Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

# Authors' response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data":

# Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples


by Jean-François Beaumont, Keven Bosa, Andrew Brennan,
Joanne Charlebois and Kenneth Chu

Release date: June 25, 2024

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**Email at** infostats@statcan.gc.ca

**Telephone,** from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

| | |
|---|---|
| • Statistical Information Service | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line | 1-514-283-9350 |

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “Contact us” > “Standards of service to the public.”

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada



**An HTML version is also available.**

*Cette publication est aussi disponible en français*.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 123-141
Statistics Canada, Catalogue No. 12-001-X


# Authors' response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data":

# Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples

**Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu**[1]

Abstract

Inspired by the two excellent discussions of our paper, we offer some new insights and developments into the problem of estimating participation probabilities for non-probability samples. First, we propose an improvement of the method of Chen, Li and Wu (2020), based on best linear unbiased estimation theory, that more efficiently leverages the available probability and non-probability sample data. We also develop a sample likelihood approach, similar in spirit to the method of Elliott (2009), that properly accounts for the overlap between both samples when it can be identified in at least one of the samples. We use best linear unbiased prediction theory to handle the scenario where the overlap is unknown. Interestingly, our two proposed approaches coincide in the case of unknown overlap. Then, we show that many existing methods can be obtained as a special case of a general unbiased estimating function. Finally, we conclude with some comments on nonparametric estimation of participation probabilities.

**Key Words:** Best linear unbiased estimation; Best linear unbiased prediction; Estimating equation; Population likelihood; Pseudo likelihood; Sample likelihood.

## 1. General remarks

We would first like to thank the discussants for taking the time to read our paper and share their thoughtful and insightful observations on inverse probability weighting for non-probability samples. Dr. Wu shed light into three common weighting methods for non-probability samples whereas Dr. Gershunskaya and Dr. Beresovsky introduced two new methods to us: Implicit Logistic Regression (ILR), see also Beresovsky (2019), Savitsky, Williams, Gershunskaya, Beresovsky and Johnson (2022) and Gershunskaya and Lahiri (2023), and Pseudo-ILR. We have greatly enjoyed reading both discussions, which have helped us improve our knowledge on the field and stimulated us to have further thoughts on the topic. In what follows, we will share these thoughts along with some new developments.

Sections 2, 3 and 4 are devoted to the methods of Chen, Li and Wu (2020), Wang, Valliant and Li (2021) and Elliott (2009), see also Elliott and Valliant (2017), respectively. We provide additional observations on

1. Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca, keven.bosa@statcan.gc.ca, andrew.brennan@statcan.gc.ca, joanne.charlebois@statcan.gc.ca and kenneth.chu@statcan.gc.ca.

these methods and make connections with the ILR and Pseudo-ILR methods. We show that all three methods are valid in the sense that they lead to unbiased estimating functions for the parameters of the participation model, regardless of the size of the probability and non-probability samples as well as the size of the overlap between both samples. However, only the Chen-Li-Wu (CLW) method has the property of reducing to the maximum likelihood method when the probability sample is a census, which we refer to as the Census Likelihood (CL) property. In Section 2, we also show that the CLW method does not fully leverage the available auxiliary information, which may result in an inefficient estimating function, particularly when the non-probability sample is larger than the probability sample. Using Best Linear Unbiased (BLU) estimation theory, we propose an improvement of the CLW method that addresses this issue and still satisfies the CL property. In Section 5, we propose a sample likelihood approach, similar in spirit to the Elliott/ILR method, that properly accounts for the overlap between both samples provided that it can be identified in one of the samples. Our sample likelihood approach satisfies the CL property. Using BLU prediction theory, we obtain an "optimal" estimating function applicable when the overlap cannot be identified in any of the samples. Interestingly, it is identical to the estimating function underlying our improved CLW method. In Section 6, we unify existing methods that do not require identifying the overlap and show their equivalence for the homogeneous group model. A brief summary is given in Section 7 along with a few comments on nonparametric estimation of participation probabilities.

# 2. A population likelihood approach and the method of Chen, Li and Wu (2020)

We use the notation in our main paper: the vector of auxiliary variables for unit $k$ of the finite population $U$ is denoted by $\mathbf{x}_k,$ and the participation indicator (indicator of participation in the non-probability sample $s_{\text{NP}} \subset U$) for population unit $k \in U$ is denoted by $\delta_k.$ The participation probability $p_k = \Pr\left(\delta_k = 1 \,|\, \mathbf{x}_k\right) > 0$ is modelled using a parametric model, such as the logistic model $p_k(\boldsymbol{\alpha}) = \left[1 + \exp\left(-\mathbf{x}_k'\boldsymbol{\alpha}\right)\right]^{-1},$ where $\boldsymbol{\alpha}$ is a vector of unknown model parameters. We make the following standard independence assumption:

A1) $\delta_k,\ k \in U,$ are mutually independent given $\mathbf{x}_k,\ k \in U.$

Ideally, we would have access to both $\left\{\mathbf{x}_k; k \in s_{\text{NP}}\right\}$ and $\left\{\mathbf{x}_k; k \in U\right\}.$ These two data sets would not need to be linked. Under that ideal scenario and assumption (A1), the population log likelihood function is

$$l(\boldsymbol{\alpha}) = \log\left\{\prod_{k \in U}\left[p_k(\boldsymbol{\alpha})\right]^{\delta_k}\left[1 - p_k(\boldsymbol{\alpha})\right]^{(1-\delta_k)}\right\} = \sum\nolimits_{k \in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right] + \sum\nolimits_{k \in U} \log\left[1 - p_k(\boldsymbol{\alpha})\right]$$

and the population likelihood estimating function (or score function) is

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]} - \sum\nolimits_{k \in U} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]}, \tag{2.1}$$

where $\mathbf{g}_k(\boldsymbol{\alpha}) = \partial p_k(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$. In particular, $\mathbf{g}_k(\boldsymbol{\alpha}) = p_k(\boldsymbol{\alpha})[1 - p_k(\boldsymbol{\alpha})]\mathbf{x}_k$ for the logistic model.

In many real cases, the vector $\mathbf{x}_k$ is not known for the entire population, but is at least available in a probability sample $s_P$ in addition to the non-probability sample $s_{\mathrm{NP}}$. As a result, the term $\sum_{k \in U} \log[1 - p_k(\boldsymbol{\alpha})]$ in the population log likelihood function $l(\boldsymbol{\alpha})$ is not computable as it depends on unknown values of $\mathbf{x}_k$. Chen, Li and Wu (2020) proposed to address this issue by estimating this term using the probability sample. This leads to the pseudo log likelihood function:

$$\hat{l}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right] + \sum_{k \in s_P} w_k \log[1 - p_k(\boldsymbol{\alpha})], \tag{2.2}$$

where $w_k = 1/\pi_k$ is a probability survey weight for unit $k \in s_P$ and $\pi_k$ is its selection probability. We focus on this basic weight for simplicity although more complex weighting methods, involving nonresponse and calibration adjustments, are often used in real surveys. Taking the derivative of (2.2) with respect to $\boldsymbol{\alpha}$, we obtain the pseudo likelihood estimating function

$$\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{[1 - p_k(\boldsymbol{\alpha})]} - \sum_{k \in s_P} w_k \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{[1 - p_k(\boldsymbol{\alpha})]}. \tag{2.3}$$

It is easy to see that the estimating function (2.3) is $md$ unbiased, conditional on $\pi_k$ and $\mathbf{x}_k$, $k \in U$, i.e., $E_{md}\left[\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha})\right] = \mathbf{0}$, provided that the following assumption holds:

A2) $E_m\left(\delta_k \middle| \pi_k, \mathbf{x}_k\right) = E_m\left(\delta_k \middle| \mathbf{x}_k\right) = p_k(\boldsymbol{\alpha}),\ k \in U.$

The subscript $m$ refers to the participation model and the subscript $d$ refers to the probability sampling design. Conditioning on $\pi_k$, $k \in U$, makes sense when $\pi_k$ is available in both samples since it can be treated as a potential auxiliary variable. Indeed, assumption (A2) is automatically satisfied if $\pi_k$ is included in the vector $\mathbf{x}_k$. In Section 2.1, we will condition on $\pi_k$, $k \in U$. Then, in Section 2.2, we will consider the case where $\pi_k$ is treated as random and inferences are conditional only on $\mathbf{x}_k$, $k \in U$.

## 2.1 Improvement of the CLW estimating function using BLU estimation theory

The second term on the right-hand side of (2.3) is an estimator of the corresponding term in (2.1), $\boldsymbol{\Phi}(\boldsymbol{\alpha}) = \sum_{k \in U} \mathbf{g}_k(\boldsymbol{\alpha}) \big/ [1 - p_k(\boldsymbol{\alpha})]$. It is an inefficient estimator of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ because it uses only probability sample data and ignores relevant non-probability sample auxiliary data. A more efficient estimator would thus use auxiliary data from both samples. Such an estimator could be obtained by applying the Missing Information Principle (MIP). The MIP was introduced by Orchard and Woodbury (1972), see also Chambers (2023) for a recent reference on applications of the MIP with survey data. The MIP consists of replacing the population likelihood estimating function (2.1) with its expectation conditional on observed data, or equivalently replacing $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ with its best predictor. However, this would involve modelling the vector of auxiliary variables, and the MIP solution would generally not be easy to implement.

As an alternative to applying the MIP, we propose to estimate $\mathbf{\Phi}(\boldsymbol{\alpha})$ using BLU estimation theory. We consider the following linear unbiased estimator that uses auxiliary data from both samples:

$$\hat{\mathbf{\Phi}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{\gamma_k}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]} + \sum\nolimits_{k \in s_P} w_k (1 - \gamma_k) \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]}, \tag{2.4}$$

where $\gamma_k, k \in U,$ are constants. It is easy to show that $\hat{\mathbf{\Phi}}(\boldsymbol{\alpha})$ is $md$ unbiased for $\mathbf{\Phi}(\boldsymbol{\alpha}),$ i.e., $E_{md}\left[\hat{\mathbf{\Phi}}(\boldsymbol{\alpha})\right] = \mathbf{\Phi}(\boldsymbol{\alpha}),$ provided that assumption (A2) holds. Replacing the second term on the right-hand side of (2.1) with the right-hand side of (2.4), we obtain the $md$ unbiased estimating function

$$\hat{\mathbf{U}}_{\pi}^{\gamma}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{(1 - \gamma_k)}{\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} w_k \frac{(1 - \gamma_k)}{\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}). \tag{2.5}$$

It is easy to see that the CLW estimating function (2.3) is the special case of (2.5) obtained by specifying $\gamma_k = 0$ for all $k \in U$.

The BLU estimator of $\mathbf{\Phi}(\boldsymbol{\alpha})$ is obtained by finding $\gamma_k, k \in U,$ that minimize $\text{var}_{md}\left[\mathbf{c}'\hat{\mathbf{\Phi}}(\boldsymbol{\alpha})\right]$ for any fixed vector $\mathbf{c} \neq \mathbf{0}$. We make the following assumptions:

A3) $I_k, k \in U,$ are mutually independent given $\pi_k$ and $\mathbf{x}_k, k \in U,$ where $I_k$ is the indicator of inclusion in the probability sample $s_P$.

A4) $E_d\left(I_k \mid \delta_k, \pi_k, \mathbf{x}_k\right) = E_d\left(I_k \mid \pi_k, \mathbf{x}_k\right) = \pi_k, k \in U$.

Assumption (A3) implies that the probability sample is selected using Poisson sampling. It is used to simplify the derivations of $\text{var}_{md}\left[\mathbf{c}'\hat{\mathbf{\Phi}}(\boldsymbol{\alpha})\right]$ even though we recognize that other sampling designs may be used in practice. Note that neither assumption (A3) nor assumptions (A1) and (A4) are needed to prove that the estimating function (2.5) is $md$ unbiased. Under (A1)-(A4), it is straightforward to show that

$$\text{var}_{md}\left[\mathbf{c}'\hat{\mathbf{\Phi}}(\boldsymbol{\alpha})\right] = \sum\nolimits_{k \in U} \frac{\left[1 - p_k(\boldsymbol{\alpha})\right]}{p_k(\boldsymbol{\alpha})} \gamma_k^2 \left(\frac{\mathbf{c}'\mathbf{g}_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right)^2 + \sum\nolimits_{k \in U} \frac{(1 - \pi_k)}{\pi_k} (1 - \gamma_k)^2 \left(\frac{\mathbf{c}'\mathbf{g}_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right)^2. \tag{2.6}$$

The variance (2.6) is minimized when

$$\gamma_k = \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = \frac{(1 - \pi_k) p_k(\boldsymbol{\alpha})}{(1 - \pi_k) p_k(\boldsymbol{\alpha}) + \left[1 - p_k(\boldsymbol{\alpha})\right] \pi_k} = \frac{w_k - 1}{(w_k - 1) + (p_k^{-1}(\boldsymbol{\alpha}) - 1)}, \; k \in U. \tag{2.7}$$

Plugging $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ in (2.4) leads to the BLU estimator of $\mathbf{\Phi}(\boldsymbol{\alpha})$.

We have the following properties associated with $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$:

i) If $\pi_k = p_k(\boldsymbol{\alpha})$ then $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = 1/2$;

ii) If $\pi_k > p_k(\boldsymbol{\alpha})$ then $0 \leq \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) < 1/2$;

iii) If $\pi_k < p_k(\boldsymbol{\alpha})$ then $1/2 < \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) \leq 1$;

iv) If $\pi_k \to 1$ or $p_k(\boldsymbol{\alpha}) \to 0$ then $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) \to 0$; and

v) If $\pi_k \to 0$ or $p_k(\boldsymbol{\alpha}) \to 1$ then $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) \to 1$.

As a result of properties (iii) and (v), when the probability sample is small compared with the non-probability sample, $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ is expected to be large for many population units, and the CLW method ($\gamma_k = 0$) may become inefficient relative to the optimal solution (2.7). The inefficiency of the CLW method in that scenario was shown in the empirical study of Savitsky et al. (2022). The explanation for this inefficiency is that the CLW method ignores the large non-probability sample for the estimation of the population total $\boldsymbol{\Phi}(\boldsymbol{\alpha})$. From properties (ii) and (iv), the CLW method should perform better in the reverse scenario where the probability sample is much larger than the non-probability sample as it possesses the CL property, i.e., the estimating function (2.3) reduces to the population likelihood estimating function (2.1) when the probability sample is a census. This scenario is not unrealistic in practice. For example, the Canadian Long-Form Census, randomly administered to 25% of the Canadian population, could be an effective probability sample for the estimation of participation probabilities for a smaller non-probability sample.

If we plug (2.7) in the estimating function (2.5), we obtain the "optimal" estimating function:

$$\begin{aligned}\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha}) \quad &= \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\pi_k}{\left[\pi_k + p_k(\boldsymbol{\alpha}) - 2\pi_k p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) \\ &- \sum\nolimits_{k \in s_P} \frac{1}{\pi_k} \frac{\pi_k}{\left[\pi_k + p_k(\boldsymbol{\alpha}) - 2\pi_k p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}).\end{aligned} \tag{2.8}$$

It is straightforward to show that $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$ is the BLU predictor of $\mathbf{U}(\boldsymbol{\alpha})$ given in (2.1). Like the CLW estimating function (2.3), it possesses the CL property.

## 2.2 Weight smoothing

As discussed above, the possible inefficiency of (2.3) can be mostly explained by the omission of relevant non-probability sample auxiliary data for the estimation of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$. Another possible source of inefficiency may be attributable to the variability of the survey weights $w_k$, $k \in s_P$. Indeed, it is well known that pseudo likelihood estimation can be inefficient for the estimation of model parameters from probability survey data (e.g., see Chambers, 2023, for a recent reference). Weight smoothing (Beaumont, 2008) can be used to address this issue. In this context, it consists of replacing the survey weight $w_k$ in (2.3) with the smoothed weight $\tilde{w}_k = E_{\xi}\left(w_k \mid k \in s_P, \mathbf{x}_k\right)$, where the subscript $\xi$ indicates that the expectation is taken with respect to a model for $\pi_k$ (or $w_k$). The smoothed weight $\tilde{w}_k$ is often unknown but can be estimated using the probability sample along with parametric or nonparametric models. If $\pi_k$ is available in the non-probability sample and included in the vector $\mathbf{x}_k$, $\tilde{w}_k = w_k$ and weight smoothing does not bring any efficiency gain.

Using a relationship in Pfeffermann and Sverchkov (1999), the smoothed weight can be expressed as

$$\tilde{w}_k = E_{\xi}\left(w_k \mid k \in s_P, \mathbf{x}_k\right) = \frac{1}{E_{\xi}\left(\pi_k \mid \mathbf{x}_k\right)} = \frac{1}{\tilde{\pi}_k}, \tag{2.9}$$

where $\tilde{\pi}_k = E_\xi\left(\pi_k \,|\, \mathbf{x}_k\right) = \Pr\left(k \in s_P \,|\, \mathbf{x}_k\right)$. Using this relationship, it is straightforward to show that the estimating function (2.3) is $\xi md$ unbiased, regardless of the validity of assumption (A2) and whether $w_k$ or $\tilde{w}_k$ is used in (2.3), i.e., $E_{\xi md}\left[\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha})\right] = \mathbf{0}$ and $E_{\xi md}\left[\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})\right] = \mathbf{0}$, where $\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})$ is the estimating function (2.3) with $w_k$ replaced by $\tilde{w}_k$. Note that $\xi md$ expectations are conditional only on $\mathbf{x}_k$, $k \in U$, so that $\pi_k$ is treated as random. Relationship (2.9) can also be used to obtain an estimator of $\tilde{\pi}_k$, i.e., a consistent estimator of $\tilde{\pi}_k = E_\xi\left(\pi_k \,|\, \mathbf{x}_k\right)$ is $\hat{\tilde{\pi}}_k = 1/\hat{\tilde{w}}_k$, where $\hat{\tilde{w}}_k$ is a consistent estimator of $\tilde{w}_k = E_\xi\left(w_k \,|\, k \in s_P, \mathbf{x}_k\right)$.

Using $\tilde{w}_k$ instead of $w_k$ in the estimating function (2.3) increases its efficiency at the expense of requiring the validity of a model for $w_k$ and estimation of $\tilde{w}_k$. A similar argument can be made to improve the efficiency of $\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})$ by replacing $w_k$ by $\tilde{w}_k$ in (2.4). The resulting estimator is $\xi md$ unbiased, i.e., $E_{\xi md}\left[\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right] = \boldsymbol{\Phi}(\boldsymbol{\alpha})$, and its variance $\operatorname{var}_{\xi md}\left[\mathbf{c}'\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right]$ takes the same form as (2.6), with $\pi_k$ replaced by $\tilde{\pi}_k$, provided that assumptions (A1)-(A4) hold as well as the following assumption:

A5) $\pi_k$, $k \in U$, are mutually independent given $\mathbf{x}_k$, $k \in U$.

As a result, the optimal value of $\gamma_k$, denoted by $\gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$, and the optimal estimating function, denoted by $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, are again given by expressions (2.7) and (2.8), respectively, with $\pi_k$ replaced by $\tilde{\pi}_k$.

Using $\pi_k$ in (2.8) is not possible if it is not observed in the non-probability sample, a likely scenario in practice. In that case, an estimate of $\tilde{\pi}_k$ can be used in (2.8) to replace $\pi_k$. If $\pi_k$ is observed in the non-probability sample but not included in $\mathbf{x}_k$, it may still be desirable to replace $\pi_k$ with an estimate of $\tilde{\pi}_k$ to improve the efficiency of the optimal estimating function (2.8).

### 2.3 Variable selection

The estimating function (2.8) is not the derivative of a pseudo log likelihood function. Therefore, the methodology that we used in our main paper to derive an Akaike Information Criterion (AIC), based on Lumley and Scott (2015), is not directly applicable. For variable selection, one option is to use our proposed AIC along with the CLW method. Once auxiliary variables are selected, the estimating function (2.8), $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$, or its smoothed version $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, could be used to estimate $\boldsymbol{\alpha}$ instead of the CLW estimating function. Otherwise, variable selection methods that are not likelihood-based could be envisioned.

As a side remark, the methodology developed in this section to obtain an optimal estimator (or BLU estimator) of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ could also be used to combine two independent probability samples from the same population. This is left for future research.

## 3. The method of Wang, Valliant and Li (2021) and Pseudo-ILR

The method of Wang-Valliant-Li (WVL) consists of creating an artificial population $U_A$ by stacking the non-probability sample $s_{\text{NP}}$ on top of the population $U$. Each element of $U_A$ is considered distinct even though non-probability sample units are present twice in $U_A$. An indicator $R_i$ is defined for each element

$i \in U_A$; $R_i = 1$, if $i \in s_{\text{NP}} \cap U_A$, and $R_i = 0$, if $i \in U \cap U_A$. We use the subscript $i$ to refer to elements of the artificial population $U_A$ so as to distinguish them from the units of the population $U$. For any given unit $k \in s_{\text{NP}}$, there are two distinct elements of $U_A$ that are labelled differently from that unit $k$; $R = 0$ for one element and $R = 1$ for the other. The probability sample is assumed to be selected from the elements in $U \cap U_A$ for which $R_i = 0$. The authors also assumed that the indicators $R_i, i \in U_A$, are mutually independent given $\mathbf{x}_i, i \in U_A$, and obtained a pseudo log likelihood function similar to CLW by modelling $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$ using a logistic model. Then, they established the relationship $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right) = p_i/(1+p_i)$, which allowed them to estimate the participation probability $p_i$. Because they used a logistic model for $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$, they implicitly modelled $p_i$ using the exponential model $p_i(\boldsymbol{\alpha}) = \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)$, which has the undesirable feature of admitting estimates greater than 1. However, nothing in their theory would have prevented them from using another model for $p_i$, such as the logistic model, and thereby implicitly obtaining a model for $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$. This is exactly what Dr. Gershunskaya and Dr. Beresovsky proposed in their discussion. They call their method Pseudo-ILR, which is simply the WVL method with a logistic model for $p_i$.

For an arbitrary parametric model for $p_i$, the WVL or Pseudo-ILR estimating function can be expressed as

$$\hat{\mathbf{U}}_{\pi}^{\text{WVL-PILR}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 + p_k(\boldsymbol{\alpha})\right]} - \sum\nolimits_{k \in s_P} w_k \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 + p_k(\boldsymbol{\alpha})\right]}. \tag{3.1}$$

Similar to the CLW method, the estimating function (3.1) can possibly be improved by replacing the survey weight $w_k$ with the smoothed weight $\tilde{w}_k$. The resulting estimating function is denoted by $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$.

Dr. Gershunskaya and Dr. Beresovsky pointed out that, unlike $\delta_k$, the indicator $R_i$ is fully observed once the probability and non-probability samples are observed. However, this characteristic of $R_i$ is deceiving. The indicators $R_i$ for elements in the probability and non-probability samples do not bring any new information about the participation mechanism than what is observable about $\delta_k$ and used in the CLW method. In other words, both methods use the same observed information: $\left\{w_k, \mathbf{x}_k; k \in s_P\right\}$ and $\left\{\delta_k, \mathbf{x}_k; k \in s_{\text{NP}}\right\}$. In addition, we see two main issues with the WVL method, which are described below.

*Issue* 1: The assumption that $R_i, i \in U_A$, are mutually independent given $\mathbf{x}_i, i \in U_A$, is not valid as each non-probability sample unit is present twice in $U_A$: $R_i = 0$ for one element and $R_i = 1$ for the other (see greater detail below).

*Issue* 2: The estimating function (3.1) does not have the CL property as it does not reduce to the population likelihood estimating function (2.1) when the probability sample is a census. When $s_P = U$, the CLW method uses the same amount of information as if $\left\{\delta_k, \mathbf{x}_k; k \in U\right\}$ were known, but the WVL log likelihood fails to recognize this information.

Despite the above two issues, it is easy to show that the estimating function (3.1) is $md$ unbiased provided that assumption (A2) holds. Both $\hat{\mathbf{U}}_{\pi}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ and $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ are also $\xi md$ unbiased, regardless of the validity of assumption (A2); the estimating function (3.1) is thus valid. It can be written in the form (2.5) with $\gamma_k = \gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha}) = 2\,p_k(\boldsymbol{\alpha}) \big/ \left[1 + p_k(\boldsymbol{\alpha})\right].$ It is easy to see that $0 < \gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha}) \le 1;$ it thus uses non-probability sample auxiliary data to some extent for the estimation of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$.

The variance (2.6) is a quadratic function of $\gamma_k$ that is minimized when $\gamma_k = \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}),\ k \in U.$ Therefore, if $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ is closer to $\gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ than to 0 for most population units, i.e., $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) > 0.5\gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha}),$ the WVL estimating function (3.1) is expected to be more efficient than the CLW estimating function ($\gamma_k = 0$), but still less efficient than the optimal estimating function (2.8). It is easy to show that $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) > 0.5\gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ is satisfied when $\pi_k < \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}.$ However, if $\pi_k > \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$ for most population units, the CLW estimating function (2.3) is expected to become more efficient than (3.1), in particular when the probability sample is a census ($\pi_k = 1, k \in U$). Since $\left[2 - p_k(\boldsymbol{\alpha})\right]^{-1} > 0.5,$ the condition $\pi_k < \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$ is typically more common in social surveys than $\pi_k > \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1},$ at least for most population units. It is also straightforward to show that $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = \gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ when $\pi_k = 1/3$ whereas $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = 0$ when $\pi_k = 1.$ The WVL estimating function should thus be close to the optimal estimating function when the probability sample size is around one third of the population size and the selection probabilities $\pi_k$ are not too variable.

Like Prof. Wu in his discussion, we had trouble understanding the probabilistic framework underlying the relationship $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right) = p_i / (1 + p_i).$ However, it appears that the clever setup proposed by Savitsky et al. (2022) provides a correct justification of this relationship. These authors imagined a fixed augmented population $U^*$ obtained by stacking population $U_1$ on top of population $U_2,$ where $U_1$ and $U_2$ are two populations identical to $U$ of size $N,$ but uniquely labelled so that each of the $2N$ elements of $U^*$ is viewed as distinct. First, one of the two populations is chosen at random with probability ½. We denote that randomly selected population by $U_{\text{NP}},$ which is either $U_1$ or $U_2.$ The other population is denoted by $U_P.$ The non-probability sample $s_{\text{NP}}$ is observed from $U_{\text{NP}},$ and the probability sample $s_P$ is randomly selected from $U_P.$ Using this setup, it is easy to show that

$$\Pr\left(R_i = 1 \middle| i \in s_{\text{NP}} \cup U_P, \mathbf{x}_i\right) = \frac{\Pr\left(i \in s_{\text{NP}} \middle| \mathbf{x}_i\right)}{\Pr\left(i \in s_{\text{NP}} \cup U_P \middle| \mathbf{x}_i\right)} = \frac{1/2\,p_i}{1/2 p_i + 1/2} = \frac{p_i}{(1 + p_i)}.$$

Note that the random splitting of $U^*$ into $U_{\text{NP}}$ and $U_P$ is not explicitly stated in Savitsky et al. (2022) but is necessary to obtain the above equation.

This setup seems to solve the problem but the two issues noted above remain. In particular, it is easy to show that the independence assumption is not satisfied since, for any pair of elements $i \in U_1$ and $j \in U_2,$

$$\Pr\left(R_i = 1, R_j = 1 \middle| i, j \in s_{\text{NP}} \cup U_P, \mathbf{x}_i, \mathbf{x}_j\right) = 0 \ne \frac{p_i}{(1 + p_i)} \frac{p_j}{(1 + p_j)}.$$

For two different elements $i$ and $j$ in the same population, either $U_1$ or $U_2,$ we have

$$\Pr\left(R_i = 1, R_j = 1 \middle| i, j \in s_{\text{NP}} \cup U_P, \mathbf{x}_i, \mathbf{x}_j\right) = \frac{p_i p_j}{1 + p_i p_j} \ne \frac{p_i}{(1 + p_i)} \frac{p_j}{(1 + p_j)}.$$

provided (A1) holds for elements $i \in U_{\mathrm{NP}}$. Therefore, the independence assumption is reasonable only when all (or at least many of) the participation probabilities are small, and thereby the overlap is a small portion of the probability sample. In this situation, the WVL and CWL estimating functions should be roughly equivalent. If most of the participation probabilities are large, the pseudo log likelihood function proposed by WVL is based on an incorrect independence assumption. In principle, an AIC based on an incorrect (pseudo) log likelihood function is not valid. How small should the participation probabilities be to make this independence assumption reasonable? The simulation study of Dr. Gershunskaya and Dr. Beresovsky is a first step in that direction, but further studies are needed. Note that the CLW pseudo log likelihood function is valid regardless of the magnitude of the participation probabilities as long as assumption (A1) holds.

# 4. The method of Elliott (2009) and ILR

In the method of Elliott (2009), see also Elliott and Valliant (2017), a combined sample $s^*$ is obtained by stacking the non-probability sample $s_{\mathrm{NP}}$ on top of the probability sample $s_P$ while ignoring the possible (unknown) overlap. A population unit $k \in U$ that is selected in $s_P$ and observed in $s_{\mathrm{NP}}$ is thus present twice in $s^*$. Elliott (2009) implicitly assumed that the overlap between both samples is negligible. Similar to Wang, Valliant and Li (2021), an indicator $z_i$, $i \in s^*$, is created such that $z_i = 1$, if $i \in s_{\mathrm{NP}} \cap s^*$, and $z_i = 0$, if $i \in s_P \cap s^*$. Elliott (2009) proposed to model $\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right)$ using a logistic model and, assuming the sampling fractions are small (Elliott and Valliant, 2017), established the relationship $p_i = K\tilde{\pi}_i \rho_i / (1-\rho_i)$ used to estimate $p_i$, where $K$ is an unknown constant of proportionality. This implies that $\rho_i = p_i / (K\tilde{\pi}_i + p_i)$. In practice, $\tilde{\pi}_i$ is typically unknown but can be estimated, as discussed in Section 2.

In this section and the next one, we condition on $\mathbf{x}_i$ and treat $\pi_i$ as random. The theory remains valid if we condition on both $\mathbf{x}_i$ and $\pi_i$ provided that $\tilde{\pi}_i$ is replaced with $\pi_i$ in the developments below and assumption (A2) holds. Conditioning on $\pi_i$ makes sense only if it is observed in both samples, so that it can be treated as a potential auxiliary variable and included in $\mathbf{x}_i$. If $\pi_i$ is included in $\mathbf{x}_i$, $\tilde{\pi}_i = \pi_i$ and assumption (A2) is satisfied. For complex probability surveys, it is unlikely that $\pi_i$ would be observed in the non-probability sample. In that case, it must be treated as random.

Using the setup introduced by Savitsky et al. (2022), also described in Section 3, it is easy to show that

$$\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right) = \frac{\Pr\left(i \in s_{\mathrm{NP}} \middle| \mathbf{x}_i\right)}{\Pr\left(i \in s^* \middle| \mathbf{x}_i\right)} = \frac{p_i}{(\tilde{\pi}_i + p_i)}.$$

Note that the relationship does not require a constant of proportionality $(K = 1)$ and is valid regardless of the size of the sampling fractions. When the logistic model $\rho_i(\boldsymbol{\alpha}) = \left[1 + \exp\left(-\mathbf{x}_i'\boldsymbol{\alpha}\right)\right]^{-1}$ is used, it is easy to see that the resulting implicit model for $p_i$ is $p_i(\boldsymbol{\alpha}) = \tilde{\pi}_i \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)$, which admits estimates greater than 1. Other models for $p_i$ can be considered, such as the logistic model. Beresovsky (2019), see also Gershunskaya and Lahiri (2023), called this method ILR, which is essentially the Elliott method with a logistic model for $p_i$ and results in an implicit model for $\rho_i$.

A log likelihood function is derived by assuming that $z_i,\ i \in s^*,$ are mutually independent given $\mathbf{x}_i,$ $i \in s^*.$ For an arbitrary parametric model for $p_i,$ the resulting Elliott/ILR estimating function is

$$\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k}{\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k}{\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}). \tag{4.1}$$

The estimating function (4.1) is $\xi md$ unbiased. If $\tilde{\pi}_k$ is replaced by $\pi_k$ in (4.1), the resulting estimating function is denoted by $\hat{\mathbf{U}}_{\pi}^{E\text{-ILR}}(\boldsymbol{\alpha}).$ It is both $md$ and $\xi md$ unbiased provided that assumption (A2) holds. The estimating function (4.1) has a form similar to the optimal estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ given by (2.8) with $\pi_k$ replaced by $\tilde{\pi}_k.$ Both $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ and $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$ are expected to be roughly equivalent in general, except for scenarios where the sampling fraction in both samples is large and the overlap is not small. It is thus not surprising that the estimating function (4.1) performed better than the CLW and WVL estimating functions in the simulation study of Savitsky et al. (2022).

The two issues noted in Section 3 for the WVL/Pseudo-ILR method also apply for the Elliott/ILR method. The estimating function (4.1) does not have the CL property since it does not reduce to the population likelihood estimating function (2.1) when the probability sample is a census. Indeed, it reduces to the WVL/Pseudo-ILR estimating function (3.1).

Also, the assumption that $z_i,\ i \in s^*,$ are mutually independent given $\mathbf{x}_i,\ i \in s^*,$ is not valid since, using the setup of Savitsky et al. (2022) along with the random splitting of $U^*$ described in Section 3,

$$\Pr\left(z_i = 1, z_j = 1 \middle| i, j \in s^*, \mathbf{x}_i, \mathbf{x}_j\right) = 0 \neq \frac{p_i}{(\tilde{\pi}_i + p_i)} \frac{p_j}{(\tilde{\pi}_j + p_j)},$$

for any pair of elements $i \in U_1$ and $j \in U_2.$ For two different elements $i$ and $j$ in the same population, either $U_1$ or $U_2,$ we also have

$$\Pr\left(z_i = 1, z_j = 1 \middle| i, j \in s^*, \mathbf{x}_i, \mathbf{x}_j\right) = \frac{p_i p_j}{\tilde{\pi}_i \tilde{\pi}_j + p_i p_j} \neq \frac{p_i}{(\tilde{\pi}_i + p_i)} \frac{p_j}{(\tilde{\pi}_j + p_j)},$$

when (A1) holds for elements $i \in U_{\text{NP}}$ as well as (A3) and (A5) for elements $i \in U_P.$ Even under these assumptions, the mutual independence of $z_i,\ i \in s^*,$ is not tenable unless many of the $p_i$'s are small, and thereby the overlap is a negligible portion of the probability sample. This condition appears to be reasonably satisfied in the simulation study of Dr. Gershunskaya and Dr. Beresovsky and may explain the good performance of the AIC for the ILR method. In principle, an AIC based on an incorrect log likelihood function is not valid and may not be effective for variable selection.

# 5. A sample likelihood approach

## 5.1 Known overlap between both samples

A sample likelihood function (e.g., Pfeffermann, Krieger and Rinott, 1998) in the data integration scenario studied in our paper is a likelihood function based on observations from sample units $k \in s_{\text{NP}} \cup s_P.$

Let us first assume that we have access to $\mathbf{X}_s = \{\mathbf{x}_k; k \in s_{\mathrm{NP}} \cup s_P\}$ in addition to $\{\mathbf{x}_k; k \in s_{\mathrm{NP}}\}$. These two data sets do not need to be linked, but the overlap between the probability and non-probability samples needs to be known to create $\mathbf{X}_s$ from auxiliary data of the two samples. We thus assume that either $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ or $\{I_k, \mathbf{x}_k; k \in s_{\mathrm{NP}}\}$ is known. This assumption will be relaxed in Section 5.2.

Under assumptions (A2) and (A4), it is easy to show that the probability of participation given $k \in s_{\mathrm{NP}} \cup s_P$ is

$$p_{s,k} = \Pr\left(\delta_k = 1 \middle| k \in s_{\mathrm{NP}} \cup s_P, \mathbf{x}_k\right) = \frac{\Pr\left(k \in s_{\mathrm{NP}} \middle| \mathbf{x}_k\right)}{\Pr\left(k \in s_{\mathrm{NP}} \cup s_P \middle| \mathbf{x}_k\right)} = \frac{p_k}{p_k + \tilde{\pi}_k - \tilde{\pi}_k p_k}. \tag{5.1}$$

This conditional participation probability reduces to $p_{s,k} \approx p_k / (p_k + \tilde{\pi}_k)$ when the overlap is negligible, which is the implicit assumption made by Elliott (2009). Note that our assumptions (A2) and (A4) do not necessarily imply a negligible overlap, in particular when the sampling fractions are large.

Under the independence assumptions (A1), (A3) and (A5), $(I_k, \delta_k)$, $k \in s_{\mathrm{NP}} \cup s_P$, are mutually independent given $\mathbf{x}_k$, $k \in s_{\mathrm{NP}} \cup s_P$. Assuming a parametric model for $p_k$, the sample likelihood function can be written as

$$L_s(\boldsymbol{\alpha}) = \prod_{k \in s_{\mathrm{NP}} \cup s_P} \left[p_{s,k}(\boldsymbol{\alpha})\right]^{\delta_k} \left[1 - p_{s,k}(\boldsymbol{\alpha})\right]^{(1-\delta_k)}, \tag{5.2}$$

where $p_{s,k}(\boldsymbol{\alpha})$ is given by (5.1) with $p_k = p_k(\boldsymbol{\alpha})$. If Poisson sampling is not used to select the probability sample, assumption (A3) does not hold. It remains to be verified if the sample likelihood function (5.2) would remain approximately valid for sampling designs used in practice beyond Poisson sampling. In the context of modelling probability sample data only, Pfeffermann, Krieger and Rinott (1998) showed the asymptotic independence of sample observations for common sampling designs provided that the population observations are independent. It is possible that a similar result would also hold when combining probability and non-probability sample data.

Using (5.2) and reorganising terms, we obtain the sample log likelihood function

$$l_s(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \log\left[p_{s,k}(\boldsymbol{\alpha})\right] + \sum_{k \in s_P} \log\left[1 - p_{s,k}(\boldsymbol{\alpha})\right] - \sum_{k \in s_{\mathrm{NP}} \cap s_P} \log\left[1 - p_{s,k}(\boldsymbol{\alpha})\right]. \tag{5.3}$$

Taking the derivative of $l_s(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, we obtain, after straightforward algebra, the sample likelihood estimating function

$$\mathbf{U}_s(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) + \boldsymbol{\Psi}(\boldsymbol{\alpha}), \tag{5.4}$$

where

$$\boldsymbol{\Psi}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}} \cup s_P} I_k \delta_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} = \sum_{k \in s_{\mathrm{NP}}} I_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} = \sum_{k \in s_P} \delta_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]}. \tag{5.5}$$

The estimating function (5.4) satisfies the CL property and, under assumptions (A2) and (A4), is $\xi md$ unbiased conditional on $\mathbf{X}_s$. From (5.5), we observe that the use of estimating function (5.4) requires knowing the overlap only in one of the two samples, i.e., observing either $\{I_k, \mathbf{x}_k; k \in s_{\mathrm{NP}}\}$ or $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ is sufficient. This information could be obtained via additional questions in the probability or non-probability survey, or via record linkage, with auxiliary variables being possible matching variables. For instance, if the vector $\mathbf{x}_k$ is distinct for each population unit $k \in U$ (e.g., there is at least one continuous auxiliary variable), it is possible to gain knowledge of $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ and $\{I_k, \mathbf{x}_k; k \in s_{\mathrm{NP}}\}$ by matching each unit of one sample with all the units of the other sample. That is, if the vector $\mathbf{x}_k$ for a unit $k \in s_P$ is identical to the vector $\mathbf{x}_l$ of a unit $l \in s_{\mathrm{NP}}$, we then know that $\delta_k = 1$ (and $I_l = 1$). Otherwise, if there is no match with $\mathbf{x}_k$, then $\delta_k = 0$. This matching can be repeated for each unit $k \in s_P$ to identify the entire overlap $\{\delta_k, \mathbf{x}_k; k \in s_P\}$. A similar procedure can be used to identify $\{I_k, \mathbf{x}_k; k \in s_{\mathrm{NP}}\}$. If sufficient information is available to implement the estimating function (5.4) then the classical AIC can be used for variable selection using the sample log likelihood function (5.3), i.e., $\mathrm{AIC} = -2l_s(\hat{\boldsymbol{\alpha}}_s) + 2q$, where $\hat{\boldsymbol{\alpha}}_s$ is the solution of $\mathbf{U}_s(\boldsymbol{\alpha}) = \mathbf{0}$ and $q$ is the number of model parameters. This is the ideal solution if the overlapping units can be accurately identified. Kim and Kwon (2024) independently proposed an unconditional propensity score model approach that appears to be very similar to our sample likelihood approach.

## 5.2 Unknown overlap between both samples

In practice, we may observe neither $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ nor $\{I_k, \mathbf{x}_k; k \in s_{\mathrm{NP}}\}$. One way to address this issue is by a direct application of the MIP. It consists of replacing the unobserved estimating function (5.4) with its expectation conditional on observed data, $\mathbf{X}_{\mathrm{obs}} = \{\{\mathbf{x}_k; k \in s_{\mathrm{NP}}\}, \{\mathbf{x}_k; k \in s_P\}\} \neq \mathbf{X}_s$. This leads to the estimating function

$$\begin{aligned} E_{\xi md}\left(\mathbf{U}_s(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\mathrm{obs}}\right) &= \sum_{k \in s_{\mathrm{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \\ &\quad - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} + E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\mathrm{obs}}\right). \end{aligned} \tag{5.6}$$

Using the last term on the right-hand side of (5.5), the expectation $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\mathrm{obs}}\right)$ in (5.6), which is the best predictor of $\boldsymbol{\Psi}(\boldsymbol{\alpha})$, can be written as

$$E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\mathrm{obs}}\right) = \sum_{k \in s_P} p_k^{\mathrm{obs}} \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]},$$

where $p_k^{\mathrm{obs}} = E_{\xi md}\left(\delta_k \middle| \mathbf{X}_{\mathrm{obs}}\right)$, $k \in s_P$. Under our assumptions, it can be shown that $p_k^{\mathrm{obs}} = E_{\xi md}\left(\delta_k \middle| n_k^{\mathrm{NP}}\right) = n_k^{\mathrm{NP}} / N_k$, where $n_k^{\mathrm{NP}}$, $k \in s_P$, is the number of units $l \in s_{\mathrm{NP}}$ for which $\mathbf{x}_l = \mathbf{x}_k$, and $N_k$, $k \in s_P$, is the number of units $l \in U$ for which $\mathbf{x}_l = \mathbf{x}_k$. The result follows by noting that $n_k^{\mathrm{NP}}$ obeys a binomial distribution with number of trials $N_k$ and probability $p_k(\boldsymbol{\alpha})$. The application of the MIP in this context requires knowing $N_k$, $k \in s_P$. This information is typically unknown, but if we can assume that the population vectors $\mathbf{x}_k$ are all distinct (i.e., $N_k = 1$, $k \in U$), we can identify the entire overlap, as explained above, and thus $p_k^{\mathrm{obs}} = \delta_k$, $k \in s_P$, and $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\mathrm{obs}}\right) = \boldsymbol{\Psi}(\boldsymbol{\alpha})$. In other less trivial cases, $N_k$ could be modelled using, for example, the Poisson distribution.

As a simpler alternative to modelling $N_k$, for situations where the overlap cannot be identified in any of the samples, we propose to replace $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ in (5.4) by its BLU predictor. This may lead to somewhat reduced efficiency compared with the best predictor $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha})\middle|\mathbf{X}_{\text{obs}}\right)$, but at least it does not depend on unknown values $N_k,\ k \in s_P$, as shown below.

We consider the following linear unbiased predictor of $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ that uses the available auxiliary data from both samples:

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \lambda_k \tilde{\pi}_k \frac{p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1-p_k(\boldsymbol{\alpha})\right]} + \sum_{k \in s_P} (1-\lambda_k) p_k(\boldsymbol{\alpha}) \frac{p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1-p_k(\boldsymbol{\alpha})\right]}, \tag{5.7}$$

where $\lambda_k,\ k \in s_{\text{NP}} \cup s_P$, are constants. The estimator (5.7) is conditionally unbiased in the sense that $E_{\xi md}\left(\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) - \boldsymbol{\Psi}(\boldsymbol{\alpha})\middle|\mathbf{X}_s\right) = \mathbf{0}$, provided that assumptions (A2) and (A4) hold. Replacing $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ in (5.4) by the right-hand side of (5.7), we obtain the estimating function

$$\begin{aligned}\hat{\mathbf{U}}_s^{\lambda}(\boldsymbol{\alpha}) &= \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})}\left(1+\lambda_k \frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right)\frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \\ &\quad - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k}\left(1+\lambda_k \frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right)\frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})}.\end{aligned} \tag{5.8}$$

It is $\xi md$ unbiased conditional on $\mathbf{X}_s$.

The BLU predictor of $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ is obtained by determining $\lambda_k,\ k \in s_{\text{NP}} \cup s_P$, that minimize the prediction variance $\text{var}_{\xi md}\left(\mathbf{c}'\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) - \mathbf{c}'\boldsymbol{\Psi}(\boldsymbol{\alpha})\middle|\mathbf{X}_s\right)$. Under our three independence assumptions, this prediction variance is minimized when $\text{var}_{\xi md}\left(\lambda_k \tilde{\pi}_k \delta_k + (1-\lambda_k)\, p_k(\boldsymbol{\alpha})\, I_k - I_k \delta_k \,\middle|\, k \in s_{\text{NP}} \cup s_P, \mathbf{x}_k\right)$ is minimized for each $k \in s_{\text{NP}} \cup s_P$. The constant $\lambda_k$ is thus determined so that $\lambda_k \tilde{\pi}_k \delta_k + (1-\lambda_k) p_k(\boldsymbol{\alpha}) I_k$ predicts as accurately as possible $I_k \delta_k$, i.e., whether unit $k$ is in the intersection of the two samples or not. Adding assumptions (A2) and (A4), it can be shown, after straightforward algebra, that the value of $\lambda_k$ that minimizes the prediction variance is $\lambda_k = \lambda_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha}) = 1 - \gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha}),\ k \in s_{\text{NP}} \cup s_P$, where $\gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$ is given by (2.7) after replacing $\pi_k$ by $\tilde{\pi}_k$. Using $\lambda_k = \lambda_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$ in (5.8), it turns out that the estimating function (5.8) reduces exactly to the optimal estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, which is given by (2.8) with $\pi_k$ replaced by $\tilde{\pi}_k$. It is thus interesting to see that using BLU estimation theory under a population likelihood approach (see Section 2) is equivalent to using BLU prediction theory under a sample likelihood approach.

If the selection probability $\pi_k$ is observed for all the probability and non-probability sample units, it can and should be considered as a potential auxiliary variable to be included in $\mathbf{x}_k$. If $\pi_k$ is included in $\mathbf{x}_k$, $\tilde{\pi}_k = \pi_k$ and thus using $\pi_k$ or $\tilde{\pi}_k$ in (2.8) does not make any difference. If $\pi_k$ is not included in $\mathbf{x}_k$, because it does not appear to explain $\delta_k$ after conditioning on $\mathbf{x}_k$, then the above theory still remains valid and $\tilde{\pi}_k$ can be used in (2.8) as if $\pi_k$ were unknown. It would also be possible to condition on both $\pi_k$ and $\mathbf{x}_k$, which would result in replacing $\tilde{\pi}_k$ by $\pi_k$ in the above developments. The optimal estimating function (2.8) would be $md$ unbiased conditional on $\mathbf{X}_s$ (and unconditionally) provided that assumption (A2) holds.

### 5.3 Variable selection

The estimating function (2.8) is not the derivative of a sample log likelihood function, and thus the classical AIC is not applicable. Further research is needed on variable selection when $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$ or $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ is used to estimate the participation probabilities. However, if many of the $p_k$'s are small, the overlap is a negligible portion of the probability sample and the sample likelihood estimating function (5.4) becomes approximately equivalent to the estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$. As a result, the sample log likelihood function (5.3), ignoring the negligible intersection term, can be used along with $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ to compute the classical AIC and select relevant auxiliary variables. It appears to be similar to the AIC Dr. Gershunskaya and Dr. Beresovsky used in their simulation study for the ILR method, except for the use of the estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$ given in (4.1). Both $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$ and $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ should be similar when the overlap is negligible. It is reassuring to see that their AIC performed well in their simulation study. We expect the performance to deteriorate as the non-probability sample size increases and the overlap becomes non-negligible.

## 6. A unified estimating function

Let us continue with the realistic scenario where neither $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ nor $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$ is known. In that scenario, we have described several methods in previous sections that led to different estimating functions. Assuming $\tilde{\pi}_k$ is used rather than $\pi_k$, they are all special cases of the general estimating function

$$\hat{\mathbf{U}}_{\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} \tilde{w}_k h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}), \tag{6.1}$$

where $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ is a function that depends on the method. Table 6.1 provides the expression of $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ for the methods described in previous sections.

**Table 6.1**
**Expression of $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ for different methods.**

| Method | Estimating function | $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ |
|---|---|---|
| CLW | $\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})$: (2.3)* | $\left[1 - p_k(\boldsymbol{\alpha})\right]^{-1}$ |
| WVL/Pseudo-ILR | $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$: (3.1)* | $\left[1 + p_k(\boldsymbol{\alpha})\right]^{-1}$ |
| Elliott/ILR | $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$: (4.1) | $\tilde{\pi}_k\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]^{-1}$ |
| BLU estimation/prediction | $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$: (2.8)** | $\tilde{\pi}_k\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha}) - 2\tilde{\pi}_k p_k(\boldsymbol{\alpha})\right]^{-1}$ |

* $w_k$ is replaced with $\tilde{w}_k$ in (2.3) and (3.1).
** $\pi_k$ is replaced with $\tilde{\pi}_k$ in (2.8).

Some authors (e.g., Beaumont, 2020; Chen, Li and Wu, 2020; and Rao, 2021) considered the calibration estimating function

$$\hat{\mathbf{U}}_{\pi}^{\text{cal}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \mathbf{x}_k - \sum\nolimits_{k \in s_P} w_k \mathbf{x}_k.$$

Its smoothed version $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{cal}}(\boldsymbol{\alpha})$, obtained by replacing $w_k$ with $\tilde{w}_k$ in the above equation, is also a special case of (6.1). For example, if a logistic model for $p_k(\boldsymbol{\alpha})$ is used, the estimating function (6.1) reduces to $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{cal}}(\boldsymbol{\alpha})$ when $h[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})] = \{p_k(\boldsymbol{\alpha})[1 - p_k(\boldsymbol{\alpha})]\}^{-1}$. The calibration estimating function does not have the CL property but has an implicit double robustness property when a linear model between the survey variables and auxiliary variables holds. It could also be easily generalized to the scenario where different auxiliary variables are available in different probability samples as long as all the auxiliary variables are observed in the non-probability sample. The calibration estimating function $\hat{\mathbf{U}}_{\pi}^{\text{cal}}(\boldsymbol{\alpha})$ is the special case of (2.5) with $\gamma_k = \gamma_k^{\text{CAL}}(\boldsymbol{\alpha}) = -[1 - p_k(\boldsymbol{\alpha})] / p_k(\boldsymbol{\alpha}) < 0.$ It is thus expected to be inefficient for the estimation of $p_k(\boldsymbol{\alpha})$.

The estimating function (6.1) is $\xi md$ unbiased, both unconditional and conditional on $\mathbf{X}_s$. The probability $\tilde{\pi}_k$ in (6.1) can also be replaced by $\pi_k$ if it is available in the non-probability sample. The estimating function (6.1) remains (conditionally) $\xi md$ unbiased provided that assumption (A2) holds (e.g., if $\pi_k$ is included in $\mathbf{x}_k$). If $\pi_k$, $k \in U$, are treated as fixed, (6.1) is also (conditionally) $md$ unbiased under assumption (A2). A hybrid estimating function that does not require the availability of $\pi_k$ in the non-probability sample is

$$\hat{\mathbf{U}}_{\pi,\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} h[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})] \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} w_k h[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})] \mathbf{g}_k(\boldsymbol{\alpha}). \tag{6.2}$$

It is (conditionally) $\xi md$ unbiased without requiring the validity of a model for $\pi_k$, but may be less efficient than (6.1) due to the variability of the probability survey weights $w_k$.

In practice, whether (6.1) or (6.2) is used, $\tilde{\pi}_k$ is unknown and must be estimated. As pointed out in Section 2, the probability sample can be used to estimate $\tilde{w}_k = E_{\xi}(w_k | k \in s_P, \mathbf{x}_k)$ by $\hat{\tilde{w}}_k$, perhaps using nonparametric methods, such as machine learning methods. Using the relationship (2.9), $\tilde{\pi}_k$ is estimated by $\hat{\tilde{\pi}}_k = 1 / \hat{\tilde{w}}_k$. Note that $\tilde{\pi}_k = E_{\xi}(\pi_k | \mathbf{x}_k)$ cannot be estimated by modelling $E_{\xi}(\pi_k | k \in s_P, \mathbf{x}_k)$ and ignoring the probability sampling design, as is sometimes suggested in the literature (e.g., Elliott, 2009; Elliott and Valliant, 2017). This is because the probability sampling design is (strongly) informative with respect to the distribution of $\pi_k$ given $\mathbf{x}_k$.

Let us now consider the homogeneous group model for which the auxiliary variables partition the population into $G$ groups and $p_k(\boldsymbol{\alpha}) = p_g$ for a unit $k$ in group $g$. The smoothed weight for a unit $k$ in group $g$ is $\tilde{w}_g = E_{\xi}(w_k | k \in s_{P,g})$, where $s_{P,g}$ is the set of probability sample units that fall in group $g$. It can simply be estimated by the average of the weights in group $g$, i.e., $\hat{\tilde{w}}_g = \hat{N}_g / n_g^P$, where $\hat{N}_g = \sum\nolimits_{k \in s_{P,g}} w_k$ and $n_g^P$ is the probability sample size in group $g$. For a unit $k$ in group $g$, $\hat{\tilde{\pi}}_k = \hat{\tilde{\pi}}_g = n_g^P / \hat{N}_g$. Replacing $\tilde{\pi}_k$ by $\hat{\tilde{\pi}}_k$ in (6.1) or (6.2) and solving the estimating equations (either $\hat{\mathbf{U}}_{\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \mathbf{0}$ or $\hat{\mathbf{U}}_{\pi,\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \mathbf{0}$) for any choice of $h[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})]$ yield $\hat{p}_g = n_g^{\text{NP}} / \hat{N}_g$, the estimate of $p_g$, where $n_g^{\text{NP}}$ is the non-probability sample size in group $g$. Instead, if $\tilde{\pi}_k$ is replaced by $\pi_k$ in (6.1) or (6.2) and $h[\pi_k, p_k(\boldsymbol{\alpha})]$ depends on both $\pi_k$

and $p_k(\boldsymbol{\alpha}),$ the resulting estimated participation probability in group $g$ is not $\hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g$ anymore and does not have a closed form.

Note that $\hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g$ can be larger than 1. It is more likely to happen for large non-probability samples and small probability samples. For a general vector $\mathbf{x}_k,$ this may suggest that a solution may exist less frequently with the logistic model ($p_k(\hat{\boldsymbol{\alpha}})$ bounded by 1) than with the exponential model ($p_k(\hat{\boldsymbol{\alpha}})$ unbounded). However, the exponential model may not be accurate for large non-probability samples.

For the homogeneous group model, solving the sample likelihood estimating equation $\mathbf{U}_s(\boldsymbol{\alpha}) = \mathbf{0},$ where $\mathbf{U}_s(\boldsymbol{\alpha})$ is given in (5.4), yields

$$\hat{p}_g^{\mathrm{SL}} = \frac{\hat{p}_g}{\hat{p}_g + \left(1 - \frac{n_g^I}{n_g^P}\right)}$$

as the estimate of $p_g,$ where $n_g^I$ is the number of units in the intersection $s_{\mathrm{NP}} \cap s_P$ that fall in group $g$ and $\hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g.$ As expected, $\hat{p}_g^{\mathrm{SL}}$ is close to $\hat{p}_g$ when $\hat{p}_g$ and the overlap rate $n_g^I \big/ n_g^P$ are small. The sample likelihood estimate $\hat{p}_g^{\mathrm{SL}}$ cannot be greater than 1 unlike $\hat{p}_g.$ This is a desirable property of the sample likelihood estimating function (5.4), which results from exploiting information on the overlap between both samples.

# 7. Concluding remarks

In previous sections, we described three likelihood approaches for the estimation of participation probabilities and selection of relevant auxiliary variables that are valid regardless of the size of the probability and non-probability samples as well as the size of the overlap between both samples: the population and pseudo likelihood approaches, described in Section 2, and the sample likelihood approach, described in Section 5. If the probability sample is a census, the population likelihood approach is the most efficient and should be the preferred choice. If the probability sample is not a census, but the overlap is known in at least one of the samples, the sample likelihood approach should be preferred over the pseudo likelihood approach for efficiency considerations. If the overlap is unknown, the pseudo likelihood approach of Chen, Li and Wu (2020) can be used both for the estimation of participation probabilities and computation of an AIC for variable selection. However, the CLW estimating function (2.3) may not be efficient, especially when the non-probability sample is larger than the probability sample, because it does not fully leverage the available auxiliary data. Our optimal estimating function (2.8), $\hat{\mathbf{U}}_\pi^{\mathrm{opt}}(\boldsymbol{\alpha}),$ or its smoothed version $\hat{\mathbf{U}}_{\tilde{\pi}}^{\mathrm{opt}}(\boldsymbol{\alpha}),$ is expected to be more efficient than existing alternatives, although it remains to be demonstrated in an empirical study. Variable selection in the case of unknown overlap requires further research when $\hat{\mathbf{U}}_\pi^{\mathrm{opt}}(\boldsymbol{\alpha})$ or $\hat{\mathbf{U}}_{\tilde{\pi}}^{\mathrm{opt}}(\boldsymbol{\alpha})$ is used, except for the case where many of the participation probabilities are small and the overlap can be neglected. In that case, the sample log likelihood function (5.3), ignoring the overlap term, along with $\hat{\mathbf{U}}_{\tilde{\pi}}^{\mathrm{opt}}(\boldsymbol{\alpha})$ can be used to compute the classical AIC.

In practice, estimated participation probabilities from a parametric model are rarely used directly to compute estimates of finite population parameters. Groups homogeneous with respect to these estimated probabilities are often created to protect against model misspecifications and extreme inverse probability weights. It is possible that the choice of an estimating function does not have a major impact on the estimates of finite population parameters if homogeneous groups are used before computing those estimates. Nevertheless, it seems reasonable to choose the most efficient estimating function for the estimation of participation probabilities before creating homogeneous groups.

Nonparametric estimation of participation probabilities using, for example, machine learning methods could be useful to protect against possible model misspecifications. Existing machine learning methods can be directly used to model $p_k$ if $\{\delta_k, \mathbf{x}_k; k \in U\}$ is known. Otherwise, if $\{\delta_k, \mathbf{x}_k; k \in s_{\mathrm{NP}} \cup s_P\}$ is known, the conditional probability $p_{s,k}$ can be modelled using existing machine learning methods and $p_k$ can then be estimated using relationship (5.1). The most difficult case is when the overlap is unknown. In our main paper, we proposed nppCART as a means of creating homogeneous groups and obtaining protection against model misspecifications. Our procedure is inspired from the pseudo likelihood approach of Chen, Li and Wu (2020) and does not require a negligible overlap between both samples. An alternative would be to consider machine learning methods along with the Elliott/ILR method, as suggested in Elliott and Valliant (2017) and Elliott (2022). The idea would consist of modelling $\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right)$ using a machine learning method, ignoring the overlap and thus the lack of independence between observations. Then, $p_i$ would be estimated for the non-probability sample units using the relationship $\rho_i = p_i / (\tilde{\pi}_i + p_i),$ whose validity was shown in Section 4 using the setup of Savitsky et al. (2022). If most of the participation probabilities are small, the overlap is a negligible portion of the probability sample and can thus be ignored. Therefore, this approach would be equivalent to our suggestion above of modelling $p_{s,k}$ using a machine learning method and then using relationship (5.1) to estimate $p_k$. It remains to be evaluated how that machine learning version of the Elliott/ILR method would perform when the overlap is not negligible.

In our main paper and in this rejoinder, we have focussed on the estimation of the participation probability $p_k$ for non-probability sample units. Once the estimates $\hat{p}_k,\ k \in s_{\mathrm{NP}},$ are computed, they can be used to estimate finite population parameters, such as population totals or means. The basic inverse probability weighted estimator of finite population parameters consists of weighting non-probability sample units by $1/\hat{p}_k$. Of course, there are estimators that more efficiently use $\hat{p}_k,$ for instance, by taking advantage of a model for the survey variables to achieve a double robustness property (e.g., Chen, Li and Wu, 2020; Chambers, Ranjbar, Salvati and Pacini, 2022). The simplest, but common, example is when the inverse probability weights $1/\hat{p}_k$ are calibrated on known or estimated population totals of auxiliary variables. The resulting estimator of population totals is doubly robust in the sense that it is valid under either the participation model or a linear model between survey variables and auxiliary variables.

Our point of view is that survey statisticians should start with the most efficient estimates of $p_k,\ k \in s_{\mathrm{NP}},$ possible before using them for the estimation of finite population parameters. This is exactly the same point of view many survey statisticians take for the estimation of finite population parameters using data from a probability sample; they start with their best estimate of the probability of selection in the sample, $\pi_k$, and then use it to derive efficient estimators of finite population parameters (e.g., using calibration techniques). It just happens that the probability $\pi_k$ is usually known for probability samples and does not require to be estimated.

In this final remark, we would like to take this opportunity to sincerely thank Prof. Partha Lahiri, the guest editor of this special issue, for all his efforts in organizing such a nice collection of papers, which were presented at the 2022 Morris Hansen lecture event, along with their discussion.

# References

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

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Beresovsky, V. (2019). On application of a response propensity model to estimation from web samples. In ResearchGate.

Chambers, R.L. (2023). The missing information principle – A paradigm for analysis of messy sample survey data. *Survey Methodology*, 49, 2, 219-256. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2023002/article/00018-eng.pdf.

Chambers, R., Ranjbar, S., Salvati, N. and Pacini, B. (2022). Weighting, informativeness and causal inference, with an application to rainfall enhancement. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 185, 1584-1612.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Elliott, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813-845.

Elliott, M.R. (2022). Comments on "Statistical inference with non-probability survey samples". *Survey Methodology*, 48, 2, 319-329. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00004-eng.pdf.

Elliott, M., and Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Gershunskaya, J., and Lahiri, P. (2023). Discussion of "Probability vs. nonprobability sampling: From the birth of survey sampling to the present day", by Graham Kalton. *Statistics in Transition New Series*, 24, 3, 31-37.

Kim, J.K., and Kwon, Y. (2024). Comment on "Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples". *Survey Methodology*, 50, 1, 57-63. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00007-eng.pdf.

Lumley, T., and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3, 1-18.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: Theory and application. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.

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

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 61, 166-186.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. and Johnson, N.G. (2022). Methods for combining probability and nonprobability samples under unknown overlaps. https://doi.org/10.48550/arXiv.2208.14541.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.