. L., A. H. Dorfman, and M. Y. Sverchkov (2003). Nonparametric regression with complex survey data. In *Analysis of survey data (Southampton, 1999)*, Wiley Ser. Surv. Methodol., pp. 151–174. Chichester: Wiley.

Chaudhuri, A. and J. W. E. Voss (1988). *Unified Theory and Strategies of Survey Sampling*. North-Holland.

Chaudhuri, S., M. S. Handcock, and M. S. Rendall (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, J. and R. R. Sitter (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, **9(2)**, 385–406.

Chen, J., R. R. Sitter, and C. Wu (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, **89(1)**, 230–237.

Cochran, W. (1963). *Sampling techniques*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley.

Fuller, W. A. (2009). *Sampling Statistics*. Wiley and Sons.

Gill, R. D., Y. Vardi, and J. A. Wellner (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics*, (2)16(3), 1069–1112.

Godambe, V. P. (1975). A reply to my critics. *Sankhya, Series C*, **37**, 53–76.

Hartley, H. O. and J. N. K. Rao (1968). A new estimation theory for sample surveys. *Biometrika* (2)55(3), 547–557.

Hartley, H. O. Rao, J. N. K. (1975). Some comments on labels: A rejoinder to the section of godambe's paper, 'a reply to my critics'. *Sankhya, Series C*, (2)37, 163–170.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, **47**, 663–685.

Kim, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, **19(1)**, 145–157.

Krieger, A. M. and D. Pfeffermann (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, **18(2)**, 225–239.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing data*. John Wiley and sons.

Lumley, T. S. (2013). **survey***: Analysis of Complex Survey Samples*. Statnet Project. Version 3.28-2.

McCullagh, P. and J. Nelder (1989). *Generalised Linear Models*. Chapman& Hall/CRC.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–174.

Oguz-Alper, M. and Y. G. Berger (2016). Modelling complex survey data with population level information: an empirical likelihood approach. *Biometrika*, **103(2)**, 447–459.

Owen, A. (2001). *Empirical Likelihood*. Chapman& Hall/CRC.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Patil, G. P. and C. R. Rao (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**(2), 179–189.

Pfeffermann, D., A. M. Krieger, and Y. Rinott (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, **8(4)**, 1087–1114.

Pfeffermann, D. and M. Sverchkov (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā Ser. B*, **61(1)**, 166–186.

Pfeffermann, D. and M. Sverchkov (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey data*, pp. 175 – 195. Chichester: Wiley.

Qin, J. (1993). Empirical likelihood in biased sample problems. *The Annals of Statistics*, **21(3)**, 1182–1196.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**, 300–325.

Qin, J., D. Leung, and J. Shao (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of American Statistical Association*, **97**(457), 193–200.

Qin, J. and B. Zhang (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of Royal Statistical Society Series B Statistical Methodology*, **69(1)**, 101–122.

Rao, J. N. K. and C. Wu (2008). Empirical likelihood methods. In R. C. Pfeffermann D. (Ed.), *Handbook of statistics, Sample Surveys: Inference and Analysis*, Volume 29B, pp. 189–207. Elsevier.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rendall, M. S., R. Admiraal, A. DeRose, P. DiGiulio, M. S. Handcock, and F. Racioppi (2008). Population constraints on pooled surveys in demographic hazard modelling. *Statistical Methods and Applications* **17(4)**, 519–539.

Särndal, C., B. Swensson, and J. Wretman (2003). *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag GmbH.

Scott, A. J. (1977). Some comments on the problem of randomisation in surveys. *Sankhyā* **39**, 1–9.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Willey & Sons.

Skinner, C. J., T. Holt, and T. F. Smith (1989). *Analysis of Complex Surveys*. Wiley and Sons.

Sugden, R. A. and T. M. F. Smith (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, **71(3)**, 495–506.

Thomas, D. R. and G. L. Grunkemeier (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, **70**(352), 865–871.

Tighe, E., D. Livert, and L. Saxe (2010). Cross-survey analysis to estimate low-incidence religious groups. *Sociological Methods and Research*, **39(1)**, 56–82.

Tille, Y. (2006). *Sampling Algorithms*. Springer-Verlag.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics*, **13(1)**, 178–205. With discussion by C. L. Mallows.

Wu, C. (2004). Weighted empirical likelihood inference. *Statistics & Probability Letters*, **66(1)**, 67–79.

Wu, C. and J. N. K. Rao (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, **34(3)**, 359–375.