A CONDITIONAL EMPIRICAL LIKELIHOOD APPROACH TO COMBINE SAMPLING DESIGN AND POPULATION LEVEL INFORMATION

Sanjay Chaudhuri,
Mark S. Handcock
and
Michael S. Rendall

Technical Report No. 3/2010
September 28, 2010

National University of Singapore
Singapore, 117546
Abstract

We consider methods to include sampling weights in an empirical likelihood based estimation procedure to augment population level information in sample-based statistical modelling. Our estimator uses conditional weights and is able to incorporate covariate information both through the weights and the usual estimating equations. We show that the estimates are strongly consistent, asymptotically unbiased and normally distributed. Moreover, they are more efficient than other methods. Our framework provides additional justification for inverse probability weighted score estimators in terms of conditional empirical likelihood. We give two applications to demographic hazard modelling by combining birth registration data with complex survey data to estimate annual birth probabilities.

Sanjay Chaudhuri: Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546. E-mail address: sanjay@stat.nus.edu.sg.

Mark S. Handcock: Department of Statistics, University of California, Los Angeles. E-mail address: handcock@ucla.edu.

Michael S. Rendall: RAND Corporation, Santa Monica. E-mail address: mrendall@rand.org.
1. Introduction

In many applications in statistics and demography, the use population level information and sample survey data in conjunction is beneficial. In sample surveys, data are collected for a large number of variables, and thus meaningful models for the behaviour of a response of interest can be specified. However, survey data suffer from sampling error and from bias due to non-response. On the other hand, population level data collected from, e.g., census and vital events registration systems, typically do not contain a sufficient range of variables to specify meaningful models, but they are collected with comparatively less error and are less biased. These complementary strengths and weaknesses suggest that a combination of population and sample data may produce more meaningful and efficient estimates of the model parameters and lead to better inference.

Several methods of incorporating population level information in sample-based modelling have been investigated. One procedure is to express the population level information as functions of the model parameters and to use them as restrictions in parameter estimation. Handcock et al. [2000, 2005] consider a constrained maximum likelihood estimator (CMLE) while Imbens and Lancaster [1994] use generalised method-of-moments (GMM) to incorporate the constraints. It is known that both these methods produce asymptotically normal and unbiased estimates. Analytic forms of their asymptotic covariance matrices are known. However, both methods, particularly the CMLE, are computationally intensive and generally difficult to handle. The likelihood is constrained by non-linear equality constraints, and though standard software to perform such optimisation exists, it may be prohibitively slow for even moderate numbers of explanatory variables and population level constraints.

Empirical likelihood, introduced by Owen [2001], provides a semi-parametric method for augmentation of the population level information. Qin and Lawless [1994] showed that the empirical likelihood can be used to define a profile empirical likelihood of the parameters, which can then be maximised to obtain the parameter estimates. They further show that these estimates are asymptotically unbiased and normally distributed. Hellerstein and Imbens [1999] apply this method to economic data. Chen and Qin [1993] and Chen and Sitter [1999] use an empirical likelihood based method to incorporate auxiliary information available in a sample drawn from a finite population. For equal probability sampling, Chaudhuri et al. [2008] used an empirical likelihood based method to augment population level information in sample-based generalised linear modelling and developed a simple two-step method to estimate the model parameters. Under usual regulatory conditions, this two-step estimator is strongly consistent, asymptotically normal and unbiased. Further, incorporation of population level information reduces the standard error of the parameter estimates. Empirical likelihood based methods maximise their objective function under linear equality constraints and so is computationally much less demanding than the CMLE. Unlike the CMLE, no parametric form of the distribution has to be specified and so it is much more flexible and easy to implement. Moreover, the efficiency of this estimator is very close to the CMLE under the correctly specified parametric model and usually much better than the CMLE under misspecified models.

In many cases, however, sample observations are drawn according to a stratified design and are accompanied by known or approximated inclusion probabilities. These inclusion probabilities contain important information about the relationship between the sample and the underlying distribution in the population. In this
article we investigate methods of combining the information contained in these
design weights with sampled observations and population level information through
empirical likelihood based methods.

Empirical likelihood based methods which take into account the design weights
in the sample has been studied by several researchers. Chen and Sitter [1999],
motivated by the Horvitz-Thompson estimator in survey sampling, proposed a
pseudo-empirical likelihood. 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 paper we develop a framework that produces a different procedure to
the above. We use the framework to compare the two procedures (Section 6). We
consider the conditional distribution of the sample given that they were selected in
the sample and estimate their distribution in the population. Patil and Rao [1978]
considered a similar parametric approach and implemented it on size-biased sam-
pling. Pfeffermann and colleagues (e.g. Pfeffermann et al. [1998], Pfeffermann and
Sverchkov [1999], Krieger and Pfeffermann [1992]) investigated its use in parametric
modelling of survey data. Non-parametric estimators of the population distribu-
tion using the same principles were investigated by Vardi [1985]. He considered
multiple samples drawn from a population through different designs and provided
conditions for the existence and uniqueness of the non-parametric estimator of the
population distribution. The asymptotic properties of this non-parametric estimator
has been studied by Gill et al. [1988]. Qin [1993] employed empirical likelihood
in a two-sample testing problem, where only one sample is 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.
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 develop an empirical likelihood method based on the conditional likelihood
used by Pfeffermann and colleagues (e.g., Pfeffermann et al. [1998]). Population
and model information are incorporated to infer from a sample drawn according to
a complex design. In Section 2 we represent the sampling weights as random vari-
able and that they contain all the information about the design. In Section 3 we
consider parametric likelihood models. The formulation of this empirical likelihood
is discussed in Section 4. In Section 5 we develop a two-step estimation procedure
to estimate the model parameters. The asymptotic properties of this conditional
likelihood based estimator are compared with the pseudo-empirical likelihood es-
timator of Chen and Sitter [1999] (CS) and the unconstrained parametric pseudo
likelihood (PL) estimator. We show that, under usual regularity conditions, the
conditional likelihood (CL) estimator is strongly consistent, asymptotically unbi-
ased and has an asymptotic normal limit (Section 7). The CL estimator is shown
to be more efficient that the CS and PL estimators in applications to demographic
hazard modelling in two complex longitudinal surveys (Section 8).
2. Design and Model Specification

We consider a “superpopulation” model with response $Y$, a set of auxiliary variables $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ and a set of design variables $D$. The population $P$ is an i.i.d. sample of size $N$ generated from the model. A random sample $S$ of $n$ observations is drawn from $P$ according to a design depending on $D$ and possibly on some unknown parameters (specified in Section 4.2.2). The sample $S$ 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 at the design stage. We collect all the explanatory variables in the model in set $V$. Further, we denote $V = \{Y\} \cup X \cup Z$ to be the $m + p + 1$ dimensional random vector observed in the dataset.

2.1. Design Specification. The $i$th element in $S$ is drawn with probability $\pi_i$ or weight $d_i = \pi_i^{-1}/\sum_{i=1}^{n} \pi_i^{-1}$ (so that $\sum_{i=1}^{n} d_i = 1$). Since the design depends on the observed values of $D$ in the population, we assume the design to be random and so are $\pi$ and $d$.

If the sampled units are drawn according to $D$, the sampling mechanism may not be ignorable. The observed distribution of $V$ in the sample $S$ may be different from its distribution in the population and may depend on the particular sample selected. Suppose $I_S$ is the event that sample $S$ was selected. Let $V_S$ and $\pi_S$ respectively, be the matrix of observations and the vector of selection probabilities in the sample. $V_S$ are the observations not in the sample, where $S = \hat{P} \cdot S$. Let $V_P$, $Z_P$, etc, denote all the observations of the corresponding variables in $P$. $P = S \cup S$. $V_P$, $Z_P$ etc denotes all the observations of the corresponding variables in $P$.

The selection probabilities $\pi_S$ may be correlated to the variables in $X_P$ and $Y_P$ through $Z_P$. We make the following assumption:

Assumption 1: Conditional independence given the design. For all possible $S$, $\pi_S$ is conditionally independent of $Y_P$ and $X_P$ given $D_P$. That is,

\[ \pi_S \perp (Y_P, X_P) \mid D_P \text{ for all } S. \]

The joint probability of $I_S$ and $V_S$ in the population $P$ can be expressed as:

\[ Pr_P (I_S, V_S) = \int Pr_P (I_S|V_S, V_S, Z_P^c, \pi_S) Pr_P (V_S, V_S, Z_P^c, \pi_S) dV_Sd\pi_SdZ_P. \]

Assumption 2: Conditional independence given the sampling probabilities. We assume that for all possible $S$, given $\pi_S$, $I_S$ is conditionally independent of $V_S$, $V_S$ and $Z_P^c$. That is,

\[ I_S \perp (V_S, V_S, Z_P^c) \mid \pi_S \text{ for all } S. \]

This implies that:

\[ Pr_P (I_S|V_S, V_S, Z_P^c, \pi_S) = Pr_P (I_S|\pi_S) = \xi (\pi_S) \text{ for all } V_S, V_S, Z_P^c \text{ and } \pi_S, \]

and some function $\xi$ of $\pi_S$.

Under this assumption:

\[ Pr_P (I_S, V_S) = \int \xi (\pi_S) Pr_P (V_S, \pi_S) d\pi_S = \int \xi (\pi_S) Pr_P (\pi_S \mid V_S) d\pi_S \int Pr_P (V_S) \]

\[ = E_P [\xi (\pi_S) \mid V_S] Pr_P (V_S). \]

Assumption 2 is stronger than the “Condition 1” in Sugden and Smith [1984]. It implies $I_S \perp D_P \mid \pi_S$. Assumptions 1 and 2 together also implies that for all $S$
the condition \((I_S, \pi_S) \perp \perp (Y_P, X_P) \mid D_P\) holds. The first condition implies that \(\pi_S\) contains all information about the actual design mechanism. The second condition assumes that \(Y_P\) and \(X_P\) influence \(\pi_S\) and \(I_S\) only through the design variables \(D_P\) in the population. A graphical representation of the assumed conditional independencies is given in Figure 1.

The assumption that the selection probability \(\pi_S\) contains all information about the design 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 datasets are constructed by merging several available datasets obtained from different surveys. Typically, each survey is based on different designs dependent on different variables. A design for the merged dataset may not be easy to specify, but weights from individual surveys can be used to provide information about the underlying designs.

From the graph in Figure 1 we note that, \(\pi_S\) and \(I_S\) may depend on whole of \(Y_P\) and \(X_P\) via \(Z_P\) which are unobserved in \(S\). Thus even though \(\pi_S\) does not depend on \(Y_S\) and \(X_S\) directly, in (4) above, \(E_P[\xi(\pi_S) \mid V_S] \neq E_P[\xi(\pi_S) \mid Z_S]\) in general.

Note that our assumptions do not require \(I_S\) to be conditionally independent of \(V_S\) given \(V_S\) (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 \(I_S \perp V_S \mid (V_S, \pi_S)\) holds.

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 known. Specifically, we assume that

\[
E_{F^0}[\psi_\theta(Y_1, A_1)] = 0.
\]

where \(\psi\) is a known function depending only on \(Y\) and \(A\) and some unknown parameter \(\theta\). There may be several choices for \(\psi\). A natural one for parametric models (e.g., *generalised linear models*) is the corresponding *score functions* \(S_\theta(Y, A)\).
Further, certain parameters in the superpopulation may be known without any error. Suppose $g$ is a given functional of $V$ not depending on $\theta$ and

$$E_{F_0} [g(V_1)] = \gamma.$$  

We then say that (6) specifies \textit{population-level auxiliary information} if $\gamma$ is known without any error.

### 3. Incorporating Population Level Information and Sampling Weights in Parametric Models

Suppose that $F_0 \in \mathcal{F}_\theta$ for some parametric family $\mathcal{F}_\theta$.

The population level information can be expressed as:

$$C(\theta) = E_{F_0} [g(V_1)] = \int g(V_1) dF_\theta = \gamma.$$  

A natural way to estimate $\theta$ satisfying (7) is considered by Handcock et al. [2005]. They maximise the log-likelihood corresponding to $dF_\theta$ under the constrained in (7).

The resulting constrained MLE (CMLE) is known to be asymptotically normally distributed and is more efficient than the corresponding unconstrained estimator. Handcock et al. [2005] consider weighted constrained likelihood estimation of $\theta$, but their estimator is not a maximum likelihood estimator for weighted designs.

In presence of sampling weights the unconstrained estimator of $\theta$ can be obtained by solving:

$$\sum_{i=1}^{n} d_i S_\theta (Y_i, A_i) = 0.$$  

Here $S_\theta (Y, A)$ is the score function corresponding to $dF_\theta$. We denote the corresponding estimator by $\hat{\theta}_{PL}$. This procedure is derived from a Horvitz-Thompson estimator [Horvitz and Thompson, 1952] of the population total of the score function. Krieger and Pfeffermann [1992] describe it as a \textit{pseudo-maximum likelihood estimator}. A conditional likelihood based unconstrained estimator of $\theta$ has been discussed by Pfeffermann and several colleagues (See Krieger and Pfeffermann [1992]). We discuss this estimator in Section 4.2.2. Pfeffermann and Sverchkov [2003] show that in certain contexts this estimator is equivalent to $\hat{\theta}_{PL}$. Under usual assumptions, $\hat{\theta}_{PL}$ is asymptotically unbiased and normally distributed. Its asymptotic variance can also be computed analytically [Chambers, 2003]. In what follows, $\hat{\theta}_{PL}$ denotes the weighted unconstrained pseudo-likelihood estimator. We compare the efficiency of this estimator to other estimators in Section 8. It is at a disadvantage as it does not incorporate any population level information.

### 4. Empirical Likelihood to Incorporate Population Level Information in Parameter Estimation

In this section we discuss various empirical likelihood based methodology to include population level information in statistical modelling based on sample data. We initially assume the sample points are all drawn with equal probability. This assumption is relaxed in Section 4.2, where we introduce our conditional likelihood based empirical likelihood estimator.
4.1. Empirical likelihood for equally weighted samples. As before, we assume that \( V_1, V_2, \ldots, V_n \) are independent random vectors following \( F^0 \in \mathcal{F} \) in the population. We do not specify a parametric form for \( \mathcal{F} \) but specify constraints on the model and population as before ((5) and (6), respectively). Using empirical likelihood, we estimate \( F^0 \) non-parametrically from the observed sample and include all the available parametric or population based informations in the analysis.

Suppose, following Owen [2001], for each \( F \in \mathcal{F} \) we let:

\[
(9) \quad w_i = F(\{V_i\})
\]

be the weights \( F \) assigns on \( V_i \) ( \( w_i = 0 \) for all \( F \) continuous at \( V_i \)). Let \( \Delta_{n-1} \) denote the \( n \) dimensional simplex and

\[
(10) \quad \mathcal{W}_\theta = \left\{ w \in \Delta_{n-1} : \sum_{i=1}^{n} w_i \psi_g (Y_i, A_i) = 0 \right\} \quad \text{for each } \theta \in \Theta,
\]

\[
(11) \quad \mathcal{W}_P = \left\{ w \in \Delta_{n-1} : \sum_{i=1}^{n} w_i g (V_i) = \gamma \right\}.
\]

\[
(12) \quad \mathcal{W} = \bigcup_{\theta \in \Theta} (\mathcal{W}_\theta \cap \mathcal{W}_P).
\]

The empirical likelihood of \( F \) is defined as:

\[
(13) \quad L(F) = \prod_{i=1}^{n} w_i
\]

Let

\[
(14) \quad \hat{w} = \arg \max_{w \in \mathcal{W}} \prod_{i=1}^{n} w_i,
\]

then the estimator of \( F^0 \) defined by

\[
(15) \quad \hat{F}_w (C) = \sum_{i=1}^{n} \hat{w}_i 1_{\{V_i \in C\}} \quad \forall \ C \in \mathbb{R}^{m+p+1},
\]

satisfies both (5) and (6). Thus \( \hat{F}_w \) is the required constrained estimator of \( F^0 \). A constrained estimator \( \theta \in \Theta \) can be obtained as [Qin and Lawless, 1994, Chaudhuri et al., 2008]

\[
(16) \quad \hat{\theta} = \arg \max_{\theta \in \Theta} \left\{ \max_{w \in \mathcal{W}_\theta \cap \mathcal{W}_P} \left( \prod_{i=1}^{n} w_i \right) \right\}.
\]

4.2. Parameter estimation in the presence of population level constraints and sampling weights. We consider the situation where the distribution of \( V_i \), \( i = 1, 2, \ldots, n \) in the sample is not the same as that in the population. In this case knowledge of the sampling design is crucial and has to be included in the framework and parameter estimation. Below we discuss ways to include information about both the sampling design and population level information in our analysis.

We assume that \( V_1, V_2, \ldots, V_n \) are drawn independently from population \( \mathcal{P} \). Suppose \( V_i \) was drawn with probability \( \pi_i \). Let \( F_S^{(i)} \) be the conditional distribution of \( V_i \) given \( i \in \mathcal{S} \). Suppose \( dF_S^{(i)} \) is the density of \( F_S^{(i)} \). Using Bayes’ rule and (4) it follows that:

\[
(17) \quad dF_S^{(i)} = \frac{E_P \left[ \xi (\pi_i) \mid V_i \right] dF^0 (V_i)}{Pr_P (i \in \mathcal{S})}.
\]
Since the $i$th unit was drawn with probability $\pi_i$, from an argument by Pfeffermann, Krieger, and Rinott [1998], it follows that $\xi(\pi_i) = \pi_i$ for all $i = 1, 2, \ldots, n$. Further from (4),

$$Pr_P(i \in S) = \int Pr_P((i \in S), V_i) dV_i = \int E_P[\pi_i | V_i] dF^0(V_i) dV_i. \tag{18}$$

We call the conditional inclusion probability $\nu_i = E_P[\pi_i | V_i]$ the conditional visibility for the $i$th element in the population and $D_i = \int \nu_i dF^0(V_i) dV_i = E_P[\pi_i] = E_{F^0(V_i)}[\nu_i]$ the visibility factor for the $i$th element in the population [Patil and Rao, 1978]. By substituting these expressions into (17) we get:

$$dF^{(i)}_S = \frac{\nu_i dF^0(V_i)}{D_i}. \tag{19}$$

Note that the conditional likelihood in (19) is invariant to the scale of $\pi$ and $\nu$. For each $i$, $\pi$ and $\nu$ may be specified up to an arbitrary positive scaling constant.

If the distribution of $\pi_i | V_i$ in $S$ is equal to that in $P$, one can directly estimate $\nu_i$ from the available data. On the other hand, if the above two distributions are different, from Pfeffermann and Sverchkov [1999], we obtain:

$$Pr_S(\pi_i^{-1} | V_i) = \frac{\pi_i Pr_P(\pi_i^{-1} | V_i)}{E_P[\pi_i | V_i]}, \tag{20}$$

$$E_P[\pi_i | V_i] = [E_S(\pi_i^{-1} | V_i)]^{-1}. \tag{21}$$

It is not clear how to check if the population and sample distribution of $\pi | V_i$ are the same. In most cases use of (21) would be appropriate.

To specify $dF^{(i)}_S$ in (19) it is typically necessary to model the conditional visibility ($E_P[\pi_i | V_i]$) and the distribution of $V_i$ in the population ($dF^0(V_i)$). Both of these model may depend on unknown parameters. As in Section 2, we parametrise the $dF^0(V_i)$ by $\theta$. The model for the conditional visibility will be parametrised by $\alpha$.

Pfeffermann et al. [1998] discuss a class of conjugate parametric models for the distribution of $V_i$ and $\pi_i | V_i$ such that $dF^{(i)}_S$ is in the same class as $dF^0(V_i)$. This avoids a complicated computation of $D$. However, estimation of $\theta$ is typically complex. The parameters in $dF^{(i)}_S$ usually depends on both $\theta$ and $\alpha$. It is possible to model $E_P[\pi | V_i]$ non-parametrically, especially in situations where estimates of $\alpha$ are not of primary interest.

In many situations $\nu_i$ may only depend on a subset of variables in $V$. Furthermore, this subset may be quite different from $A$. This is particularly true if the sample $S$ was obtained by merging several subsamples drawn from different designs. $\nu_i$ may in that case depend on the particular sample the $i$th observation belongs to. Such sample indicator variables in most cases shall not be useful in modelling the response and consequently won’t be in $A$.

Computation of the conditional likelihood requires knowledge of the dependence structure of the $F^{(i)}_S$. As an alternative, we construct a pseudo-conditional likelihood for $\alpha$ and $\theta$, $i = 1, 2, \ldots, n$ as:

$$L_{CL}(V, \alpha, \theta) = \prod_{i=1}^{n} F^{(i)}_S. \tag{22}$$

Parametric estimation of $F^0$ by maximising (22) 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, with no constraints, Vardi [1985] shows that if $\pi_i > 0$,
for all \( i \), the non-parametric estimator of \( F_0 \) is unique. Further, if \( \pi \) is fixed this estimator is sufficient for \( F_0 \). Gill et al. [1988] have studied the large sample properties of this estimator. Finally, Peiffermann, 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 \( N \rightarrow \infty \). These results suggest that the pseudo-conditional likelihood may be a useful surrogate for the conditional likelihood in this setting.

4.3. Empirical likelihood based estimation from the conditional likelihood. As before, we assume that \( V_1, V_2, \ldots, V_n \) are independent random vectors following \( F_0 \in \mathcal{F}_\theta \) in the population. We estimate the parameter \( \theta \) of \( F_0 \) by maximising the corresponding non-parametric likelihood under constraints. Our approach follows Kim [2009] and Qin [1993]. Similar approaches have been taken by Qin et al. [2002], Qin and Zhang [2007] to include auxiliary information in the presence of non-ignorable data.

Under our assumptions, \( E_P[\pi_i] = D_i = D \) for all \( i = 1, 2, \ldots, n \). The natural non-parametric likelihood for \( F_0 \) corresponding to (19) is obtained by substituting \( dF_i = F(\{V_i\}) \) by \( w_i \) and \( D \) by \( \sum_{i=1}^n \nu_i w_i \).

The empirical pseudo-conditional log-likelihood function corresponding to (22) takes the form:

\[
L_{CL}(w, Z, \nu) = \sum_{i=1}^n \log(w_i) - n \log(\sum_{i=1}^n \nu_i w_i).
\]

In the presence of parametric and population level information the weights \( w \) can be estimated by

\[
\hat{w}_{CL} = \arg \max_{\hat{w} \in W} \left\{ \sum_{i=1}^n \log(w_i) - n \log(\sum_{i=1}^n \nu_i w_i) \right\}.
\]

The corresponding estimate \( \hat{\theta}_{CL} \) of \( \theta \) follows similarly from (16).

Kim [2009] considers estimation of the mean under Poisson sampling and uses expression (23) with \( \nu_i \) replaced by \( \pi_i \). In the context of two sample testing, Qin [1993] maximises (23) w.r.t. \( w_i \) and \( D \) with the additional constraint \( \sum_{i=1}^n w_i \nu_i = D \).

The choice of \( D \) in the second term of (23) is crucial. Our choice involves both \( \nu_i \) and \( w_i \). The use of the sample mean of \( \pi \) or \( \nu \) is not appropriate and would lead to unweighted estimator of the parameters.

A joint maximum empirical pseudo-conditional likelihood estimator for the conditional visibility model parameter can be obtained by maximising (23) over \( \alpha \) as well as \( \theta \). A simple alternative is to use \( \hat{\alpha} \), the maximum likelihood estimator for \( \alpha \) obtained under the model for \( E_P[\pi|V] \) used to obtain \( \nu \). However, maximising the empirical likelihood over \( \alpha \) may be computationally difficult. The estimator \( \hat{\alpha} \) would be asymptotically very close to the joint estimator of \( \alpha \) under mild assumptions. For computational ease, we will use \( \hat{\alpha} \) as the estimator for \( \alpha \).

An estimate \( \hat{F}_{CL} \) of \( F_0 \) can be obtained by substituting \( \hat{w}_{CL} \) in (15). In fact \( \hat{w}_{CL} \) has the desirable property that when \( \mathcal{W}_P = \Delta_{n-1} \), the first step of the two-step method gives \( \hat{w}_{CL} = \nu_i^{-1}/\sum_{i=1}^n \nu_i^{-1} \), for all \( i \).

**Theorem 1.** Suppose \( \mathcal{W}_P = \Delta_{n-1} \). The estimate of \( F_0 \) obtained by maximising (23) above over \( \mathcal{W}_P \) is given by:

\[
\hat{F}_{CL}(C) = \sum_{i=1}^n \frac{(1/\nu_i)}{\sum_{i=1}^n (1/\nu_i)} 1\{V_i \in C \subseteq \mathbb{R}^{m+p+1}\}.
\]
Proof. See Appendix.

Notice that, when \( \hat{w}_{CLi} = \nu_i^{-1}/\sum_{i=1}^{n} \nu_i^{-1} \), for all \( i \), \( \hat{\theta}_{CL} \) satisfies:

\[
\sum_{i=1}^{n} \frac{1}{\nu_i} \psi_{\theta}(y_i, a_i) = 0
\]

(26)

Estimators based on inverse probability weighted score functions, as in (26) have been studied in detail. They occur very often in connection to with missing data, two-phase designs, etc. However, since the weights \( \nu \) are random, their justification as a Horvitz-Thompson type estimator is not entirely appropriate. Our framework avoids invoking Horvitz-Thompson estimators, and provides a better explanation in terms of conditional empirical likelihood. Furthermore, the 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 sampling where the design in the later phases depend on the observed variables in the earlier phase [Breslow and Wellner, 2006].

5. Computational issues in the estimation of model parameters

Note that \( W \) is specified by linear constraints on weights. Thus, empirical likelihood based methodology has a clear computational advantage over the corresponding constrained maximum likelihood estimator. In this section we show how the computational burden can further be reduced by using a two-step procedure. In this procedure, the first step is to maximise the empirical likelihood over \( WP \). The maximised weights are then substituted in (10). These equations can then be solved to obtain \( \theta \). In (16) the outer maximisation w.r.t. \( \theta \) is unconstrained. However it is potentially a non-convex problem. We adapt the two-step procedure in Chaudhuri et al. [2008] to estimate \( w \) and \( \theta \).

In this adaption, the first maximisation is done over \( WP \). These maximising weights are then substituted in the estimating equation for \( \theta \) and in the second step these equations are solved to obtain the parameter estimates. Since the log-empirical likelihood in (23) is concave on a closed convex set \( WP \), it has a unique maximiser on \( WP \). Clearly if this maximising weights are in \( W_{\theta_0} \), for some \( \theta_0 \in \Theta \), \( \hat{\theta} = \theta_0 \). The two-step method may fail for a situation where there is a small sample size and when a solution to the second step does not exist. Such situations are rare in practice.

The two-step procedure is quite easy to compute. Once the maximising weights are available, one can solve the parametric equations by a simple Newton-Raphson method. The maximisation of the empirical likelihood over \( WP \) is not difficult either. Usually one solves the corresponding lower dimensional dual problem. It can be shown that the dual problem is restricted by inequality constraints. Owen [2001] defines an approximation of the log function with which the dual can be minimised without any constraints. [Chen et al., 2002] discusses a modified Newton-Raphson method with guaranteed convergence. Standard packages for computation exist [R Development Core Team, 2007].
5.1. A two-step procedure to compute $\hat{\theta}_{CL}$. The two-step method is straightforward to apply to $\hat{w}_{CL}$. However, maximising (23) over $W_P$ requires some discussion. Consider the objective function

$$L(w, \alpha, \lambda) = \sum_{i=1}^{n} \log(w_i) - n \log(\sum_{i=1}^{n} \nu_i w_i) - \alpha \left( \sum_{i=1}^{n} w_i - 1 \right) - n\lambda \sum_{i=1}^{n} w_i h_i,$$

where $\alpha$ and $\lambda$ are Lagrange multipliers and $h_i = g(Y_i, A_i) - \gamma$.

By differentiating w.r.t. $w_i$ for an extremum one obtains

$$0 = \frac{1}{w_i} - \frac{n \nu_i}{\sum_{i=1}^{n} \nu_i w_i} - \alpha - n\lambda h_i.$$

Now multiplication of (27) by $w_i$ on both sides and summation over $i$ yields $\alpha = 0$. So by writing $\kappa = \lambda \sum_{i=1}^{n} \nu_i w_i$ we obtain

$$w_i = \frac{\sum_{i=1}^{n} \nu_i w_i}{n} \frac{1}{\nu_i + \kappa h_i}.$$

Clearly $w_i \leq 1$, for each $i$ implies the restriction

$$n \{ \nu_i + \kappa h_i \} \geq \sum_{i=1}^{n} \nu_i w_i = D.$$

By substituting these values of $w_i$ into (23) we obtain

$$L_{CL}(w, V, \nu) = -\sum_{i=1}^{n} \log(\nu_i + \kappa h_i).$$

The weights can be estimated by minimising $L_{CL}(w, V, \nu)$ w.r.t. $\kappa$ under the restriction in (29) for all $i = 1, 2, \ldots, n$.

From (29) it is clear that the lower bound on $(\nu_i + \kappa h_i)$ depends on the weights which are unknown and thus direct constrained minimisation of $L_{CL}(w, V, \nu)$ does not follow from Owen [2001] in a straightforward manner. However, $\hat{w}_{CLLi}$ can be obtained, by solving a similar but easier problem, as we now see.

5.2. Constrained minimisation of $L_{CL}(w, V, \nu)$. We consider $L_{CL}(w, V, \nu)$ from (30). Notice that, we need to find $\kappa$ which solves

$$\sum_{i=1}^{n} \frac{h_i}{\nu_i + \kappa h_i} = 0$$

subject to (29).

By dividing the numerator and the denominator of each term by $w_i$ this is equivalent to solving

$$\sum_{i=1}^{n} \frac{h_i / \nu_i}{1 + \kappa h_i / \nu_i} = 0,$$

subject to the constraint in (29). Notice that the expression in (32) is analogous to the ones described in Owen [2001]. So available software can be used to maximise the product of weights. The following Lemma connects solution of (32) to $\hat{w}_{CLLi}$.

**Lemma 1.** Suppose

$$w^* = \arg \max_w \sum_{i=1}^{n} \log w_i$$

where $\alpha$ and $\lambda$ are Lagrange multipliers and $h_i = g(Y_i, A_i) - \gamma$. By differentiating w.r.t. $w_i$ for an extremum one obtains

$$0 = \frac{1}{w_i} - \frac{n \nu_i}{\sum_{i=1}^{n} \nu_i w_i} - \alpha - n\lambda h_i.$$
subject to
\[ w \in \Delta_{n-1} \text{ and } \sum_{i=1}^{n} w_i h_i / \nu_i = 0. \] (34)

Then
\[ \hat{w}_{CLi} = \frac{w_i^* / \nu_i}{\sum_{i=1}^{n} w_i^* / \nu_i}. \] (35)

Proof. See Appendix. \qed

We show later (see (56)) that \( \hat{w}^* = \nu_i \hat{w}_{CLi} / \hat{D} \). Thus \( \hat{w}^* \) correspond to constrained empirical likelihood estimator of \( F_S \). Parameter constraints can also be included. The detailed formulation is as follows:

\[ \hat{w}^* = \arg \max_{w} \sum_{i=1}^{n} \log(w_i) \] subject to
\[ w \in \Delta_{n-1}, \]
\[ \sum_{i=1}^{n} w_i \psi_0(Y_i, A_i) / \nu_i = 0, \] (37)
\[ \sum_{i=1}^{n} w_i h_i / \nu_i = 0. \] (38)

The constrained estimate of \( F_S \) can be obtained in usual way.

6. MINIMUM DIVERGENCE BASED METHODS

The maximum empirical pseudo-conditional likelihood estimator \( \hat{\theta}_{CL} \) can be viewed as a profile likelihood estimator. As an alternative to this, Chen and Qin [1993] and Chen and Sitter [1999] use an empirical likelihood based method to incorporate auxiliary information available in a sample drawn from a finite population. Chen and Sitter [1999] introduced a pseudo-empirical likelihood estimator to include fixed sampling weights \( d_i \). They estimate the total of \( \log(w_i) \) in the population through the design unbiased Horvitz-Thompson estimator from the sample. Their estimator for \( w \) is given by:

\[ \hat{w}_{CS} = \arg \max_{w \in W} \sum_{i=1}^{n} d_i \log w_i. \] (40)

This estimator has been frequently used in sampling literature in several context. Rao and Wu [2008] interpret of (40) as a “backward” Kullback-Leibler divergence between \( w \) and \( d \). \( \hat{w}_{CS} \) in (40) minimises this divergence. \( \hat{\theta}_{CS} \) can be obtained similarly as in (16). It is known that for population mean, under certain conditions pseudo-empirical likelihood based estimator (PELE) is asymptotically equivalent to the generalised regression (GREG) estimator. For stratified single stage and multi-stage sampling PELE is equivalent to the optimal regression estimator (ORE), but in many other cases PELE may be substantially better than the ORE. Further discussion on this estimate may be found in Glenn and Zhao [2007], Wu and Rao [2006], Fu et al. [2008] among others.

Note that in the absence of any population level information (i.e. \( W_p = \Delta_{n-1} \)), the two-step method gives \( w_i = d_i \), for \( i = 1, 2, \ldots, n \). Thus \( \hat{\theta}_{CS} \) is the solution of
the estimating equation $\sum_{i=1}^{n} d_i \psi_0 (Y_i, A_i) = 0$. Thus if $\mathcal{W}_p = \Delta_{n-1}$, $\theta_{CS}$ reduces to the usual weighted unconstrained estimator $\hat{\theta}_{PL}$. This is considered desired, but is not shared by weighted empirical type likelihoods described in Wu [2004], Chen and Qin [1993] etc.

Note that the minimum divergence based methods and the conditional empirical likelihood in (23) lead to different profile likelihoods for $\theta$.

The two-step estimation method developed for $\theta_{CL}$ can be adapted to $\theta_{CS}$. If $\sum_{i=1}^{n} d_i \psi_0 (Y_i, A_i) = 0$ has a solution, $d_i \in \mathcal{W}_0$ for some $\theta \in \Theta$. Thus $\hat{\theta}_{CS}$ obtained in the first step of the two-step procedure minimises the divergence between $d$ and $w$ over $\mathcal{W}_p$. Since $\mathcal{W}_p$ is closed and convex subset of $\Delta_{n-1}$ (and therefore compact), an unique minimiser of the divergence exists. The two-step method will fail in this case if $\hat{\theta}_{CS} \notin \bigcup_{\theta \in \Theta} \mathcal{W}_0$. The solutions to both the first and the second steps are easy to obtain.

The CL estimator in (36) is fundamentally different from the minimum divergence estimator. In the minimum divergence estimator the objective function depends on the design weights but the restricting equations do not. In (36) the sampling weights appear in the restricting equations but not in the objective function. Two estimators use different sets of weights though, unless $\nu = \pi$.

It can be shown that, for all $i$, $\hat{\theta}_{CS} = d_i/n(1 + \lambda h_i)$. These can be compared to those in (28) which are the estimates of the weights for the CL estimator. Similarly, the profile likelihood estimator [Qin and Lawless, 1994] for $\theta$ is different in these two cases.

7. ASYMPTOTIC PROPERTIES OF THE ESTIMATORS

In this section we discuss the asymptotic properties of the two parameter estimates of $\theta$ obtained from the two empirical likelihood based methods (CL and CS).

We show that, as $n \to \infty$ and under the usual assumptions, both $\hat{\theta}_{CS}$ and $\hat{\theta}_{CL}$ are strongly consistent and asymptotically normal. They have different asymptotic covariance matrices. We compare their asymptotic standard errors with the unconstrained estimator. Further, when $\nu = \pi$, $\hat{\theta}_{CL}$ has smaller asymptotic standard error than $\hat{\theta}_{CS}$.

We first discuss notation and specify the assumptions. Let us denote:

$$f_1 (v, d, \theta, \lambda) = \frac{d}{1 + \lambda h(v, \gamma)} (\psi_0 (y, a), h(v, \gamma))$$

$$f_2 (v, \nu, \theta, \kappa) = \frac{1}{\nu + \kappa h(v, \gamma)} (\psi_0 (y, a), h(v, \gamma))$$

Suppose $\theta_0$ is the true value of $\theta$. Following Qin and Lawless [1994] and Serfling [1980] we make the following assumptions.

A.1. We assume that both $f_1 (v_i, d_i, \theta, \lambda)$, $1 \leq i \leq n$ and $f_2 (v_i, \nu_i, \theta, \kappa)$, $1 \leq i \leq n$ are i.i.d. random vectors.

A.2. Suppose for all $d$ and $\nu$, $E [f_1 (v, d, \theta_0, 0)] = E [f_2 (v, \nu, \theta_0, 0)] = 0$.

A.3. Both Jacobians $\partial f_1 (v, d, \theta, \lambda) / \partial (\theta, \lambda)$ and $\partial f_2 (v, \nu, \theta, \kappa) / \partial (\theta, \kappa)$ and Hessians $\partial^2 f_1 (v, d, \theta, \lambda) / \partial^2 (\theta, \lambda)$ and $\partial^2 f_2 (v, \nu, \theta, \kappa) / \partial^2 (\theta, \kappa)$ exists for all $\theta$, $\lambda$ and $\kappa$ and the Jacobian matrices are continuous in the neighbourhood of the true value $(\theta_0, 0)$.

A.4 a. With $\| \cdot \|$ denoting the Euclidean norm, suppose that $\| \partial f_1 (v, d, \theta, \lambda) / \partial (\theta, \lambda) \|$, $\| f_1 (v, d, \theta, \lambda) \|^2$ and $\| \partial^2 f_1 (v, d, \theta, \lambda) / \partial^2 (\theta, \lambda) \|$ are bounded by $\mathcal{G}(v, d)$ for some integrable function $\mathcal{G}$ in the neighbourhood of $(\theta_0, 0)$.
A.4 b. $|\partial f_2 (v, \nu, \theta, \kappa) / \partial (\theta, \kappa)|$, $|f_2 (v, \nu, \theta, \kappa)|^3$ and $|\partial^2 f_2 (v, \nu, \theta, \kappa) / \partial^2 (\theta, \kappa)|$
are bounded by $\delta f(v, \nu)$ for some integrable function $\delta f$ in the neighbourhood of $(\theta_0, 0)$.

A.5. Both $E \left[ f_1 (v, d, \theta_0, 0) f_1 (v, d, \theta_0, 0)^T \right]$ and $E \left[ f_2 (v, \nu, \theta_0, 0) f_2 (v, \nu, \theta_0, 0)^T \right]$ are positive definite matrices.

A.6. Both $E \left[ \partial f_1 (v, d, \theta_0, 0) / \partial (\theta, \lambda) \right]$ and $E \left[ \partial f_2 (v, \nu, \theta_0, 0) / \partial (\theta, \kappa) \right]$ have full ranks.

Further we denote:

(43) \( \psi' (y, a, \theta) = \partial \psi (y, a, \theta) / \partial \theta, G = E [d_1 \psi' (y_1, a_1, \theta_0)] \), \( G^* = E [d_1^2 \psi^2 (y_1, a_1, \theta_0)] \),
(44) \( K_1 = E [d_2 \psi (y_1, a_1, \theta_0) h(v_1, \gamma)], K_2 = E [d_1^2 \psi (y_1, a_1, \theta_0) h(v_1, \gamma)] \),
(45) \( H_1 = E \left[ d_1^2 h^2 (v_1, \gamma) \right], H_2 = E \left[ d_1^4 h^2 (v_1, \gamma) \right] \),
(46) \( G = E [\psi' (y_1, a_1, \theta_0) / \nu_1], G^* = E \left[ \psi^2 (y_1, a_1, \theta_0) / \nu_1^2 \right] \),
(47) \( \mathbb{K}_2 = E \left[ \psi (y_1, a_1, \theta_0) h(v_1, \gamma) / \nu_1^2 \right], \mathbb{H}_2 = E \left[ h^2 (v_1, \gamma) / \nu_1^2 \right] \).

We first prove the strong consistency and the asymptotic normality of \( \hat{\theta}_{CS}^{(n)} \).

**Theorem 2.** Suppose that assumptions A 1-4 a., A 5. and A 6. hold. Then, almost surely the equation \( \sum_{i=1}^n f_1 (v_i, d_i, \theta, \lambda) = 0 \) admits a sequence of solutions \( (\hat{\theta}_{CS}^{(n)}, \hat{\lambda}^{(n)}) \) such that

1. \( (\hat{\theta}_{CS}^{(n)}, \hat{\lambda}^{(n)}) \longrightarrow (\theta_0, 0) \) as \( n \to \infty \),
2. \( n^{1/2} (\hat{\theta}_{CS}^{(n)} - \theta_0) \Rightarrow N (0, V_1) \) distribution, where \( V_{CS} = G^{-1} \left( G^* - K_1 H_1^{-1} K_2^T - K_2 H_1^{-1} K_1^T + K_1 H_1^{-1} H_2 H_1^{-1} K_1^T \right) (G^T)^{-1} \),
3. \( n^{1/2} \hat{\lambda}^{(n)} \Rightarrow N (0, H_1^{-1} H_2 H_1^{-1}) \) distribution,
4. Asymptotic covariance of \( \hat{\theta}_{CS}^{(n)} \) and \( \hat{\lambda}^{(n)} \) is given by \( G^{-1} \left( K_1 H_1^{-1} H_2 - K_2 \right) H_1^{-1} \).

**Proof.** See Appendix.

We next prove that \( \hat{\theta}_{CL}^{(n)} \) is strongly consistent and asymptotically normally distributed.

**Theorem 3.** Suppose that assumptions A 1-4 b., A 5. and A 6. hold. Then, almost surely the equation \( \sum_{i=1}^n f_2 (v_i, u_i, \theta, \kappa) = 0 \) admits a sequence of solutions \( (\hat{\theta}_{CL}^{(n)}, \hat{\kappa}^{(n)}) \) such that

1. \( (\hat{\theta}_{CL}^{(n)}, \hat{\kappa}^{(n)}) \longrightarrow (\theta_0, 0) \) as \( n \to \infty \),
2. \( n^{1/2} (\hat{\theta}_{CL}^{(n)} - \theta_0) \Rightarrow N (0, V_2) \) distribution where \( V_{CL} = G^{-1} \left( G^* - \mathbb{K}_2 H_2^{-1} \mathbb{K}_1^T \right) (G^T)^{-1} \),
3. \( n^{1/2} \hat{\kappa}^{(n)} \Rightarrow N (0, H_2^{-1}) \) distribution,
4. \( \hat{\theta}_{CL}^{(n)} \) and \( \hat{\kappa}^{(n)} \) are asymptotically independent.

**Proof.** See Appendix.

Suppose \( \hat{\theta}_{PL} \) is the unconstrained pseudo maximum likelihood estimator. It can be shown that the variance-covariance matrix of \( \sqrt{n} (\hat{\theta}_{PL}^{(n)} - \theta_0) \) is given by \( G^{-1} G^* (G^T)^{-1} \). \( \hat{\theta}_{CS} \) and \( \hat{\theta}_{CL} \) are both constrained by population level information. Thus it is natural to expect that these two estimator would be more efficient than \( \hat{\theta}_{PL} \). Theorem 2 does not ensure any reduction in the standard error of \( \hat{\theta}_{CS} \). \( \hat{\theta}_{PL} \) does not use the same weights as \( \hat{\theta}_{CL} \), so no conclusion can be drawn based on
Theorem 3 either. However, in the majority of cases $\hat{\theta}_{CL}$ is more efficient than $\hat{\theta}_{CS}$. Further, as $\hat{\theta}_{CL}^{(n)}$ and $\hat{\lambda}^{(n)}$ are asymptotically independent, the covariance of $\hat{\theta}_{CS}^{(n)}$ and $\hat{\lambda}^{(n)}$ does not converge to 0.

7.1. $\nu = \pi$. We know consider the important special case where $\nu = \pi$. If $\pi_i \perp V_i$ for each $i$, then it is clear that $\nu = \pi$. This independence holds for equally weighed samples. It also holds if, $\pi_i$ is a function of $Z_S^i$ and $Z_R^i \perp (Y_S, X_S)$. If $\pi_i$ is a function of $Z_i$, then by assumption (see Figure 1), $\mu_i = E[\pi_i \mid V_i] = E[\pi_i \mid Z_i] = \pi_i$. So $\nu = \pi$.

Clearly if $\nu = \pi$, $\hat{\theta}_{PL}$ and $\hat{\theta}_{CL}$ both depend on the same weights and from Theorem 3 it follows that $\hat{\theta}_{PL}$ is less efficient than the latter. In fact, in the following theorem we show that $\hat{\theta}_{CL}$ is more efficient than $\hat{\theta}_{CS}$ as well.

**Theorem 4.** If $\nu_i = \pi_i$, for all $i = 1, 2, \ldots, n$, the asymptotic standard error of $\hat{\theta}_{CS}$ is larger than $\hat{\theta}_{CL}$.

*Proof.* See Appendix. \qed

If the design puts equal probability on all sampled observations, $\hat{\theta}_{CS} = \hat{\theta}_{CL}$. If $\pi_i \perp V_i$, $\hat{\theta}_{CL}$ is only slightly more efficient than $\hat{\theta}_{CS}$. If all design variables are observed, i.e. the third situation described above, $\hat{\theta}_{CL}$ gains a lot over $\hat{\theta}_{CS}$ in terms of efficiency. In fact, even if $\nu \neq \pi$, $\hat{\theta}_{CL}$ usually has lower standard error than $\hat{\theta}_{CS}$. Kim [2009] considers estimation of population mean and discusses conditions when $\hat{\theta}_{CS}$ could be more efficient than $\hat{\theta}_{CL}$. However, his simulation studies as well as ours (not presented here) show no major gain in efficiency for $\hat{\theta}_{CS}$ in any situation.

### 7.2. Estimating asymptotic covariance matrices.

The asymptotic covariance matrices of $\hat{\theta}_{CS}$ and $\hat{\theta}_{CL}$ can be estimated directly from their respective expressions. In particular the estimates are given by:

$$
\hat{\mathbf{G}} = \sum_{i=1}^{n} \hat{w}_{CSi} d_i \psi' \left( y_i, a_i, \hat{\theta}_{CS} \right), \hat{\mathbf{G}}^* = \sum_{i=1}^{n} \hat{w}_{CSi}^2 d_i^2 \psi^2 \left( y_i, a_i, \hat{\theta}_{CS} \right),
$$

$$
\hat{\mathbf{K}}_1 = \sum_{i=1}^{n} \hat{w}_{CSi} d_i \psi \left( y_i, a_i, \hat{\theta}_{CS} \right) h(v_i, \gamma), \hat{\mathbf{K}}_2 = \sum_{i=1}^{n} \hat{w}_{CSi}^2 d_i^2 \psi \left( y_i, a_i, \hat{\theta}_{CS} \right) h(v_i, \gamma),
$$

$$
\hat{\mathbf{H}}_1 = \sum_{i=1}^{n} \hat{w}_{CSi} d_i h^2(v_i, \gamma), \hat{\mathbf{H}}_2 = \sum_{i=1}^{n} \hat{w}_{CSi}^2 d_i^2 h^2(v_i, \gamma),
$$

$$
\hat{\mathbf{G}} = \sum_{i=1}^{n} \hat{w}_{CLi} \psi' \left( y_i, a_i, \hat{\theta}_{CL} \right) / \nu_i, \hat{\mathbf{G}}^* = \sum_{i=1}^{n} \hat{w}_{CLi}^2 \psi^2 \left( y_i, a_i, \hat{\theta}_{CL} \right) / \nu_i^2,
$$

$$
\hat{\mathbf{K}}_2 = \sum_{i=1}^{n} \hat{w}_{CLi} \psi \left( y_i, a_i, \hat{\theta}_{CL} \right) h(v_i, \gamma) / \nu_i^2, \hat{\mathbf{K}}_2 = \sum_{i=1}^{n} \hat{w}_{CLi}^2 h^2(v_i, \gamma) / \nu_i^2.
$$

The estimated values $\hat{V}_{CS}$ and $\hat{V}_{CL}$ at $\hat{\theta}_{CS}$ and $\hat{\theta}_{CL}$ can be found by substituting the above expression in the formulas in Theorems 2 and 3.

### 8. Application to demographic hazard modelling in complex longitudinal surveys

In this section we apply conditional empirical likelihood based methods to two major datasets. We compare the performance of $\hat{\theta}_{CL}$ with several estimators including $\hat{\theta}_{PL}$ and $\hat{\theta}_{CS}$. In each case $\hat{\theta}_{CL}$ turns out to be a better estimator than the other two in terms of efficiency. We compare the estimators first on a dataset with
a moderate but ignorable amount of inequality in sampling weights due to its multi-stage clustered design \(\text{(the British Household Panel Survey, BHPS)}\), and second on a dataset with a high amount of inequality in sample weights due to its inclusion of a low-income “oversample” \(\text{(the Panel Study of Income Dynamics, PSID)}\). For the BHPS we use a simple specification of the model and with a single constraint value. The computational complexity of this model under the CMLE framework means that we can compare only the different empirical likelihood estimators, notably the CL estimator with the CS estimator.

8.1. Application to the British Household Panel Survey data. We combine the survey data from the British Household Panel Survey \(\text{(BHPS)}\) \[\text{[Taylor et al., 1995]}\] with population-level information on the general fertility Rate \(\text{(GFR)}\) obtained from the birth registry system population level information. The BHPS data were not collected with equal probability, but because there was no sociodemographic or economic oversampling often the weights are ignored \(\text{(see, e.g., Handcock et al. [2000], Chaudhuri et al. [2008]}\)).

The dependent variable \(Y\) represents the indicator of birth for a woman between time points \(t - 1\) and \(t\). The explanatory variable \(X\) is the indicator of existence of at least one child of the woman at time point \(t - 1\). The GFR for the years 1992 – 1996 was calculated to be 0.06179. We omit the details which can be found in Chaudhuri et al. [2008].

We fit a logistic regression model to the data and constraint on the GFR. In particular the model is represented as

\[
\log\text{-odds}(Y = 1 | X = x) = \beta_0 + \beta_1 x.
\]

The population level information is given by:

\[
\sum_{i=1}^{n} w_i y_i = .06179.
\]

We use this example mainly to compare various parametric and empirical likelihood based estimators with or without constraints and weights.

In addition to the unconstrained weighted \(\text{(PL)}\) estimator and Chen-Sitter \(\text{(CS)}\) estimator, we also consider the constrained but unweighted estimators obtained from constrained parametric \(\text{(CNWP)}\) \[\text{[Handcock et al., 2005]}\] and empirical likelihood \(\text{(CNWE)}\) \[\text{[Chaudhuri et al., 2008]}\] based methods. The unconstrained and unweighted estimator \(\text{(NCNW)}\) is also considered. Two forms of CL estimator are considered. The first one is denoted CL-P, which assumes the distribution of \(\pi_i | V_i\) in the sample is not different from the population. CL-S on the other hand assumes to distributions are different and uses \(21\). In both cases \(\nu\) is obtained through a Gamma regression with canonical inverse link and \(X\) as covariate. In the first case \(\pi\) is the response and \(\nu\) is taken to be the fitted values. In the second the response is \(1/\pi\) and the inverse fitted values were taken to be \(\nu\).

In Figure 2 and Table 1 we present the estimated values of \(\beta_0\) and \(\beta_1\) together with their standard errors obtained from various methodologies. We use standard errors of these estimates as a measure for comparison among the various methodologies.

The population level information directly constrains \(\beta_0\). Thus when compared to un-constrained estimators \(\text{(NCNW and PL)}\) the standard errors of the constrained estimators \(\hat{\beta}_0\) are reduced more than those of \(\hat{\beta}_1\).

From Table 1 it follows that the divergence based methods \(\text{(i.e., PL and CS)}\) have in general higher variance than the likelihood based methods. The unweighted
Figure 2. Comparison of the two constrained and the unconstrained estimates (error bars are 1 standard errors above and below the point estimates).

Table 1. Parameter estimates and their standard errors obtained from the BHPS data using various methods. PL: Weighted pseudo-likelihood method, CL-S: CL estimator with \( \nu \) estimated from (21), CL-P: CL estimator with \( \nu \) estimated from the sample, CS: Chen-Sitter estimator, CNWP: Unweighted CMLE \( (Pr(x = 1) \text{ was estimated as a parameter}) \), CNW: Constrained un-weighted ELE, NCNW: Not constrained un-weighted estimator.

<table>
<thead>
<tr>
<th>Methods</th>
<th>( \hat{\beta}_0 )</th>
<th>( se(\hat{\beta}_0) )</th>
<th>( \hat{\beta}_1 )</th>
<th>( se(\hat{\beta}_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL</td>
<td>-3.181</td>
<td>0.072</td>
<td>0.5475</td>
<td>0.093</td>
</tr>
<tr>
<td>CL-S</td>
<td>-2.935</td>
<td>0.049</td>
<td>0.4151</td>
<td>0.087</td>
</tr>
<tr>
<td>CL-P</td>
<td>-2.947</td>
<td>0.050</td>
<td>0.4265</td>
<td>0.087</td>
</tr>
<tr>
<td>CS</td>
<td>-3.011</td>
<td>0.055</td>
<td>0.5476</td>
<td>0.093</td>
</tr>
<tr>
<td>CNWP</td>
<td>-2.962</td>
<td>0.052</td>
<td>0.4393</td>
<td>0.087</td>
</tr>
<tr>
<td>CNW</td>
<td>-2.962</td>
<td>0.052</td>
<td>0.4392</td>
<td>0.087</td>
</tr>
<tr>
<td>NCNW</td>
<td>-3.129</td>
<td>0.067</td>
<td>0.4392</td>
<td>0.087</td>
</tr>
</tbody>
</table>

constrained empirical likelihood based estimator (CNW) is essentially as efficient as the constrained parametric estimator (CNWP). The constraint improves the estimate for \( \beta_0 \). However CS seems to be much worse than the two CL estimators as well as CNWP and CNW. Both CL-P and CL-S are better than CNWP and CNW. Further CL-S dominates CL-P.

8.2. Application to the Panel Study of Income Dynamics data. We use 1985-97 years of the data from the Panel Study of Income Dynamics (PSID, Institute for Social Research [2010]) in combination with population information from the National Center for Health Statistics (NCHS) age-specific first birth probabilities in the US to estimate the relationship between the probability of first birth to age and other socio-demographic factors.

We use data from both the Survey Research Centre (SRC) and Survey of Economic Opportunity (SEO) subsamples from the PSID. When used with the sample
weights supplied with the data, the combined subsamples are designed to be nationally representative. While the SRC subsample is close to equal probability sampling, in the SEO subsample the low-income group was over sampled. Thus the weights of the SEO-low-income observations (Family IDs > 5,000) are about one-tenth of those in the main SRC sample (Family IDs < 3,000, see Figure 3(b)). We see that the age specific first birth probabilities in USA, obtained from NCHS, have a bimodal shape (Figure 3(a)). The probability initially peaks at the age 20 then
drops off and then climbs again to a maximum at age 28 (see also Sullivan [2005]).
In contrast, the same probabilities obtained from the weighted PSID sample show
an upward trend but less clear bimodality.

The probability of giving birth are modelled as:

$$\text{log-odds}(I^{(B)}_i) = \beta_0 + \beta_M I^{(M)}_i + \beta_{PM} I^{(PM)}_i + \beta_E I^{(E)}_i + \beta_W I^{(W)}_i + \beta_{NHS} I^{(NHS)}_i + \beta_{NHNS} I^{(NHNS)}_i + \beta_{HNS} I^{(HNS)}_i + \beta_{COLL} I^{(COLL)}_i + \sum_{k=18}^{30} \beta_k I^{(k)}_i + \beta_{MD} \cdot MD_i + \beta_{MD^2} \cdot MD_i^2.$$  

(50)

Here $I^C_i = 1$ if observation $i$ is in category $C$. The notations for different
categories denote:

- $B =$ Birth in the year $t-1$ to $t$
- $M =$ Married at $t-1$
- $PM =$ Previously married at $t-1$
- $E =$ Employed at $t-1$
- $W =$ White
- $NHS =$ Non high school graduate and in school at $t-1$
- $NHNS =$ Non high school graduate and not in school at $t-1$
- $HS =$ High School graduate at $t-1$
- $COLL =$ Any college education at $t-1$
- $k =$ Age, $k = 18, \ldots, 30$
- $MD =$ Marital duration at $t-1$

Both SRC and SEO samples recorded if a person is Hispanic or not. However
there are only 100 Hispanic people in those two samples. Due to the small sample
size we only consider two categories for the race variable, namely white and non-
white.

The weight ($d_i$) of an observation depends on its race (white/non-white) and also
on the sample (i.e. SRC/SEO) in comes from. Suppose $I^{(S)}_i = 1$ if the observation
is from the SRC sample.

We first specify $\nu$. From the context the sample indicator and the race variable
play crucial roles in determining the weights. Other design variables are not avail-
able. We assume $V$ consists of all the variables defined above including $I^{(S)}$ and
that $A$ is the set of all variables appearing in (50).

It is also not clear if the distribution of $\pi \mid V_i$ in the sample is same as in the
population. Thus we use (21) to estimate $\nu$. We take $E_S [d_i \mid V_i]$ to be the fitted
values from a Gamma regression of $d$ on $V$. The link function is taken to be the
inverse (canonical) link function. The model for the mean function $\mu$ was taken to be:

$$\mu_i = \alpha_0 + \sum_{k=18}^{30} \alpha_k I^{(k)}_i + \alpha_W I^{(W)}_i + \alpha_S I^{(S)}_i + \alpha_{SW} \left( I^{(W)}_i \cdot I^{(S)}_i \right).$$  

(51)

We use NCHS values of the age specific first birth probability Schoen [2005] as
our population level constraints. For $k = 17, 18, \ldots, 30$ the constraints are given by:

$$\sum_{i=1}^{n} w_i I^{(k)}_i \left( I^{(B)}_i - \gamma_k \right) = 0.$$  

(52)
Following Handcock et al. [2005], we expect only the “directly-constrained” coefficients $k$ for ages (plus the intercept representing the reference age) to have their standard errors substantially reduced by the introduction of population information in the constraint, even while all standard errors will be at least as low as for the unconstrained model.

**Figure 4.** Parameter estimates and their asymptotic standard errors for PSID data.
The parameter estimates and their standard errors are presented in Figure 4. From the figure it is clear that $\hat{\theta}_{CL}$ is more efficient than both $\hat{\theta}_{PL}$ and $\hat{\theta}_{CS}$ and that $\hat{\theta}_{CS}$ is less efficient than $\hat{\theta}_{PL}$ for some variables. In fact, other than for $I^{(M)}$, $MD$ and $MD^2$ the standard errors for $\hat{\theta}_{CL}$ are much less than those for $\hat{\theta}_{CS}$. Note that here $\pi \neq \nu$, thus Theorem 4 does not hold. However, $\hat{\theta}_{CL}$ turns out to be better than $\hat{\theta}_{CS}$. On average, the standard errors for the age coefficients are about 10% lower for our CL estimator than the CS estimator.

The weights $d_i$ have a large variation (see again Figure 3(b)). In the presence of such extreme weights the usual weighted unconstrained estimator $\hat{\theta}_{PL}$ is known to produce unstable estimates of the parameters. We validate these asymptotic values with bootstrap re-sampling. Based on 5000 re-samples it appears that the bootstrap standard errors for most of the components of $\hat{\theta}_{PL}$, $\hat{\theta}_{CS}$ and $\hat{\theta}_{CL}$ are slightly larger but close to their analytical values. The bootstrap standard errors of the coefficients corresponding to the variables $NHNS$, $NHS$ and $COLL$ for the unconstrained estimate is larger than their asymptotic estimates. However the relative ordering between the estimates in terms of standard errors remains the same.

9. Discussion

In this article we present a new method to include design weights in an empirical likelihood based estimation procedure. We also incorporate population level information in statistical modelling based on sample data. Typically, in sample surveys, observations are selected with unequal probabilities due to purposive “oversampling”, clustering, stratification, post-stratification, attrition and other non-response adjustments. For such surveys, the observed distribution of the sampled observations are different from their distribution in the population. We adapt a parametric conditional likelihood (Pfeffermann et al. [1998] etc.) to empirical likelihood and include population level information in the analysis. Information about the model and the population are introduced as equality restrictions through estimating equations. The parameter estimates are obtained by maximising the empirical likelihood under these constraints by a two-step procedure. The product of the weights can be interpreted as a non-parametric likelihood of the sample under the true population distribution. We assume that the sampling weights contain all information about the design. The expectation of these sampling weights conditional on the observed variables are used in the analysis. In this way, we can incorporate all design variables, even those such as the cluster identifiers that may not be available to the researcher in the given survey dataset.

Our estimator is related to the inverse probability weighted general Horvitz-Thompson type estimator. In fact, without any population level restriction, the parameter estimates are obtained by solving the score equations weighted by inverse of the conditional expectation of weights. This justifies the Horvitz-Thompson estimator with random weights and shows that it can be derived from a likelihood perspective.

Empirical likelihood based methodologies have huge computational and implementational advantages over the corresponding constrained maximum likelihood estimators. Direct non-linear equality constraints on the parameters often make computation infeasible. Empirical likelihood based methods put linear constraints on the weights, which can be implemented quite easily. As we show in Section 5, the two-step estimator used by Chaudhuri et al. [2008] can be adapted to obtain the
estimates. Analytic expressions of the asymptotic standard errors of the estimates are also known.

It is known that empirical likelihood and the estimators based on them (ELE) have many desirable properties. Owen [2001] shows that the corresponding Wald statistics has an asymptotic Chi-squared limit for i.i.d. observations. Similar results follow for various kinds of dependence as well. The constrained empirical likelihood can be expressed as a profile likelihood for $\theta$. Qin and Lawless [1994] show that, under usual regularity conditions, $\hat{\theta}$ is asymptotically unbiased and normally distributed.

Handcock et al. [2005] and Chaudhuri et al. [2008] show that it is beneficial to include available population level information in statistical modelling. Such information is guaranteed to reduce the standard error of the estimates.

Our method has one more advantage over its parametric counterpart. The parametric conditional likelihood involves a high-dimensional integral. In most cases, the analytic form of this integral cannot be found. We substitute the integral with a sum which can be computed instantly.

For empirical likelihood based estimators, one does not need specify any parametric form for the likelihood. They are therefore more flexible avoid unnecessary assumptions on the distribution of $A$. They also do not lose much efficiency compared to the CMLE of a correctly specified parametric model. In fact, if the underlying distribution is misspecified, often our ELE would be more efficient than the corresponding CMLE. For illustrative examples, we refer to Chaudhuri, Drton, and Richardson [2007]. The relation between our estimator to the inverse probability weighted general Horvitz-Thompson estimator is particularly interesting in this respect. It is known that Horvitz-Thompson estimators are usually robust against model misspecification. It is an intriguing possibility that our estimate inherits a part of this robustness as well. The robustness against model misspecification is particularly beneficial, since it is often difficult to specify correct distributions for the design variables. The proposed estimator requires specification of two models. One for the response of interest and the auxiliary variables, the other for the sampling probabilities. Both can be specified from the background information.

Our estimator differs from the Horvitz Thompson type minimum divergence estimator (Chen and Sitter [1999] etc.). In real data applications, we show that our estimator is likely to be more efficient than this minimum divergence estimator in most situations. The design information is incorporated through conditional expectation of the sampling probabilities. This makes it different from the estimator used by Kim [2009].

The constrained maximisation problems to estimate the weights in the empirical likelihood can be easily solved using the algorithms in Owen [2001] and Chen et al. [2002]. For generalised linear models both have been implemented in open-source software developed by the authors [Chaudhuri et al., 2010]. We will extend the package to make the methods developed in this paper freely available also.

The parametric model for the conditional expectation of the sampling weights is usually of secondary interest. It is probably more profitable to use non-parametric procedures instead. Such a procedure may produce a better estimate of the conditional expectation and also lead to a better approximation to the probability of inclusion of an element in a sample. However, in the presence of a large number of auxiliary variables, it is not clear which non-parametric procedure would be optimal for such purpose.
Our likelihood based semi-parametric approach has several advantages over the fully parametric methods. For a fully parametric approach one needs to specify a parametric candidate for $F_0$. This is usually difficult and in practice the model can be misspecified. $\hat{\theta}_{CL}$ is more efficient than a misspecified CMLE. To use the empirical likelihood based method, one only needs to specify $(Y, A)$ and $E_P[\pi | V]$. These can be done using the background knowledge.

Even if a parametric model for $F_0$ is correctly specified, $\hat{\theta}_{CL}$ is almost as efficient as the corresponding likelihood based estimator. Computationally the empirical likelihood based method is far less demanding. It is more so since we replace $D$ by $P_i w_i$ avoiding direct computation of a high-dimensional integral. Further, the population level constraints are nonlinear functions of the model parameters. Maximisation under nonlinear inequality constraints usually involves heavy computation. The conditional empirical likelihood can be used as a likelihood in Bayesian procedures. In particular, this may be applied to Bayesian analysis in problems in sample surveys, small area estimation, epidemiology, case-control studies, among others.

**Appendix A. Proofs**

In this section we present the proofs of the theorems.

**Proof of Lemma 1**

Proof. Consider the objective function in (23). Clearly for an extremum

\[
0 = \frac{\partial L}{\partial w_i} = \frac{1}{w_i} - \frac{m
u_i}{\sum_{i=1}^n w_i \nu_i}
\]

(53)

\[w_i \Rightarrow \frac{\sum_{i=1}^n w_i \nu_i}{m
u_i}
\]

(54)

Now from $\sum_{i=1}^n w_i = 1$ we get $\sum_{i=1}^n w_i \nu_i / n = 1/ \sum_{i=1}^n 1/ \nu_i$. Thus $w_i = (1/ \nu_i) / \sum_{i=1}^n 1/ \nu_i$ and the result follows. \hfill $\Box$

**Proof of Lemma 1**

Proof. Following Owen [2001] it can be shown that $\hat{w}_i^* = \{n(1 + \xi^* h_i / \nu_i)\}^{-1}$ with $(1 + \xi^* h_i / \nu_i) > n^{-1}$ for all $i = 1, 2, \ldots, n$, where $\xi^*$ is the unique optimal value of the Lagrange multiplier $\xi$.

Further $\xi^*$ satisfies:

\[
\sum_{i=1}^n \frac{h_i / \nu_i}{1 + \xi^* h_i / \nu_i} = 0,
\]

(55)

which by uniqueness implies $\kappa = \xi^*$.

By denoting $\hat{D} = \sum_{i=1}^n \nu_i \hat{w}_{CL,i}$ and by comparing $\hat{w}_i^*$ and $\hat{w}_{CL,i}$ from (28) we notice that

\[
\hat{w}_i^* = \frac{\nu_i \hat{w}_{CL,i}}{\hat{D}}
\]

(56)

Summing over $i$ and noting that $\sum_{i=1}^n \hat{w}_{CL,i} = 1$ it follows that $\hat{D} = (\sum_{i=1}^n \hat{w}_i^* / \nu_i)^{-1}$. Substituting these results in (28) it follows that

\[
\hat{w}_{CL,i} = \frac{\hat{w}_i^* / \nu_i}{\sum_{i=1}^n \hat{w}_i^* / \nu_i}
\]

(57)

Further note that by the non negativity of $w_i^*$

\[
\hat{D}^{-1} = \sum_{i=1}^n \frac{\hat{w}_i^*}{\nu_i} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\nu_i + kh_i} > \frac{1}{n (\nu_i + kh_i)}
\]

(58)

for all $i = 1, 2, \ldots, n$, which is the restriction on $\hat{w}_{CL,i}$ in (28). \hfill $\Box$
The next two proofs closely follow Chaudhuri et al. [2008]. We only present the sketch of the arguments here. The details can be found in the above reference.

**Proof of Theorem 2**

**Proof.** Note that:

\[
(-1) \cdot \frac{\partial f_1(v_1, d_1, \theta, \lambda)}{\partial (\theta, \lambda)} = \left( \begin{array}{c}
-\frac{d_1 \varphi(v_1, \theta)}{1 + \lambda h(v_1, \gamma)} & \frac{d_1 \varphi(v_1, \theta) h(v_1, \gamma)}{(1 + \lambda h(v_1, \gamma))^2} \\
0 & \frac{d_1 h^2(v_1, \gamma)}{(1 + \lambda h(v_1, \gamma))^2}
\end{array} \right).
\]

Thus

\[
(-1) \cdot E \left\{ \frac{\partial f_1(v_1, d_1, \theta, \lambda)}{\partial (\theta, \lambda)} \bigg|_{\theta = \theta_0, \lambda = 0} \right\} = \left( \begin{array}{c}
-G & K_1 \\
0 & H_1
\end{array} \right).
\]

Further

\[
\text{Var} \{ f_1(v_1, d_1, \theta_0, 0) \} = \left( \begin{array}{c}
G^T & K_2 \\
K_2^T & H_2
\end{array} \right).
\]

Now by expanding \( \sum_{i=1}^{n} f_1(v, d, \hat{\theta}_n, \hat{\lambda}_n) / n \) around \((\theta_0, 0)\), under the assumptions, the results can be shown via standard techniques.

In particular, \( \sqrt{n} \left( (\hat{\theta}_n - \theta_0), \hat{\lambda}_n \right)^T \) converges to a normal distribution with covariance matrix:

\[
\begin{pmatrix}
-G & K_1 \\
0 & H_1
\end{pmatrix}^{-1}
\begin{pmatrix}
G^T & K_2 \\
K_2^T & H_2
\end{pmatrix}
\begin{pmatrix}
-G & 0 \\
0 & H_1
\end{pmatrix}^{-1}
\begin{pmatrix}
G^* - K_1 H_1^{-1} K_2^* - K_2 H_1^{-1} K_1^* + K_1 H_1^{-1} H_2 H_1^{-1} K_1^* \\
H_1^{-1} (K_2^* - H_2^{-1} H_1 K_1^*) G_1^{-1}
\end{pmatrix}
\begin{pmatrix}
G^* - K_1 H_1^{-1} H_2 - K_2 \\
H_1^{-1} H_2 H_1^{-1}
\end{pmatrix}
\]

\[=\]

\[\square\]

**Proof of Theorem 3**

**Proof.** The proof is similar to that of Theorem 2. Note that :

\[
(-1) \cdot \frac{\partial f_2(v_1, \nu, \theta, \kappa)}{\partial (\theta, \kappa)} = \left( \begin{array}{c}
-\frac{\varphi(v_1, \nu)}{1 + \kappa h(v_1, \gamma)} & \frac{\varphi(v_1, \nu) h(v_1, \gamma)}{(1 + \kappa h(v_1, \gamma))^2} \\
0 & \frac{d_1 h^2(v_1, \gamma)}{(1 + \kappa h(v_1, \gamma))^2}
\end{array} \right).
\]

Thus

\[
(-1) \cdot E \left\{ \frac{\partial f_2(v_1, \nu, \theta, \kappa)}{\partial (\theta, \kappa)} \bigg|_{\theta = \theta_0, \kappa = 0} \right\} = \left( \begin{array}{c}
-G & K_2 \\
0 & H_2
\end{array} \right).
\]

Similar to Theorem 2 we know that:

\[
\text{Var} \{ f_2(v_1, \nu, \theta_0, 0) \} = \left( \begin{array}{c}
G_2^T & K_2 \\
K_2^T & H_2
\end{array} \right)
\]

Now from the expansion of \( \sum_{i=1}^{n} f_2(v, \nu, \theta_0, 0) / n \) in the neighbourhood of \((\theta_0, 0)\) as before, the results follow.

Furthermore, as before the asymptotic variance of \( \sqrt{n} \left( (\hat{\theta}_n - \theta_0), \hat{\lambda}_n \right)^T \) is given by:

\[
\begin{pmatrix}
-G & K_2 \\
0 & H_2
\end{pmatrix}^{-1}
\begin{pmatrix}
G_2^T & K_2 \\
K_2^T & H_2
\end{pmatrix}
\begin{pmatrix}
-G & 0 \\
0 & H_2
\end{pmatrix}^{-1}
\begin{pmatrix}
G_2^* - K_2 H_2^{-1} K_2^* \\
H_2^{-1} K_2^*
\end{pmatrix}
\begin{pmatrix}
G_2^* - K_2 H_2 \\
H_2^{-1} H_2 H_2^{-1}
\end{pmatrix}
\]

\[=\]

\[\square\]

**Proof of Theorem 4**
Proof. Since $\nu = \pi$, $G = G^* = G^*$, $K_2 = K_2$ and $H_2 = H_2$.
\[
G(V_{CS} - V_{CL})G^T = K_2H_2^{-1}K_2^T - K_1H_1^{-1}K_1^T + K_2H_2^{-1}K_2^T - K_2H_2^{-1}K_2^T + K_1H_1^{-1}H_2H_1^{-1}K_1^T
\]
\[
= K_2\left(H_2^{-1}K_2^T - H_1^{-1}K_1^T\right) + K_1H_1^{-1}\left(H_2H_1^{-1}K_1^T - K_2^T\right)
\]
\[
= (K_2 - K_1H_1^{-1}H_2)\left(H_2^{-1}K_2^T - H_1^{-1}K_1^T\right)
\]
\[
= (K_2H_2^{-1} - K_1H_1^{-1})H_2\left(H_2^{-1}K_2^T - H_1^{-1}K_1^T\right)
\]
Clearly $G(V_{CS} - V_{CL})G^T$ is a non-negative definite matrix. So $V_{CS} - V_{CL}$ is n.n.d. as well. 

\section*{References}


Institute for Social Research. \textit{The panel Study of Income dynamics}. University of Michigan, 2010. URL \url{http://psidonline.isr.umich.edu/}.


