Electronic Journal of Statistics
Vol. 18 (2024) 4198–4234
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2295

Probability-weighted clustered coefficient regression models in complex survey sampling*

Mingjun Gang[1], Xin Wang[2], Zhonglei Wang[3], and Wei Zhong[3]

[1]*Department of Statistics, National University of Singapore,*
e-mail: gangmingjun@u.nus.edu

[2]*Department of Mathematics and Statistics, San Diego State University,*
e-mail: xwang14@sdsu.edu

[3]*MOE Key Lab of Econometrics, WISE, Department of Statistics and Data Science and School of Economics, Xiamen University,*
e-mail: wangzl@xmu.edu.cn; wzhong@xmu.edu.cn

Abstract: Regression analysis is commonly conducted in survey sampling. However, existing methods fail when regression models vary across different clusters of domains. In this paper, we propose a unified framework to study the cluster-wise covariate effect under complex survey sampling based on pairwise penalties, and the associated objective function is solved by the alternating direction method of multipliers. Theoretical properties of the proposed method are investigated under regularity conditions. Numerical experiments demonstrate that the proposed method outperforms its alternatives in terms of identifying the cluster structure and estimation efficiency for both linear regression and logistic regression models. American Community Survey is used as an example to illustrate the advantages of the proposed approach.

MSC2020 subject classifications: Primary 62D05, 62J07; secondary 62J12, 62J05.
Keywords and phrases: Clustered coefficient, generalized linear regression models, pairwise penalties, survey sampling.

Received November 2023.

1. Introduction

Regression models are commonly used in survey sampling to analyze the relationship among different variables [13, 26, 36]. Traditional regression models assume common regression coefficients across all domains, but this assumption fails and leads to biased estimators if regression coefficients vary across different clusters of domains. In this paper, we propose a probability-weighted clustered coefficients (PCC) regression model to solve this problem under complex survey sampling.

arXiv: 2210.09339
*All authors contributed equally to this work.

A concave fusion approach was proposed to identify a cluster structure and estimate model parameters with common regression coefficients but cluster-specific intercepts [28]. Extension to models with clustered regression coefficient heterogeneity was investigated by [29]. Estimation accuracy of [28] can be further improved by incorporating spatial information for areal data with repeated measures [49]. Various models with cluster-specific regression coefficients were explored for different types of data, such as binary data [55], count data [6], survival data [15], functional data [46, 52], Poisson process [47]; also see [25, 54]. These works have shown the advantages of models with cluster-specific regression coefficients when estimating population parameters. However, none considers identifying a cluster structure of regression coefficients under complex survey sampling. Under complex survey sampling, an empirical-likelihood based method was proposed for variable selection under high dimensional setups [53], but it cannot be used to identify cluster structures; also see [8, 45].

Hierarchical models are widely used in survey sampling and small area estimation [36]. A multi-level Bayesian model was proposed by [21] for small area estimation. A robust data-driven transformation technique was proposed by [38], and a two-level linear regression model with random intercepts was used to obtain a pseudopopulation [31]. [48] proposed new model-based estimators using a linear mixed model with cluster-specific regression coefficients, where random effects were considered in the model compared to [29] without random effects. [23] proposed a model based on linear mixed models with heterogeneity in regression coefficients and variance components. Also see [3, 7, 10, 16, 19, 41]. Generalized linear models with random effects were also proposed for categorical responses [4, 18, 42, 51]. However, existing works did not incorporate sampling weights. A two-level model with random cluster effects was proposed for population mean estimation [34], and sampling weights are incorporated; also see [30]. Under complex survey sampling, a generalized linear regression model with random effect was considered by [9]. The efficiency of existing works can be further improved if we can successfully identify cluster structures, but none pursued in this direction.

Under complex survey sampling, sampling weights are essential to guarantee unbiased inference [33], and statistical efficiency can be improved by incorporating cluster structures. In this work, we consider a PCC regression model with the smoothly clipped absolute deviation (SCAD) penalty [11] to identify the cluster structure by the estimated regression coefficients. We use the alternating direction method of multipliers (ADMM) [5] to implement the proposed method. In theory, we show that the oracle estimator, when the cluster structure is available, is a local minimizer of the objective function with probability approaching one under regularity conditions. The asymptotic distribution of our estimator is established accordingly. Theoretical properties are investigated under a general model setup, so they apply to a wide range of situations, including linear and logistic regression models.

The article is organized as follows. In Section 2, we propose the PCC regression model incorporating sampling weights and introduce an ADMM-based algorithm to solve the objective function. In Section 3, asymptotic properties of

the proposed estimators are investigated under regularity conditions. Two simulation studies are conducted to illustrate the advantage of the proposed method for both linear and logistic regression models in Section 4. The proposed method is applied to a real survey dataset in Section 5. Finally, a summary and conclusion are provided in Section 6.

2. Methodology and algorithm

2.1. Basic setup

In this paper, we consider a finite population $\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_m$ of size N, where $\{\mathcal{F}_1, \ldots, \mathcal{F}_m\}$ are m mutually exclusive domains, $\mathcal{F}_i = \{(y_{ih}, \boldsymbol{x}_{ih}, \boldsymbol{z}_{ih}) : h = 1, \ldots, N_i\}$ for $i = 1, \ldots, m$, y_{ih} is the response of interest for hth element in the ith domain, $\boldsymbol{x}_{ih}$ is a p-dimensional covariate vector associated with the domain-specific part in the regression model, $\boldsymbol{z}_{ih}$ is a q-dimensional covariate vector with respect to the population-level part, N_i is the size of $\mathcal{F}_i$, and $N = \sum_{i=1}^m N_i$. In this paper, we consider the following generalized linear regression model,

$$g\{E(y_{ih} \mid \boldsymbol{x}_{ih}, \boldsymbol{z}_{ih})\} = \boldsymbol{x}_{ih}^{\mathrm{T}} \boldsymbol{\beta}_i^0 + \boldsymbol{z}_{ih}^{\mathrm{T}} \boldsymbol{\eta}^0 \quad (i = 1, \ldots, m; h = 1, \ldots, N_i), \tag{1}$$

where $g(\cdot)$ is a known link function, $\{\boldsymbol{\beta}_i^0 : i = 1, \ldots, m\}$ are domain-specific regression coefficients, and $\boldsymbol{\eta}^0$ is common for different domains. For example, $g(x) = x$ corresponds to a linear regression model, and $g(x) = \log\{x/(1-x)\}$ to a logistic regression model.

In (1), $\{\boldsymbol{\beta}_i^0 : i = 1, \ldots, m\}$ may not be distinct across different domains, and clustering elements with the same domain-specific regression coefficient can effectively improve estimation efficiency. Among $\{\boldsymbol{\beta}_i^0 : i = 1, \ldots, m\}$, assume there are K mutually different cluster-specific regression coefficients $\{\boldsymbol{\alpha}_k : k = 1, \ldots, K\}$, and denote $\mathcal{G}_k = \{i : \boldsymbol{\beta}_i = \boldsymbol{\alpha}_k, 1 \le i \le m\}$ to the domain index set associated with $\boldsymbol{\alpha}_k$. In practice, we neither know $\mathcal{G}_k$ nor K in advance, and we are interested in identifying the partition $\hat{\mathcal{G}}$ and the number of clusters $\hat{K}$, where $\hat{\mathcal{G}} = \{\hat{\mathcal{G}}_1, \ldots, \hat{\mathcal{G}}_{\hat{K}}\}$, $\hat{\mathcal{G}}_k = \{i : \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_k, 1 \le i \le m\}$, and $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\alpha}}_k$ are estimated regression coefficients; see Section 2.2 for details.

If the finite population $\mathcal{F}$ were available, the loss function would be

$$L_N(\boldsymbol{\eta}, \boldsymbol{\beta}) = \sum_{i=1}^m W_i L_i(\boldsymbol{\eta}, \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_m^{\mathrm{T}}\right)^{\mathrm{T}}$, $W_i = N_i/N$, $L_i(\boldsymbol{\eta}, \boldsymbol{\beta}) = N_i^{-1} \sum_{h=1}^{N_i} L_{ih}(\boldsymbol{\eta}, \boldsymbol{\beta})$, and $L_{ih}(\boldsymbol{\eta}, \boldsymbol{\beta})$ is a loss function corresponding to the hth element in the ith domain. For example, $L_{ih}(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2}(y_{ih} - \boldsymbol{z}_{ih}^{\mathrm{T}}\boldsymbol{\eta} - \boldsymbol{x}_{ih}^{\mathrm{T}}\boldsymbol{\beta}_i)^2$ corresponds to a linear regression model, and $L_{ih}(\boldsymbol{\eta}, \boldsymbol{\beta}) = -y_{ih}\left(\boldsymbol{z}_{ih}^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_{ih}^{\mathrm{T}}\boldsymbol{\beta}_i\right) + \log\left\{\exp(\boldsymbol{z}_{ih}^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_{ih}^{\mathrm{T}}\boldsymbol{\beta}_i) + 1\right\}$ to a logistic regression model.

However, $\mathcal{F}$ is never fully observable in practice due to time and budget constraints. Instead, we can only observe a probability sample. Let $n_0 = \sum_{i=1}^m n_i$ be the sample size, where n_i is the size with respect to the ith domain. Denote

δ_{ih} as a sampling indicator with $\delta_{ih} = 1$ if the hth element in the ith domain is observed and $\delta_{ih} = 0$ otherwise, and let π_{ih} be the associated inclusion probability. Then, a probability-weighted loss function is $L_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) = \sum_{i=1}^{m} W_i \hat{L}_i(\boldsymbol{\eta}, \boldsymbol{\beta})$, where $\hat{L}_i(\boldsymbol{\eta}, \boldsymbol{\beta}) = N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-1} L_{ih}(\boldsymbol{\eta}, \boldsymbol{\beta})$. It can be shown that $L_\omega(\boldsymbol{\eta}, \boldsymbol{\beta})$ is design-unbiased for $L_N(\boldsymbol{\eta}, \boldsymbol{\beta})$ [14]. To estimate the domain-specific regression coefficients $\{\boldsymbol{\beta}_i : i = 1, \ldots, m\}$ and identify the cluster structure $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$, we consider the following objective function,

$$Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) = mL_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) + \sum_{1 \le i < j \le m} p_\gamma \left(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda \right), \tag{2}$$

where $p_\gamma(\cdot, \lambda)$ is a penalty function imposed on all distinct pairs of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ with $i \neq j$, γ is a fixed constant, and $\lambda \ge 0$ is a tuning parameter. In this paper, we use the SCAD penalty with the following form:

$$p_\gamma(t, \lambda) = \lambda \int_0^{|t|} \min \left\{ 1, (\gamma - x/\lambda)_+ / (\gamma - 1) \right\} \mathrm{d}x, \tag{3}$$

and the tuning parameter is determined by a BIC criterion; see Section 2.2 for details. Following a practical rule [28, 29, 49], we fix γ to be 3. There are also other options for the penalty function, such as the minimax concave penalty (MCP) [50]. [28] compared different penalty functions and concluded that SCAD and MCP perform similarly, and they are better than an L_1 penalty.

2.2. Algorithm

In this section, we use an ADMM-based algorithm to minimize the objective function (2). First, we fix the tuning parameter λ and show the algorithm to minimize (2). Details regarding the selection of λ are relegated to the end of this section.

Let $\boldsymbol{\zeta}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j$ be the slack parameter associated with $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$. Then, the optimization problem is equivalent to minimizing

$$\begin{aligned} Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\zeta}) &= mL_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) + \sum_{1 \le i < j \le m} p_\gamma \left(\|\boldsymbol{\zeta}_{ij}\|, \lambda \right), \\ \text{subject to} \quad & \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\zeta}_{ij} = \mathbf{0}, \quad 1 \le i < j \le m, \end{aligned} \tag{4}$$

where $\boldsymbol{\zeta} = \left(\boldsymbol{\zeta}_{ij}^{\mathrm{T}}, 1 \le i < j \le m \right)^{\mathrm{T}}$. The constrained objective function (4) can be further formatted as below based on augmented Lagrangian:

$$Q(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{v}) = Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\zeta}) + \sum_{i<j} \langle \boldsymbol{v}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\zeta}_{ij} \rangle + \frac{\vartheta}{2} \sum_{i<j} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\zeta}_{ij}\|^2, \tag{5}$$

where $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ is the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\boldsymbol{v} = \left(\boldsymbol{v}_{ij}^{\mathrm{T}}, 1 \le i < j \le m \right)^{\mathrm{T}}$ are Lagrange multipliers, and $\vartheta > 0$ is a tuning parameter for the penalty. In

this paper, we fix $\vartheta = 1$ following [29]. Then, we use an iterative algorithm to update $\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\zeta}$ and $\boldsymbol{v}$.

Denote $(\boldsymbol{\beta}^{(r)}, \boldsymbol{\eta}^{(r)}, \boldsymbol{\zeta}^{(r)}, \boldsymbol{v}^{(r)})$ as the values at the rth iteration, and they can be updated as

$$\left(\boldsymbol{\eta}^{(r+1)}, \boldsymbol{\beta}^{(r+1)}\right) = \underset{\boldsymbol{\eta},\boldsymbol{\beta}}{\arg\min}\, Q\left(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\zeta}^{(r)}, \boldsymbol{v}^{(r)}\right), \tag{6}$$

$$\boldsymbol{\zeta}^{(r+1)} = \underset{\boldsymbol{\zeta}}{\arg\min}\, Q\left(\boldsymbol{\eta}^{(r+1)}, \boldsymbol{\beta}^{(r+1)}, \boldsymbol{\zeta}, \boldsymbol{v}^{(r)}\right), \tag{7}$$

$$\boldsymbol{v}_{ij}^{(r+1)} = \boldsymbol{v}_{ij}^{(r)} + \vartheta\left(\boldsymbol{\beta}_i^{(r+1)} - \boldsymbol{\beta}_j^{(r+1)} - \boldsymbol{\zeta}_{ij}^{(r+1)}\right). \tag{8}$$

For a classical linear regression model, we can derive closed forms for $\boldsymbol{\eta}^{(r+1)}$ and $\boldsymbol{\beta}^{(r+1)}$ by adding sampling weights to the algorithm in [49]; see Appendix A.1 for details. When a logistic regression model is considered, we derive a new algorithm based on the Newton-Raphson method based on [55]; see Appendix A.2 for details.

To update $\boldsymbol{\zeta}_{ij}$, it is equivalent to minimizing

$$\frac{\vartheta}{2}\left\|\boldsymbol{\kappa}_{ij}^{(r)} - \boldsymbol{\zeta}_{ij}\right\|^2 + p_\gamma\left(\|\boldsymbol{\zeta}_{ij}\|, \lambda\right), \tag{9}$$

where $\boldsymbol{\kappa}_{ij}^{(r+1)} = (\boldsymbol{\beta}_i^{(r+1)} - \boldsymbol{\beta}_j^{(r+1)}) + \vartheta^{-1}\boldsymbol{v}_{ij}^{(r)}$. For the SCAD penalty, we have

$$\boldsymbol{\zeta}_{ij}^{(r+1)} = \begin{cases} S(\boldsymbol{\kappa}_{ij}^{(r+1)}, \lambda/\vartheta) & \text{if } \left\|\boldsymbol{\kappa}_{ij}^{(r+1)}\right\| \le \lambda + \lambda/\vartheta, \\ \frac{S\left(\kappa_{ij}^{(r+1)}, \gamma\lambda/(\gamma-1)\vartheta\right)}{1-1/(\gamma-1)\vartheta} & \text{if } \lambda + \lambda/\vartheta < \left\|\boldsymbol{\kappa}_{ij}^{(r+1)}\right\| \le \gamma\lambda, \\ \boldsymbol{\kappa}_{ij}^{(r+1)} & \text{if } \left\|\boldsymbol{\kappa}_{ij}^{(r+1)}\right\| > \gamma\lambda, \end{cases} \tag{10}$$

where $\gamma > 1 + 1/\vartheta$, and $S(\boldsymbol{w}, t) = (1 - t/\|\boldsymbol{w}\|)_+\boldsymbol{w}$ with $(t)_+ = t$ if $t > 0$ and 0 otherwise.

Algorithm 1 summarizes the estimation procedure.

Algorithm 1 The ADMM algorithm.

Require: : Initialize $\boldsymbol{\beta}^{(0)}, \boldsymbol{\zeta}^{(0)}$ and $\boldsymbol{v}^{(0)}$
1: **for** $r = 0, 1, 2, \ldots$ **do**
2: Obtain $\boldsymbol{\beta}^{(r+1)}$ and $\boldsymbol{\eta}^{(r+1)}$ by solving (6).
3: Obtain $\boldsymbol{\zeta}^{(r+1)}$ by (10).
4: Obtain $\boldsymbol{v}^{(r+1)}$ by (8).
5: **if** convergence criterion is met **then**
6: Stop and get the estimates
7: **else**
8: $r = r + 1$
9: **end if**
10: **end for**

To find reasonable initial values, we set $p_\gamma(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|; \lambda) = \lambda\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$ with a small value for λ. Then, let the solution to (2) be the initial values for $\boldsymbol{\beta}^{(0)}$

and $\boldsymbol{\zeta}^{(0)}$. For a classical linear regression model, the problem becomes a type of ridge regression [29]. For a logistic regression model, estimates can be obtained by [55]. Then, set $\boldsymbol{\zeta}_{ij}^{(0)} = \boldsymbol{\beta}_i^{(0)} - \boldsymbol{\beta}_j^{(0)}$ and $\boldsymbol{v}^{(0)} = \mathbf{0}$.

Denote the solution as

$$(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\eta}\in\mathbb{R}^q, \boldsymbol{\beta}\in\mathbb{R}^{mp}}{\arg\min}\ Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}), \tag{11}$$

and $\hat{\boldsymbol{\zeta}}_{ij} = \hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j$. If $\hat{\boldsymbol{\zeta}}_{ij} = \mathbf{0}$, we conclude that the ith and jth domains are in the same cluster. Then, the corresponding estimated partition $\hat{\mathcal{G}}$ and the estimated number of clusters $\hat{K}$ are identified automatically.

Remark 1. *The convergence criterion used is the same as that in [28] and [49] based on the primal residual* $\boldsymbol{r}^{(r+1)} = \boldsymbol{A}\boldsymbol{\beta}^{(r+1)} - \boldsymbol{\zeta}^{(r+1)}$. *We stop the algorithm if* $\left\|\boldsymbol{r}^{(r+1)}\right\| < \varepsilon$, *where* ϵ *is a small positive number.*

The tuning parameter λ is selected based on modified BIC [44], which is widely used in clustered coefficient regression models [28, 49]. Based on the weighted loss function discussed in [27, 43], the proposed PCC method selects the tuning parameter λ by minimizing

$$BIC(\lambda) = \hat{l}(\hat{\boldsymbol{\eta}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda) + C_m \frac{\log n_0}{n_0} \left\{\hat{K}(\lambda)p\right\}, \tag{12}$$

where $(\hat{\boldsymbol{\eta}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda)$ are the optimal estimators given λ, and C_m is a positive number depending on the number of domains m. In our paper, we set $C_m = \log(mp+q)$ as in [29]. $\hat{l}(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}) = \log(L_\omega(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}))$ for linear regression models and $\hat{l}(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}) = 2L_\omega(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}})$ for logistic regression models.

3. Theoretical properties

Recall that $\mathcal{G}_k$ is the domain index set for the kth cluster with cardinality $|\mathcal{G}_k|$. Denote $\tilde{n}_k = \sum_{i\in\mathcal{G}_k} n_i$ as the number of observations in $\mathcal{G}_k$, and $n_0 = \sum_{i=1}^m n_i$ as the sample size. Let $\widetilde{\boldsymbol{W}} = (w_{ik})$ be an $m \times K$ matrix, where $w_{ik} = 1$ if $i \in \mathcal{G}_k$ and 0 otherwise. Define an $mp \times Kp$ matrix $\boldsymbol{W} = \widetilde{\boldsymbol{W}} \otimes \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, and $\otimes$ is the Kronecker product. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_K^T)^T$ be the cluster-specific regression parameters. Then, we have $\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{\alpha}$ if the true cluster information is available.

For simplicity, denote $L_{ih}^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = L_{ih}(\boldsymbol{\eta}, \boldsymbol{W}\boldsymbol{\alpha})$, $L_i^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = L_i(\boldsymbol{\eta}, \boldsymbol{W}\boldsymbol{\alpha})$, $\hat{L}_i^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \hat{L}_i(\boldsymbol{\eta}, \boldsymbol{W}\boldsymbol{\alpha})$, and $L_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = L_\omega(\boldsymbol{\eta}, \boldsymbol{W}\boldsymbol{\alpha})$. Let $(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0)$ be the true parameters, and its oracle estimator, the one when the true cluster information is known, is obtained by

$$\begin{aligned}(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}) &= \underset{(\boldsymbol{\eta},\boldsymbol{\alpha})\in\Theta}{\arg\min}\, L_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) \\ &= \underset{(\boldsymbol{\eta},\boldsymbol{\alpha})\in\Theta}{\arg\min}\, L_\omega(\boldsymbol{\eta}, \boldsymbol{W}\boldsymbol{\alpha}) = \underset{(\boldsymbol{\eta},\boldsymbol{\alpha})\in\Theta}{\arg\min} \sum_{i=1}^m W_i N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih}\pi_{ih}^{-1} L_{ih}(\boldsymbol{\eta}, \boldsymbol{W}\boldsymbol{\alpha}),\end{aligned} \tag{13}$$

and $\hat{\boldsymbol{\beta}}^{or} = \boldsymbol{W}\hat{\boldsymbol{\alpha}}^{or}$, where Θ is the parameter space.

More notations are needed to discuss the theoretical properties of the proposed method. Let $\boldsymbol{S}_N(\boldsymbol{\eta},\boldsymbol{\alpha}) = \sum_{i=1}^{m} W_i N_i^{-1} \sum_{h=1}^{N_i} \boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})$ be the score function, where $\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha}) = \partial L_{ih}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) / \partial(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ is with respect to the hth element in the ith domain. Let $\hat{\boldsymbol{S}}_w(\boldsymbol{\eta},\boldsymbol{\alpha}) = \sum_{i=1}^{m} W_i \hat{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\alpha})$ be the sample-level score function, where $\hat{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\alpha}) = N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih}\pi_{ih}^{-1} \boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})$. It can be shown that $\hat{\boldsymbol{S}}_w(\boldsymbol{\eta},\boldsymbol{\alpha})$ is design-unbiased for $\boldsymbol{S}_N(\boldsymbol{\eta},\boldsymbol{\alpha})$. Given the cluster information, the oracle estimator $(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or})$ solves

$$\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) = \partial L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) / \partial(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}} = \sum_{i=1}^{m} W_i \hat{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\alpha}) = \boldsymbol{0}.$$

Denote $\boldsymbol{I}_i(\boldsymbol{\eta},\boldsymbol{\alpha}) = N_i^{-1} \sum_{h=1}^{N_i} \boldsymbol{I}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})$ to be the information matrix for the ith domain, where $\boldsymbol{I}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha}) = -\partial \boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha}) / \partial(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}})$. It can be shown that $\hat{\boldsymbol{I}}_i(\boldsymbol{\eta},\boldsymbol{\alpha}) = N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih}\pi_{ih}^{-1} \boldsymbol{I}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})$ is design-unbiased for $\boldsymbol{I}_i(\boldsymbol{\eta},\boldsymbol{\alpha})$.

The following conditions are required to derive the asymptotic properties of the proposed PCC method.

(C1) The number of clusters K, and the dimensions of covariates p and q are fixed. There exists a positive constant C_1 such that

$$\begin{aligned}
\max\{N_i : i = 1,\ldots,m\} / \min\{N_i : i = 1,\ldots,m\} < C_1,\\
\max\{n_i : i = 1,\ldots,m\} / \min\{n_i : i = 1,\ldots,m\} < C_1,\\
\max\{\tilde{n}_k : k = 1,\ldots,K\} / \min\{\tilde{n}_k : k = 1,\ldots,K\} < C_1.
\end{aligned}$$

(C2) The parameter space Θ is compact, containing $(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$ as an interior point. For $(\boldsymbol{\eta},\boldsymbol{\alpha}) \in \Theta$, $L_{ih}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ is convex, twice continuously differentiable with respect to $(\boldsymbol{\eta},\boldsymbol{\alpha})$ and $E\{|L_{ih}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})|\} < \infty$ for $i = 1,\ldots,m$ and $h = 1,\ldots,N_i$. Furthermore, $(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$ uniquely minimizes $E\{L_{ih}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\}$.

(C3) For $(\boldsymbol{\eta},\boldsymbol{\alpha}) \in \Theta$, $\boldsymbol{V}\{L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\,|\mathcal{F}\} \to 0$ in probability as $m \to \infty$, where $\boldsymbol{V}\{L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\,|\mathcal{F}\}$ is the variance of $L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ conditional on the finite population $\mathcal{F}$.

(C4) A central limit theorem holds for the weighted score function $\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$,

$$n_0^{1/2} \hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) \to N\left(\boldsymbol{0}, \boldsymbol{\Sigma}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)\right)$$

in distribution as $n_0 \to \infty$ with respect to the model (1) and the sampling mechanism, where $\boldsymbol{\Sigma}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)$ is positively definitive.

(C5) There exists a compact set $\mathcal{B} \subset \Theta$, such that $\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})$ converges to $\boldsymbol{\mathcal{I}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ in probability with respect to the model (1) and the sampling mechanism uniformly over $\mathcal{B}$ as $n_0 \to \infty$, where $\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) = \sum_{i=1}^{m} W_i \hat{\boldsymbol{I}}_i(\boldsymbol{\eta},\boldsymbol{\alpha})$, $\boldsymbol{\mathcal{I}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ is the limit of $\sum_{i=1}^{m} W_i E\{\boldsymbol{I}_i(\boldsymbol{\eta},\boldsymbol{\alpha})\}$ as $m \to \infty$, and $\boldsymbol{\mathcal{I}}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$ is invertible.

(C6) There exist two positive constants C_2, C_3 such that $C_2 < N_i n_i^{-1} \pi_{ih} < C_3$ for $i = 1,\ldots,m, h = 1,\ldots,N_i$ almost surely.

(C7) Denote $\rho_\gamma(t) = \lambda^{-1}p_\gamma(t,\lambda)$ as the scaled penalty function. The function $\rho_\gamma(t)$ is symmetric, non-decreasing, concave on $[0,\infty)$, and $\rho_\gamma(0) = 0$. The value of $\rho_\gamma(t)$ is a constant for $t \geq a\lambda$ and some constant $a > 0$. Its derivative $\rho'_\gamma(t)$ exists and is continuous except for a finite number of values of t, and $\rho'_\gamma(0+) = 1$.

(C8) Denote $\boldsymbol{y} = (y_{11},\ldots,y_{1,N_i},\ldots,y_{m,1},\ldots,y_{m,N_m})^{\mathrm{T}}$ and $\boldsymbol{\mu}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$ as the expected value vector of $\boldsymbol{y}$. The conditional distribution of $\boldsymbol{y}$ given covariates belongs to the canonical exponential family. For any vector $\boldsymbol{a}$ and $x > 0$, $P(|\boldsymbol{a}^{\mathrm{T}}\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\}| > \|\boldsymbol{a}\|x) \leq 2\exp(-c_1x^2)$, where $0 < c_1 < \infty$.

(C9) There exists a positive constant $c_x < \infty$ such that $\max_{i,h,l}|x_{ih,l}| \leq c_x$ for $l = 1,\ldots,p$ and $\max_{i,h,l}|z_{ih,l}| \leq c_x$ for $l = 1,\ldots,q$.

Condition (C1) contains some regularity assumptions for the model setup, and we can show that $W_i \asymp m^{-1}$ and $mn_i \asymp n_0$ for $i = 1,\ldots,m$, where $a_n \asymp b_n$ is equivalent to $a_n = O(b_n)$ and $b_n = O(a_n)$. The asymptotic orders of W_i and n_i are required to guarantee the convergence rate in Condition (C4). By Conditions (C2) and (C3), the convexity of $L_{ih}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ and the pointwise convergence of $V\{L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})|\mathcal{F}\}$ guarantee $L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) \to \sum_{i=1}^m W_iE\{L_i^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\}$ in probability uniformly with respect to the model (1) and the sampling mechanism; see Theorem 10.8 of [37] for details. Such uniform convergence guarantees the consistency of the proposed estimators; see Theorem 2.1 of [32] for details. The twice-continuously differentiable property of $L_{ih}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ guarantees that the weighted score function $\hat{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\alpha})$ is continuously differentiable, so the mean value theorem applies. Condition (C4) assumes a central limit theorem for $\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})$, which is used to derive the asymptotic properties of the oracle estimator and can be verified under general complex sampling designs; see Section C for details of Poisson sampling and stratified multistage cluster sampling. In Condition (C5), the convergence of $\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})$ is used to derive the limiting distribution of $(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or})$ [1, Theorem 4.1.2]. Condition (C6) is a regularity condition about inclusion probabilities, and it is used to establish the convergence rate of the estimating equation; see Theorem 1.3.3 of [13] for details. Conditions (C7)–(C9) are common assumptions for penalized regression in high-dimensional settings [29, 12].

Remark 2. *Under the assumption $m \to \infty$, our theoretical results apply as long as that domain sizes are of the same order, including the case where they are bounded. The analysis can be easily extended to the scenario where m is bounded but the domain sizes diverge at the same rate. Besides, we implicitly consider a stratified single-stage sampling design in this paper; see Appendix C.2 for a brief discussion under stratified multistage cluster sampling.*

The following Theorem 1 and Theorem 2 show theoretical properties of the oracle estimator in (13).

Theorem 1. *Under Conditions (C1)–(C3), we have $(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or}) \to (\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$ in probability as $m \to \infty$ with respect to the model (1) and the sampling mechanism.*

Theorem 2. *Under Conditions (C1)–(C5), we have*

$$n_0^{1/2}\left(\left(\hat{\boldsymbol{\eta}}^{or}-\boldsymbol{\eta}^0\right)^{\mathrm{T}},\left(\hat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^0\right)^{\mathrm{T}}\right)^{\mathrm{T}}\to N\left(\mathbf{0},\boldsymbol{\Sigma}\right) \tag{14}$$

in distribution as $m\to\infty$ *with respect to the model (1) and the sampling mechanism, where* $\boldsymbol{\Sigma}=\boldsymbol{\mathcal{I}}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)\boldsymbol{\mathcal{I}}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)^{-1}$, *and* $\boldsymbol{\mathcal{I}}(\boldsymbol{\eta},\boldsymbol{\alpha})=\sum_{i=1}^{m}W_iE\{\boldsymbol{I}_i\left(\boldsymbol{\eta},\boldsymbol{\alpha}\right)\}$.

The proofs of Theorems 1–2 are relegated to Appendices B.1–B.2, respectively. Theorem 1 and Theorem 2 assume the cluster structure is known and consider the theoretical properties of the oracle estimator. Theorem 1 shows the consistency of $\boldsymbol{\eta}^{or}$ and $\boldsymbol{\alpha}^{or}$, and Theorem 2 establishes the corresponding limiting distribution. Theorems 1–2 show that the oracle estimator can converge to the true parameters in a rate of $\sqrt{n}_0$.

Let $b=\min_{i\in\mathcal{G}_k,j\in\mathcal{G}_{k'},k\neq k'}\left\|\boldsymbol{\beta}_i^0-\boldsymbol{\beta}_j^0\right\|=\min_{k\neq k'}\left\|\boldsymbol{\alpha}_k^0-\boldsymbol{\alpha}_{k'}^0\right\|$ be the minimal difference between two distinct cluster-specific regression parameters. Theorem 3 establishes theoretical properties of our proposed estimator in (11).

Theorem 3. *Under Conditions (C1)–(C9), suppose* $b>a\lambda$ *for some constant* a, $\lambda\to 0$ *and* $n_0^{1/2}\lambda\to\infty$ *as* $m\to\infty$. *Then, there exists a local minimizer* $(\hat{\boldsymbol{\eta}}^{\mathrm{T}},\hat{\boldsymbol{\beta}}^{\mathrm{T}})$ *of the objective function* $Q_\omega(\boldsymbol{\eta},\boldsymbol{\beta})$ *such that* $P\{(\hat{\boldsymbol{\eta}}^{\mathrm{T}},\hat{\boldsymbol{\beta}}^{\mathrm{T}})=((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}},(\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})\}\to 1$.

The proof of Theorem 3 is in Appendix B.3. The theoretical properties are investigated under complex survey sampling, which is different from the existing studies. Theorem 3 implies that the true cluster structure can be recovered with probability approaching 1, and that the estimated number of clusters $\hat{K}$ satisfies $P(\hat{K}=K)\to 1$ under complex survey sampling.

Let $\hat{\boldsymbol{\alpha}}$ be the distinct cluster vectors of $\hat{\boldsymbol{\beta}}$. According to Theorem 2 and Theorem 3, we have the following result.

Corollary 1. *Suppose the conditions in Theorem 3 hold, we have*

$$n_0^{1/2}\left(\left(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}^0\right)^{\mathrm{T}},\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0\right)^{\mathrm{T}}\right)^{\mathrm{T}}\to N\left(\mathbf{0},\boldsymbol{\Sigma}\right)$$

in distribution as $m\to\infty$.

When λ satisfies the conditions in Theorem 3, and the cluster structure is recovered with probability approaching 1, Corollary 1 gives the support for conducting statistical inferences for $\hat{\boldsymbol{\alpha}}$.

4. Simulation studies

In this section, we conduct simulation studies to evaluate the performance of the proposed PCC method. We focus on the performance of identifying the cluster information and only consider cases without the global regression coefficient $\boldsymbol{\eta}$. We consider two types of models: linear regression models in Section 4.1 and

logistic regression models in Section 4.2, and compare the proposed PCC method with the one without using sampling weights, referred to as the "Clustered Coefficients (CC)" method.

To evaluate the cluster performance of the proposed PCC method, we report the estimated cluster number $\hat{K}$ and the adjusted Rand index (ARI) [35]. ARI measures the degree of agreement between two partitions; the largest value is 1. A larger ARI value indicates a better cluster performance. We report the average $\hat{K}$ and ARI across 1 000 simulations, along with standard deviation values in the parentheses. To evaluate the estimation accuracy of $\boldsymbol{\beta}$, we report the average root mean square error (RMSE), which is defined as

$$\left(\frac{1}{m} \sum_{i=1}^{m} \left\| \hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i \right\|^2 \right)^{1/2}.$$

4.1. Linear regression models

In this section, we consider a linear regression model with regression coefficient heterogeneity. We generate a finite population $\{(\boldsymbol{x}_{ih}, y_{ih}) : i = 1, \ldots, m, h = 1, \ldots, H\}$ with $m = 100$ and $H = 300$ by

$$y_{ih} = \boldsymbol{x}_{ih}^{\mathrm{T}} \boldsymbol{\beta}_i + \epsilon_{ih},$$

where $\boldsymbol{x}_{ih} = (1, x_{ih})^{\mathrm{T}}$, $x_{ih} \sim N(0, 1)$ independently, and conditionally on x_{ih}, ϵ_{ih} is independently generated from $N(0, \sigma_{ih}^2)$ with $\sigma_{ih} = \exp(0.5|\boldsymbol{x}_{ih}^{\mathrm{T}} \boldsymbol{\beta}_i|)$. Assume that there are three true clusters, denoted as $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$. The cluster parameters are $\boldsymbol{\beta}_i = (-1, -1)^{\mathrm{T}}$ for $i \in \mathcal{G}_1, \boldsymbol{\beta}_i = (0.5, 0.5)^{\mathrm{T}}$ for $i \in \mathcal{G}_2$ and $\boldsymbol{\beta}_i = (2, 2)^{\mathrm{T}}$ for $i \in \mathcal{G}_3$. Each domain has an equal probability of being assigned to one of these three clusters.

We conduct Poisson sampling to generate samples from the finite population with expected sample size $n_i = 10$, $n_i = 30$ and $n_i = 50$. For $i \in \mathcal{G}_1$ and $i \in \mathcal{G}_3$, we conduct Poisson sampling with unequal inclusion probabilities, setting $\pi_{ih} \propto \exp(0.3 x_{ih})$ for $i \in \mathcal{G}_1$, and $\pi_{ih} \propto \exp(0.7 x_{ih})$ for $i \in \mathcal{G}_3$. For $i \in \mathcal{G}_2$, We conduct Poisson sampling with equal inclusion probabilities.

Table 1 and Figure 1 summarize the results of the linear regression model. For $n_i = 10$, the proposed PCC method has a larger ARI than the CC method. A similar observation occurs when $n_i = 30$, with the proposed PCC method having the average cluster number closer to the true number of clusters. As n_i increases, both methods have larger ARI values, indicating a better identification of cluster structures. From the results in Figure 1, we can also observe that compared with the CC method, the proposed PCC method provides better estimation accuracy, as indicated by smaller RMSE values.

4.2. Logistic regression models

In this section, we consider a logistic regression model with regression coefficient heterogeneity. We generate a finite population $\{(\boldsymbol{x}_{ih}, y_{ih}) : i = 1, \ldots, m, h =$

TABLE 1
Summary of $\hat{K}$ and average ARI for the linear regression model based on 1 000 simulations, along with standard deviations in the parentheses. CC and PCC represent the methods without and with sampling weights, respectively.

	$\hat{K}$		ARI	
	CC	PCC	CC	PCC
$n_i = 10$	5.04(1.111)	3.40(0.582)	0.94(0.043)	0.99(0.02)
$n_i = 30$	5.00(0.816)	3.07(0.282)	0.94(0.048)	1(0.015)
$n_i = 50$	5.13(0.695)	3.06(0.248)	0.95(0.045)	1(0.01)

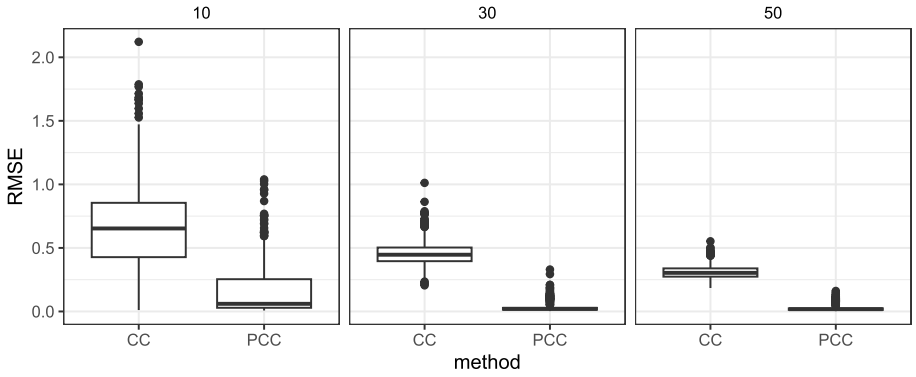

FIG 1. *RMSE of $\hat{\beta}_i$ for the linear regression model based on 1 000 simulations with expected sample size $n_i = 10$ (left), $n_i = 30$ (middle) and $n_i = 50$ (right).*

$1, \ldots, H\}$ with $m = 100$ and $H = 300$ by

$$y_{ih} \sim \text{Bernoulli}(p_{ih})$$

independently with $\text{logit}(p_{ih}) = \boldsymbol{x}_{ih}^{\mathrm{T}}\boldsymbol{\beta}_i$ where $\boldsymbol{x}_{ih} = (1, x_{ih})^{\mathrm{T}}$, x_{ih} is simulated from Uniform(−2,2), and $\text{logit}(x) = \log x - \log(1-x)$ for $x \in (0,1)$. As in the preceding section, we assume there are 3 true clusters, denoted as $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$. The cluster parameters are $\boldsymbol{\beta}_i = (-1, 0.5)^{\mathrm{T}}$ for $i \in \mathcal{G}_1, \boldsymbol{\beta}_i = (0.5, 1.5)^{\mathrm{T}}$ for $i \in \mathcal{G}_2$ and $\boldsymbol{\beta}_i = (2, -0.5)^{\mathrm{T}}$ for $i \in \mathcal{G}_3$. We still conduct Poisson sampling to generate samples from the finite population with expected sample size $n_i = 10$, $n_i = 30$ and $n_i = 50$. For $i \in \mathcal{G}_1$ and $i \in \mathcal{G}_3$, we conduct Poisson sampling with unequal inclusion probabilities and set $\pi_{ih} \propto \exp(0.3x_{ih} + 0.5y_{ih})$ for $i \in \mathcal{G}_1$, and $\pi_{ih} \propto \exp(0.7x_{ih} + 0.5y_{ih})$ for $i \in \mathcal{G}_3$. For $i \in \mathcal{G}_2$, we conduct Poisson sampling with equal inclusion probabilities.

Table 2 shows the results for $\hat{K}$ and ARI, and Figure 2 shows the results for RMSE. It can be seen that when the $n_i = 10$, the local sample size is not large, both methods cannot identify the cluster structure well, but the performance of the proposed PCC method is better. When local sample sizes increase to $n_i = 30$, both methods have better results, and the proposed PCC method outperforms the CC method in terms of larger ARI and smaller RMSE values. When $n_i = 50$, both methods have similar ARI values, but the estimated number of clusters in

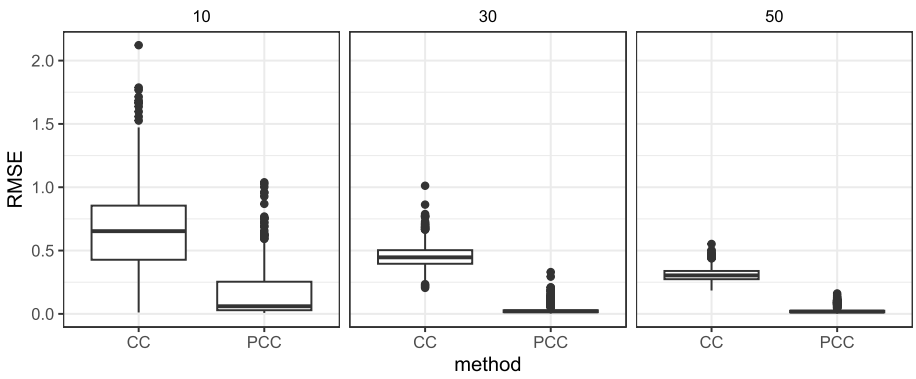

FIG 2. *RMSE of $\hat{\beta}_i$ for the logistic regression model based on 1 000 simulations with expected sample size $n_i = 10$ (left), $n_i = 30$ (middle) and $n_i = 50$ (right).*

PCC is closer to the truth number of clusters, and PCC also has smaller RMSE values.

TABLE 2
Summary of $\hat{K}$ and average ARI for the logistic regression model based on 1 000 simulations, along with standard deviations in the parentheses. CC and PCC represent the methods without and with sampling weights, respectively.

	$\hat{K}$		ARI	
	CC	PCC	CC	PCC
$n_i = 10$	4.46(1.013)	3.76(0.899)	0.26(0.135)	0.36(0.093)
$n_i = 30$	2.85(1.011)	3.92(0.939)	0.63(0.137)	0.84(0.108)
$n_i = 50$	3.65(0.806)	3.20(0.46)	0.95(0.072)	0.95(0.045)

We also consider three additional examples where the datasets are not simulated from the true models in Appendix D. In example 1, we simulate data from linear mixed models. We find that when the variance of random effects is large, our PCC can capture the random effects using clustered fixed effects, while CC cannot. In example 2 and example 3, we simulate data from linear regression models with an additional logarithm scale effect. The results show that a large sample size is needed to recover the cluster structure accurately when this additional effect is large. Our proposed PCC still performs better than CC without sampling weights.

5. Application

In this section, we apply the proposed PCC method to the American Community Survey (ACS), which is a demographics survey program conducted by the U.S. Census Bureau. This program regularly collects information previously contained only in the long form of the decennial census, such as ancestry, citizenship, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics.

The Public Use Microdata Sample (PUMS) files are publicly available data at both individual and household levels. The most detailed unit of geography contained in the PUMS files is the Public Use Microdata Area (PUMA). PUMA contains special non-overlapping areas that partition each state into contiguous geographic units containing roughly 100 000 people at the time of their creation. In this section, we use the household level PUMS data and model the relationship between the logarithm of the monthly gross rent and the logarithm of the household income across 43 PUMAs in Minnesota based on ACS 2021 data. In this dataset, the number of observations in each PUMA ranges from 26 to 163.

We consider the following model,

$$y_{ih} = \beta_{0,i} + \beta_{1,i} x_{ih} + \epsilon_{ih},$$

where y_{ih} is the logarithm of the monthly gross rent for the hth household in the ith PUMA, and x_{ih} is the logarithm of household income for the hth household in the ith PUMA. We assume there exists coefficient heterogeneity.

Tables 3–4 provide the estimated regression coefficients in different clusters with standard error in the parenthesis for the CC and PCC methods, respectively. By the CC method, all PUMAs are divided into three clusters. All clusters have positive estimates of β_1. This result implies that the monthly gross is expected to increase as the household income increases in all clusters. Among these three clusters, the second cluster has the largest $\hat{\beta}_1$. By using the proposed PCC method, four clusters are identified with different estimated regression coefficients. The estimated coefficients also show a positive relationship between household income and gross rent, which is the strongest in the third cluster, followed by the second cluster, and then the fourth cluster. Table 5 presents a contingency table that displays the number of PUMAs in each cluster. The PUMAs in CC(1) are mainly divided into two clusters by the proposed PCC method. From the estimates in Table 4, it can be seen that the estimated regression coefficients of PCC(1) and PCC(2) are quite different, which indicates that it is more reasonable to have two clusters when considering sampling weights. Furthermore, the PUMAs in CC(3) are separated into PCC(3) and PCC(4), which have different estimates for $\hat{\beta}_0$.

Table 3
Summary of estimated regression coefficients by the CC method.

cluster	1	2	3
$\hat{\beta}_0$	−0.387(0.023)	−0.003(0.029)	0.347(0.023)
$\hat{\beta}_1$	0.283(0.022)	0.62(0.03)	0.439(0.023)

Table 4
Summary of estimated coefficients by the proposed PCC method.

cluster	1	2	3	4
$\hat{\beta}_0$	−0.292(0.032)	−0.456(0.033)	0.112(0.021)	0.479(0.035)
$\hat{\beta}_1$	0.201(0.03)	0.395(0.035)	0.57(0.021)	0.353(0.036)

TABLE 5
The number of PUMAs in each cluster for the CC method and the proposed PCC method. CC(i) represents the ith cluster in the CC method, where $i = 1, 2, 3$. PCC(j) represents the jth cluster in the proposed PCC method, where $j = 1, 2, 3, 4$.

cluster	PCC(1)	PCC(2)	PCC(3)	PCC(4)
CC(1)	7	5	1	0
CC(2)	0	2	8	0
CC(3)	0	0	8	12

6. Summary and conclusion

In this paper, we propose a unified framework (PCC) to estimate regression coefficients as well as identify the cluster structure for domains simultaneously under complex survey sampling. Theoretical properties of the proposed PCC method are investigated under regularity conditions without assuming an explicit form for the model. We also developed algorithms for linear regression and logistic regression models. In the simulation study, the proposed PCC method is compared with a CC method under the setup of a linear regression model and a logistic regression model, and the results demonstrate the importance of sampling weights in identifying the cluster structure under complex survey sampling. Future research can explore new estimators based on probability weighted clustered coefficient regression models for different types of survey data, such as continuous data, binary data, and count data.

Appendix A: Computation

In this section, the detailed algorithms for updating $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are introduced. In Section A.1, updates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ for linear regression models are described. In Section A.2, updates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ for logistic regression models are described.

A.1. Linear regression models

In linear regression model, $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ are updated by the minimizing

$$f(\boldsymbol{\beta}, \boldsymbol{\eta}) = \left\| \boldsymbol{\Omega}^{1/2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta}) \right\|^2 + \vartheta \left\| \boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{\delta}^{(r)} + \vartheta^{-1}\boldsymbol{v}^{(r)} \right\|^2,$$

where

$$\boldsymbol{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{m,n_m})^{\mathrm{T}}$$
$$\boldsymbol{Z} = (\boldsymbol{z}_{11}, \ldots, \boldsymbol{z}_{1n_1}, \ldots, \boldsymbol{z}_{m1}, \ldots, \boldsymbol{z}_{m,n_m})^{\mathrm{T}}$$
$$\boldsymbol{X} = \mathrm{diag}\,(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m)$$
$$\boldsymbol{\Omega} = m\,\mathrm{diag}(W_1/N_1\pi_{11}^{-1}, \ldots W_1/N_1\pi_{1n_1}, \ldots, W_m/N_m\pi_{m1}^{-1}, \ldots, W_m/N_m\pi_{mn_m}^{-1})$$

where $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n_i})^{\mathrm{T}}$, $\boldsymbol{A} = \boldsymbol{D} \otimes \boldsymbol{I}_p$ with an $m(m-1)/2 \times m$ matrix $\boldsymbol{D} = \{(\boldsymbol{e}_i - \boldsymbol{e}_j) : 1 \le i < j \le m\}^{\mathrm{T}}$, $\otimes$ is the Kronecker product, $\boldsymbol{I}_p$ is a $p \times p$

identity matrix, and $\boldsymbol{e}_i$ is an $m\times 1$ vector with ith element 1 and other elements 0. Then, the solutions for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are

$$\begin{aligned}&\boldsymbol{\beta}^{(r+1)}\\ &=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}_{Z,\Omega}\boldsymbol{X}+\vartheta\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}_{Z,\Omega}\boldsymbol{y}+\vartheta\,\mathrm{vec}\left\{\left(\boldsymbol{H}^{(r)}-\vartheta^{-1}\boldsymbol{\Upsilon}^{(r)}\right)\boldsymbol{D}\right\}\right],\\ &\boldsymbol{\eta}^{(r+1)}=\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Pi}^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Pi}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(r+1)}\right),\end{aligned}$$

where $\boldsymbol{Q}_{Z,\Omega}=\boldsymbol{\Omega}-\boldsymbol{\Omega}\boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Omega}$, $\boldsymbol{H}^{(r)}=\{\boldsymbol{\zeta}_{ij}^{(r)},1\le i<j\le m\}$ and $\boldsymbol{\Upsilon}^{(r)}=\{\boldsymbol{v}_{ij}^{(r)}:1\le i<j\le m\}$. $\boldsymbol{H}^{(r)}$ and $\boldsymbol{\Upsilon}^{(r)}$ are two matrices with dimension $p\times m(m-1)/2$.

A.2. Logistic regression models

When updating $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$, consider the following objective function,

$$f\left(\boldsymbol{\eta},\boldsymbol{\beta}\right)=L_{\omega}(\boldsymbol{\eta},\boldsymbol{\beta})+\frac{1}{2}\nu\|\boldsymbol{A}\boldsymbol{\beta}-\boldsymbol{\delta}^{(r)}+\nu^{-1}\boldsymbol{v}^{(r)}\|^2,$$

where

$$L_{\omega}(\boldsymbol{\eta},\boldsymbol{\beta})=-m\sum_{i=1}^{m}\sum_{h=1}^{n_i}\frac{W_i}{N_i}\frac{1}{\pi_{ih}}\left(y_{ih}(\boldsymbol{z}_{ih}^{T}\boldsymbol{\eta}+\boldsymbol{x}_{ih}^{T}\boldsymbol{\beta}_i)-\log(1+e^{\boldsymbol{z}_{ih}^{T}\boldsymbol{\eta}+\boldsymbol{x}_{ih}^{T}\boldsymbol{\beta}_i})\right).$$

Let $\boldsymbol{W}=\mathrm{diag}\left(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m\right)$, where $\boldsymbol{w}_i=\frac{mW_i}{N_i}(\pi_{i1},,\ldots,\pi_{i,n_i})$. Denote $\mu_{ih}=\frac{\exp(\boldsymbol{z}_{ih}^{T}\boldsymbol{\eta}+\boldsymbol{x}_{ih}^{T}\boldsymbol{\beta}_i)}{1+\exp(\boldsymbol{z}_{ih}^{T}\boldsymbol{\eta}+\boldsymbol{x}_{ih}^{T}\boldsymbol{\beta}_i)}$, $\boldsymbol{\mu}\left(\boldsymbol{\eta},\boldsymbol{\beta}\right)=(\mu_{ih},i=1\ldots,m,h=1,\ldots n_i)$, and $\boldsymbol{V}=\mathrm{diag}\left(\boldsymbol{\mu}\left(\boldsymbol{\eta},\boldsymbol{\beta}\right)\left(1-\boldsymbol{\mu}\left(\boldsymbol{\eta},\boldsymbol{\beta}\right)\right)\right)$. Based on the Newton-Raphson algorithm, we can get updates $\boldsymbol{\eta}^{(r+1)}$ and $\boldsymbol{\beta}^{(r+1)}$ at sth inner step as below,

$$\begin{aligned}&\boldsymbol{\beta}^{(r+1,s+1)}\\ =&\boldsymbol{\beta}^{(r+1,s)}+\\ &\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{V}\boldsymbol{X}+\nu\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\\ &\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{y}-\boldsymbol{\mu}\left(\boldsymbol{\eta}^{(r+1,s)},\boldsymbol{\beta}^{(r+1,s)}\right)\right)-\nu\boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{\beta}^{(r+1,s)}-\boldsymbol{\delta}^{(r)}+\nu^{-1}\boldsymbol{v}^{(r)}\right)\right),\end{aligned}$$

and

$$\boldsymbol{\eta}^{(r+1,s+1)}=\boldsymbol{\eta}^{(r+1,s)}+\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{V}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{y}-\boldsymbol{\mu}\left(\boldsymbol{\eta}^{(r+1,s)},\boldsymbol{\beta}^{(r+1,s)}\right)\right).$$

Appendix B: Proofs of theoretical results

B.1. Proof of Theorem 1

The proof of Theorem 1 follows Theorem 2.1 in [32]. In this paper, we provide detailed proof of Theorem 1.

Recall that $L_N^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) = \sum_{i=1}^m W_i N_i^{-1} \sum_{h=1}^{N_i} L_i^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$, where $L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ is the loss function for the first element in the ith unit. By Condition (C2) and the weak law of large numbers, we have

$$L_N^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) - \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} \to 0$$

in probability with respect to the joint distribution of $\{(\boldsymbol{y}_{ih}, \boldsymbol{x}_{ih}, \boldsymbol{z}_{ih}) : i = 1,\ldots,m; h = 1,\ldots,N_i\}$. Since $E\{L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\,|\mathcal{F}\} = L_N^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$, it follows from Condition (C3) that, for $(\boldsymbol{\eta},\boldsymbol{\alpha}) \subset \Theta$,

$$L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) - L_N^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) \to 0$$

in probability with respect to the sampling mechanism as $m \to \infty$. Applying Theorem 2 of [17], we can show that

$$L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) - \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} \to 0 \tag{15}$$

as $m \to \infty$ in probability with respect to the super-population model and the sampling mechanism uniformly over Θ, where Θ is the overall parameter space.

Since $(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}) = \underset{(\boldsymbol{\eta},\boldsymbol{\alpha})\in\Theta}{\arg\min} L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$, we have

$$L_\omega^{\mathcal{G}}(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}) \le L_\omega^{\mathcal{G}}(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0).$$

By the uniform convergence in (15) over Θ, for any $\epsilon > 0$, we have

$$\left| L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) - \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} \right| < \epsilon/10$$

for sufficiently large m with probability approaching to 1 with respect to the super-population model and the sampling mechanism. More specifically, we have the following two results

$$\sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} < L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) + \epsilon/10, \tag{16}$$

$$L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) < \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} + \epsilon/10. \tag{17}$$

For any open set $\mathscr{O}$ containing $\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)$ as an interior point, $\mathscr{O}^C \cap \Theta$ is compact. Since $E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\}$ is continuous, by extreme value theorem, we have $\min_{(\boldsymbol{\eta},\boldsymbol{\alpha})\in(\mathscr{O}^C\cap\Theta)} L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})$ exists.

By Condition (C2), $E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\}$ is uniquely minimized at the true parameter $\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)$. We have

$$\min_{(\boldsymbol{\eta},\boldsymbol{\alpha})\in(\mathscr{O}^C\cap\Theta)} \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} > \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\}.$$

Denote $\epsilon = \min_{(\boldsymbol{\eta},\boldsymbol{\alpha})\in(\mathscr{O}^C\cap\Theta)} \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} - \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\}$. By (16) and (17), we have

$$\begin{aligned}\min_{(\boldsymbol{\eta},\boldsymbol{\alpha})\in(\mathscr{O}^C\cap\Theta)} L_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) &\geq \min_{(\boldsymbol{\eta},\boldsymbol{\alpha})\in(\mathscr{O}^C\cap\Theta)} \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha})\} - \epsilon/10\\ &= \sum_{i=1}^m W_i E\{L_{i1}^{\mathcal{G}}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\} + \epsilon - \epsilon/10\\ &> L_\omega^{\mathcal{G}}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) + 8\epsilon/10\\ &> L_\omega^{\mathcal{G}}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\\ &\geq L_\omega^{\mathcal{G}}(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or}).\end{aligned}$$

This shows that $(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or}) \notin (\mathscr{O}^C\cap\Theta)$, which also means $(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or}) \in \mathscr{O}$ with probability approaching 1. Since $\mathscr{O}$ is arbitrary, we can conclude that $(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or})$ is consistent for $(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$ by letting $\mathscr{O}$ be small enough. This completes the proof of Theorem 1.

B.2. Proof of Theorem 2

In Theorem 1, we have shown that

$$\begin{pmatrix}\hat{\boldsymbol{\eta}}^{or}\\ \hat{\boldsymbol{\alpha}}^{or}\end{pmatrix} \to \begin{pmatrix}\boldsymbol{\eta}^0\\ \boldsymbol{\alpha}^0\end{pmatrix}$$

in probability as $m\to\infty$ with respect to superpopulation model and sampling mechanism. By Condition (C2), $\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})$ is continuously differentiable. Based on the definition of the oracle estimator, we have $\hat{\boldsymbol{S}}_\omega(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or}) = \mathbf{0}$.

By mean value theorem, there exists $(\boldsymbol{\eta}^\dagger,\boldsymbol{\alpha}^\dagger)$ on the segment joining $(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or})$ and $(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$ such that

$$\begin{aligned}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) &= -\frac{\partial}{\partial(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^\dagger,\boldsymbol{\alpha}^\dagger)\left\{\left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}},(\hat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}} - \left((\boldsymbol{\eta}^0)^{\mathrm{T}},(\boldsymbol{\alpha}^0)^{\mathrm{T}}\right)^{\mathrm{T}}\right\}\\ &= \hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta}^\dagger,\boldsymbol{\alpha}^\dagger)\left\{\left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}},(\hat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}} - \left((\boldsymbol{\eta}^0)^{\mathrm{T}},(\boldsymbol{\alpha}^0)^{\mathrm{T}}\right)^{\mathrm{T}}\right\}.\end{aligned}$$

By Condition (C5) and Theorem 1, we have

$$\boldsymbol{\mathcal{I}}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)^{-1}\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta}^\dagger,\boldsymbol{\alpha}^\dagger) \to \boldsymbol{I} \tag{18}$$

in probability with respect to the super-population model and sampling mechanism. Then,

$$\begin{aligned}&n_0^{1/2}\boldsymbol{\mathcal{I}}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)^{-1}\boldsymbol{I}_\omega(\boldsymbol{\eta}^\dagger,\boldsymbol{\alpha}^\dagger)\left\{(\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}},(\hat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}} - ((\boldsymbol{\eta}^0)^{\mathrm{T}},(\boldsymbol{\alpha}^0)^{\mathrm{T}})^{\mathrm{T}}\right\}\\ &= n_0^{1/2}\boldsymbol{\mathcal{I}}\left(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0\right)^{-1}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0).\end{aligned} \tag{19}$$

By Conditions (C4)–(C5), (18), (19) and Slutsky's theorem, we have completed the proof of Theorem 2.

B.3. Proof of Theorem 3

Recall that $L_\omega(\boldsymbol{\eta}, \boldsymbol{\beta})$ is the loss function when the cluster information is unknown, and $L_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$ is the loss function when the cluster information is known. In addition, the penalty terms with unknown cluster information and known cluster information have the following forms,

$$P_m(\boldsymbol{\beta}) = \lambda \sum_{i<j} \rho\left(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|\right);$$
$$P_m^{\mathcal{G}}(\boldsymbol{\alpha}) = \lambda \sum_{k<k'} |\mathcal{G}_k \| \mathcal{G}_{k'}| \, \rho\left(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|\right).$$

Then, the corresponding objective functions, $Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta})$, and $Q_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$, have the following forms, respectively,

$$Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) = mL_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) + P_m(\boldsymbol{\beta}),$$
$$Q_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = mL_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) + P_m^{\mathcal{G}}(\boldsymbol{\alpha}).$$

Define $\mathcal{M}_{\mathcal{G}} = \{\boldsymbol{\beta} \in \mathbb{R}^{mp} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j, \text{ for } i, j \in \mathcal{G}_k, 1 \le k \le K\}$. From the definition, we know that for $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$, we can write $\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{\alpha}$. Let $T : \mathcal{M}_{\mathcal{G}} \to \mathbb{R}^{Kp}$ be the mapping that $T(\boldsymbol{\beta})$ is the $Kp \times 1$ vector consisting of K vectors with dimension p and its k^{th} vector component equals to the common value of $\boldsymbol{\beta}_i$ for $i \in \mathcal{G}_k$. Let $T^* : \mathbb{R}^{mp} \to \mathbb{R}^{Kp}$ be the mapping such that $T^*(\boldsymbol{\beta}) = \{|\mathcal{G}_k|^{-1} \sum_{i \in \mathcal{G}_k} \boldsymbol{\beta}_i^{\mathrm{T}}, k = 1, \ldots, K\}^{\mathrm{T}}$. Clearly, when $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}, T(\boldsymbol{\beta}) = T^*(\boldsymbol{\beta})$.

For every $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$, we have $P_m(\boldsymbol{\beta}) = P_m^{\mathcal{G}}(T(\boldsymbol{\beta}))$. And for every $\boldsymbol{\alpha} \in \mathbb{R}^{Kp}$, we have $P_m\left(T^{-1}(\boldsymbol{\alpha})\right) = P_m^{\mathcal{G}}(\boldsymbol{\alpha})$. Thus, we have

$$Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) = Q_\omega^{\mathcal{G}}(\boldsymbol{\eta}, T(\boldsymbol{\beta})), \quad Q_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = Q_\omega\left(\boldsymbol{\eta}, T^{-1}(\boldsymbol{\alpha})\right).$$

By Theorem 2, there exists a positive constant M, such that

$$\left\|\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0\right\| \le M n_0^{-1/2}, \max_i \left\|\hat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0\right\| \le M n_0^{-1/2}.$$

Denote the neighborhood of $\left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0\right)$ as

$$\Theta = \left\{\boldsymbol{\eta} \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathbb{R}^{mp} : \left\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\right\| \le M n_0^{-1/2}, \max_i \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right\| \le M n_0^{-1/2}\right\}.$$

Obviously, $\left(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or}\right) \in \Theta$. For any $\boldsymbol{\beta} \in \mathbb{R}^{mp}$, let $\boldsymbol{\beta}^* = T^{-1}\left(T^*(\boldsymbol{\beta})\right)$. We will show that $\left(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or}\right)$ is a strictly local minimizer of the objective function with probability approaching 1 through the following two steps.

(i) For any $\left(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}} \in \Theta$ and $\left(\boldsymbol{\eta}^{\mathrm{T}}, (\boldsymbol{\beta}^*)^{\mathrm{T}}\right)^{\mathrm{T}} \neq \left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$,

$$Q_\omega\left(\boldsymbol{\eta}, \boldsymbol{\beta}^*\right) > Q_\omega(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or}).$$

(ii) There is a neighborhood of $\left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$, denoted by Θ_m such that for any $\left(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}} \in \Theta_m \cap \Theta$ and sufficiently large m,

$$Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) \geq Q_\omega\left(\boldsymbol{\eta}, \boldsymbol{\beta}^*\right)$$

Therefore, by the results in (i) and (ii), we have $Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) > Q_\omega(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ for any $\left(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}} \in \Theta_m \cap \Theta$ and $\left(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}} \neq \left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$, so that $\left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$ is a strict local minimizer of $Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta})$ for sufficiently large m.

In the following we prove the result in (i). We first show $P_m^{\mathcal{G}}\left(T^*(\boldsymbol{\beta})\right) = C_m$ for any $\boldsymbol{\beta} \in \Theta$, where C_m is a constant which does not depend on $\boldsymbol{\beta}$. Let $T^*(\boldsymbol{\beta}) = \boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_K^{\mathrm{T}}\right)^{\mathrm{T}}$. It suffices to show that $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\| > a\lambda$ for all $k \neq k'$. Then, by Condition (C7), $\rho_\gamma\left(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|\right)$ is a constant, and as a result $P_m^{\mathcal{G}}\left(T^*(\boldsymbol{\beta})\right)$ is a constant. Since

$$\begin{aligned}
\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\| &= \left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0 + \boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0 + \boldsymbol{\alpha}_{k'}^0 - \boldsymbol{\alpha}_{k'}\right\| \\
&\geq \left\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\right\| - \left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0 + \boldsymbol{\alpha}_{k'}^0 - \boldsymbol{\alpha}_{k'}\right\| \\
&\geq \left\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\right\| - 2\max_k \left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\right\|,
\end{aligned}$$

and we have

$$\begin{aligned}
&\max_k \left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\right\|^2 \\
&= \max_k \left\| |\mathcal{G}_k|^{-1} \sum_{i \in \mathcal{G}_k} \boldsymbol{\beta}_i - \boldsymbol{\alpha}_k^0 \right\|^2 = \max_k \left\| |\mathcal{G}_k|^{-1} \sum_{i \in \mathcal{G}_k} \left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right) \right\|^2 \\
&= \max_k |\mathcal{G}_k|^{-2} \left\| \sum_{i \in \mathcal{G}_k} \left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right) \right\|^2 \leq \max_k |\mathcal{G}_k|^{-1} \sum_{i \in \mathcal{G}_k} \left\|\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right)\right\|^2 \\
&\leq \max_i \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right\|^2 \leq M^2 n_0^{-1}.
\end{aligned}$$

Then for all k and k',

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\| \geq \left\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\right\| - 2\max_k \left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\right\| \geq b - 2Mn_0^{-1/2} > a\lambda,$$

where the last inequality follows from the assumption that $b > a\lambda$ and $\lambda \gg n_0^{-1/2}$. Therefore, we have $P_m^{\mathcal{G}}\left(T^*(\boldsymbol{\beta})\right) = C_m$.

Since $\left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$ is the unique global minimizer of $L_\omega^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$, then $L_\omega^{\mathcal{G}}\left(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})\right) > L_\omega^{\mathcal{G}}\left(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}\right)$ for all $\left(\boldsymbol{\eta}^{\mathrm{T}}, \left(T^*(\boldsymbol{\beta})\right)^{\mathrm{T}}\right)^{\mathrm{T}} \neq \left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$ and hence $Q_\omega^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) > Q_\omega^{\mathcal{G}}(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or})$ for all $T^*(\boldsymbol{\beta}) \neq \hat{\boldsymbol{\alpha}}^{or}$ due to that $P_m^{\mathcal{G}}(T^*(\boldsymbol{\beta})) = P_m^{\mathcal{G}}(\hat{\boldsymbol{\alpha}}^{or})$. Since we have $Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) = Q_\omega^{\mathcal{G}}(\boldsymbol{\eta}, T(\boldsymbol{\beta}))$ for $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$

and $Q_\omega^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) = Q_\omega\left(\boldsymbol{\eta}, T^{-1}(\boldsymbol{\alpha})\right)$, we have $Q_\omega^{\mathcal{G}}(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\alpha}}^{or}) = Q_\omega(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\beta}}^{or})$ and $Q_\omega^{\mathcal{G}}\left(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})\right) = Q_\omega\left(\boldsymbol{\eta}, T^{-1}\left(T^*(\boldsymbol{\beta})\right)\right) = Q_\omega\left(\boldsymbol{\eta},\boldsymbol{\beta}^*\right)$. Therefore,

$$Q_\omega\left(\boldsymbol{\eta},\boldsymbol{\beta}^*\right) > Q_\omega(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\beta}}^{or})$$

for all $\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}^{or}$, and the result in (i) is proved.

Next, we prove the result in (ii).

For a positive sequence t_m, let $\Theta_m = \left\{\boldsymbol{\beta}_i : \max_i \left\|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\right\| \leq t_m\right\}$.

For $\left(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}} \in \Theta_m \cap \Theta$, by the mean value theorem, we have

$$Q_\omega(\boldsymbol{\eta},\boldsymbol{\beta}) - Q_\omega\left(\boldsymbol{\eta},\boldsymbol{\beta}^*\right) = m\Gamma_1 + \Gamma_2,$$

where

$$\Gamma_1 = \frac{\partial L_\omega(\boldsymbol{\eta},\boldsymbol{\beta}^s)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right),$$

$$\Gamma_2 = \sum_{i=1}^{m}\frac{\partial P_m\left(\boldsymbol{\beta}^s\right)}{\partial\boldsymbol{\beta}_i^{\mathrm{T}}}\left(\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*\right),$$

where $\boldsymbol{\beta}^s = \alpha\boldsymbol{\beta} + (1-\alpha)\boldsymbol{\beta}^*$ for some constant $\alpha \in (0,1)$.

We know that,

$$\max_i \left\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0\right\|^2 = \max_k \left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\right\|^2 \leq M^2 n_0^{-1}.$$

Since $\boldsymbol{\beta}_i^s$ is between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i^*$, then,

$$\begin{aligned}\max_i \left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_i^0\right\| &\leq \alpha \max_i \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right\| + (1-\alpha)\max_i \left\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0\right\| \\ &\leq \alpha M n_0^{-1/2} + (1-\alpha)M n_0^{-1/2} = M n_0^{-1/2}.\end{aligned}$$

We will first show the bound of Γ_1. The partial derivative with respect to $\boldsymbol{\beta}$ is denoted as $\tilde{\boldsymbol{S}}(\boldsymbol{\eta},\boldsymbol{\beta}) = \partial L_\omega(\boldsymbol{\eta},\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}$. Then, we have

$$\begin{aligned}\frac{\partial L_\omega(\boldsymbol{\eta},\boldsymbol{\beta}^s)}{\partial\boldsymbol{\beta}^{\mathrm{T}}} &= \sum_{i=1}^{m} W_i N_i^{-1}\sum_{h=1}^{N_i}\delta_{ih}\pi_{ih}^{-1}\frac{\partial L_{ih}(\boldsymbol{\eta},\boldsymbol{\beta}^s)}{\partial\boldsymbol{\beta}^{\mathrm{T}}} \\ &= \sum_{i=1}^{m} W_i\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s),\end{aligned}$$

where $\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s) = N_i^{-1}\sum_{h=1}^{N_i}\delta_{ih}\pi_{ih}^{-1}\partial L_{ih}(\boldsymbol{\eta},\boldsymbol{\beta}^s)/\partial\boldsymbol{\beta}^{\mathrm{T}}$. By Condition (C5) and the mean value theorem, there exists $(\tilde{\boldsymbol{\eta}},\tilde{\boldsymbol{\beta}})$ lying between $(\boldsymbol{\eta},\boldsymbol{\beta}^s)$ and $(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)$ such that $\left\|\partial^2 L_{ih}(\boldsymbol{\eta},\boldsymbol{\beta})/\partial(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}\partial(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})|_{(\tilde{\boldsymbol{\eta}},\tilde{\boldsymbol{\beta}})}\right\| \leq C_4$ for sufficiently large m, where C_4 is a positive constant. Combined with Condition (C6), we have

$$\begin{aligned}&\max_i \left\|\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s) - \tilde{\boldsymbol{S}}_i(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)\right\| \\ &\leq \max_i \frac{1}{N_i}\sum_{h=1}^{N_i}\delta_{ih}\pi_{ih}^{-1}\{\left\|\partial L_{ih}(\boldsymbol{\eta},\boldsymbol{\beta}^s)/\partial\boldsymbol{\beta}^{\mathrm{T}} - \partial L_{ih}(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)/(\boldsymbol{\beta}^0)^{\mathrm{T}}\right\|\}\end{aligned}$$

$$
\begin{aligned}
&\leq \max_i \frac{1}{N_i} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-1} \left\| \frac{\partial^2 L_{ih}(\boldsymbol{\eta}, \boldsymbol{\beta})}{\partial (\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \partial (\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})} \Big|_{(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\beta}})} \{ (\boldsymbol{\eta}^{\mathrm{T}}, (\boldsymbol{\beta}^s)^{\mathrm{T}}) - ((\boldsymbol{\eta}^0)^{\mathrm{T}}, (\boldsymbol{\beta}^0)^{\mathrm{T}}) \} \right\| \\
&\leq N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-1} C_4 M (q + Kp) n_0^{-1/2} \\
&\leq n_0^{-1/2} N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih} C_4 C_2^{-1} N_i n_i^{-1} M (q + Kp) \\
&= O_p(n_0^{-1/2}). \qquad (20)
\end{aligned}
$$

Moreover, it is known that $\left\| \tilde{\boldsymbol{S}}_i \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) \right\| \leq \sum_{l=1}^{p} \left| \tilde{\boldsymbol{S}}_i^l \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) \right|$, where

$$
\tilde{\boldsymbol{S}}_i^l \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) = \frac{1}{N_i} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-1} x_{ih,l} \left(y_{ih} - \mu_{ih} \right).
$$

We know that $\frac{1}{N_i^2} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-2} x_{ih,l}^2 \leq \frac{1}{n_i^2 C_2} \sum_{h=1}^{N_i} \delta_{ih} x_{ih,l}^2 \leq \frac{c_x^2}{n_i C_2} \leq \frac{c_x^2}{\min_i n_i C_2}$ by Condition (C9). Then, by Condition (C8), we have

$$
\begin{aligned}
&P \left(\left| \tilde{\boldsymbol{S}}_i^l \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) \right| \geq \sqrt{2 c_1^{-1} C_2^{-1} c_x^2} \sqrt{\frac{\log m}{\min_i n_i}} \right) \\
&= P \left(\left| \frac{1}{N_i} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-1} x_{ih,l} \left(y_i - \mu_i \right) \right| \geq \sqrt{2 c_1^{-1} C_2^{-1} c_x^2 \frac{\log m}{\min_i n_i}} \right) \\
&\leq P \left(\left| \frac{1}{N_i} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-1} x_{ih,l} \left(y_{ih} - \mu_{ih} \right) \right| \geq \sqrt{\frac{1}{N_i^2} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-2} x_{ih,l}^2} \sqrt{2 c_1^{-1} \log m} \right) \\
&\leq 2 \exp \left(-c_1 2 c_1^{-1} \log m \right) = 2 \exp \left(-2 \log m \right) = \frac{2}{m^2}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&P \left(\max_i \left\| \tilde{\boldsymbol{S}}_i \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) \right\| \geq \sqrt{2 c_1^{-1} C_2^{-1} c_x^2} \sqrt{\frac{\log m}{\min_i n_i}} \right) \\
&\leq \sum_{i=1}^{m} \sum_{l=1}^{p} P \left(\left| \tilde{\boldsymbol{S}}^l{}_i \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) \right| \geq \sqrt{2 c_1^{-1} C_2^{-1} c_x^2} \sqrt{\frac{\log m}{\min_i n_i}} \right) \\
&\leq \sum_{i=1}^{m} \sum_{l=1}^{p} P \left(\left| \tilde{\boldsymbol{S}}_i^l \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) \right| \geq \sqrt{2 c_1^{-1} C_2^{-1} c_x^2} \sqrt{\frac{\log m}{\min_i n_i}} \right) \\
&\leq mp \times \frac{2}{m^2} = \frac{2p}{m}.
\end{aligned}
$$

The result implies that $\max_i \| \tilde{\boldsymbol{S}}_i \left(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0 \right) \| = O_p \left(\sqrt{\frac{\log m}{\min_i n_i}} \right)$. Combined with

(20), we have

$$\begin{aligned}
&\max_i \left\|\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s)\right\| \\
=&\max_i \left\|\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s)-\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)+\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)\right\| \\
\le&\max_i \left\|\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s)-\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)\right\|+\max_i \left\|\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)\right\| \\
=&O_p\left(\sqrt{\frac{\log m}{\min_i n_i}}\right).
\end{aligned}$$

Since when $i,j\in\mathcal{G}_k, \boldsymbol{\beta}_i^*=\boldsymbol{\beta}_j^*$, Γ_1 can be written as

$$\Gamma_1=\sum_{k=1}^{K}\sum_{\{i,j\in\mathcal{G}_k,i<j\}}\frac{\{W_i\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s)-W_j\tilde{\boldsymbol{S}}_j(\boldsymbol{\eta},\boldsymbol{\beta}^s)\}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_j)}{|\mathcal{G}_k|},$$

Then, we have

$$\begin{aligned}
&\left\|\frac{\{W_i\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s)-W_j\tilde{\boldsymbol{S}}_j(\boldsymbol{\eta},\boldsymbol{\beta}^s)\}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_j)}{|\mathcal{G}_k|}\right\| \\
\le&2\max_i W_i\,|\mathcal{G}_k|^{-1}\max_i\left\|\tilde{\boldsymbol{S}}_i(\boldsymbol{\eta},\boldsymbol{\beta}^s)\right\|\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\| \\
=&2\max_i W_i\,|\mathcal{G}_k|^{-1}O_p\left(\sqrt{\frac{\log m}{\min_i n_i}}\right)\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|.
\end{aligned}$$

Based on Condition (C1), we know that all $\tilde{n}_k$'s are in the same order and all n_i's are in the same order. Since K is fixed, then $|\mathcal{G}_k|$ has the same order as m. These imply that $O_p(\sqrt{\frac{\log m}{\min_i n_i}}|\mathcal{G}_k|^{-1})=O_p(\sqrt{\frac{\log m}{m}}n_0^{-1/2})$. Also, combined with Condition (C1), then, we have

$$\Gamma_1\ge-\frac{1}{m}\sum_{k=1}^{K}\sum_{\{i,j\in\mathcal{G}_k,i<j\}}O_p\left(\sqrt{\frac{\log m}{m}}n_0^{-1/2}\right)\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|. \tag{21}$$

Next we consider Γ_2.

$$\begin{aligned}
\Gamma_2=&\lambda\sum_{\{j>i\}}\rho'_\gamma\left(\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|\right)\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|^{-1}\left(\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\right)^{\mathrm{T}}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*) \\
&+\lambda\sum_{\{j<i\}}\rho'_\gamma\left(\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|\right)\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|^{-1}\left(\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\right)^{\mathrm{T}}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*) \\
=&\lambda\sum_{\{j>i\}}\rho'_\gamma\left(\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|\right)\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|^{-1}\left(\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\right)^{\mathrm{T}}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*) \\
&+\lambda\sum_{\{i<j\}}\rho'_\gamma\left(\|\boldsymbol{\beta}_j^s-\boldsymbol{\beta}_i^s\|\right)\|\boldsymbol{\beta}_j^s-\boldsymbol{\beta}_i^s\|^{-1}\left(\boldsymbol{\beta}_j^s-\boldsymbol{\beta}_i^s\right)^{\mathrm{T}}(\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*) \\
=&\lambda\sum_{\{j>i\}}\rho'_\gamma\left(\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|\right)\|\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\|^{-1}\left(\boldsymbol{\beta}_i^s-\boldsymbol{\beta}_j^s\right)^{\mathrm{T}}\left\{(\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*)-(\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*)\right\}.
\end{aligned}$$

When $i, j \in \mathcal{G}_k, \boldsymbol{\beta}_i^* = \boldsymbol{\beta}_j^*$, and $\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s = \alpha(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$. Thus,

$$\begin{aligned}\Gamma_2 = \lambda \sum_{k=1}^{K} \sum_{\{i,j \in \mathcal{G}_k, i<j\}} \rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|\right) \left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|^{-1} \left(\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right)^{\mathrm{T}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) \\ + \lambda \sum_{k<k'} \sum_{\{i \in \mathcal{G}_k, j' \in \mathcal{G}_{k'}\}} \rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|\right) \left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|^{-1} \\ \cdot \left(\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right)^{\mathrm{T}} \left\{(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) - \left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\right)\right\}.\end{aligned}$$

Hence, for $k \neq k', i \in \mathcal{G}_k, j' \in \mathcal{G}_{k'}$,

$$\left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\| \geq \min_{i \in \mathcal{G}_k, j' \in \mathcal{G}_{k'}} \left\|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\right\| - 2 \max_i \left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_i^0\right\| \geq b - 2Mn_0^{-1/2} > a\lambda,$$

and thus $\rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|\right) = 0$ by Condition (C7). Therefore,

$$\begin{aligned}\Gamma_2 &= \lambda \sum_{k=1}^{K} \sum_{\{i,j \in \mathcal{G}_k, i<j\}} \rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|\right) \left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|^{-1} \left(\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right)^{\mathrm{T}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) \\ &= \lambda \sum_{k=1}^{K} \sum_{\{i,j \in \mathcal{G}_k, i<j\}} \rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|\right) \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|,\end{aligned}$$

where the last step follows from $\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s = \alpha(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$.

When $i, j \in \mathcal{G}_k$, we have $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_j^*$ and

$$\max_i \|\boldsymbol{\beta}_i^* - \widehat{\boldsymbol{\beta}}_i^{or}\| = \max_k \|\boldsymbol{\alpha}_k - \widehat{\boldsymbol{\alpha}}_k^{or}\| \leq \max_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\|.$$

Then,

$$\begin{aligned}\max_i \left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\| &\leq 2 \max_i \|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_i^*\| \leq 2 \max_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\| \\ &\leq 2 \left(\max_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| + \max_i \|\boldsymbol{\beta}_i^* - \widehat{\boldsymbol{\beta}}_i^{or}\|\right) \\ &\leq 4 \max_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| \\ &\leq 4t_m.\end{aligned}$$

Hence $\rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_j^s\right\|\right) \geq \rho_\gamma'(4t_m)$ by concavity of $\rho_\gamma(\cdot)$. As a result,

$$\Gamma_2 \geq \sum_{k=1}^{K} \sum_{\{i,j \in \mathcal{G}_k, i<j\}} \lambda \rho_\gamma'(4t_m) \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \tag{22}$$

Based on (21) and (22), we have

$$\begin{aligned}&Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}^*) \\ &\geq \sum_{k=1}^{K} \sum_{\{i,j \in \mathcal{G}_k, i<j\}} \left\{\lambda \rho_\gamma'(4t_m) - O_p\left(\sqrt{\frac{\log m}{m}} n_0^{-1/2}\right)\right\} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|.\end{aligned}$$

As $t_m = o(1), \rho'_\gamma(4t_m) \to 1$ by Condition (C7). Since $\lambda \gg n_0^{-1/2}$, therefore, $Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) > Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$. The result in (ii) is proved.

By combining (i) and (ii), we will have that $Q_\omega(\boldsymbol{\eta}, \boldsymbol{\beta}) > Q_\omega(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ for any $\left(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}} \in \Theta_n \cap \Theta$ and $\left(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}} \neq \left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$. This shows that $\left((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}}\right)^{\mathrm{T}}$ is a strict local minimizer of the objective function for sufficiently large m.

Appendix C: Conditions verification

C.1. Poisson sampling

To demonstrate the rationality of the proposed method, we verify the generality conditions in Section 3 for the case where Poisson sampling is conducted for each domain; see [20] for a similar treatment. In this section, we verify conditions specifically for Poisson sampling, including (C3)–(C5). We now assume the following specific conditions.

(A1) The parameter space Θ is compact, containing $\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)$ as an interior point. For $(\boldsymbol{\eta}, \boldsymbol{\alpha}) \subset \Theta$, $L_{ih}^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$ is convex and twice continuously differentiable with respect to $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ for $i = 1, \ldots, m$ and $h = 1, \ldots, N_i$, where $m \to \infty$. Besides, $E[L_{ih}^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})]^2 < C_0$ for a constant C_0. $\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)$ uniquely minimizes $E\{L_{ih}^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})\}$.

(A2) The pair $(\pi_{ih}, \boldsymbol{x}_{ih}, \boldsymbol{z}_{ih} y_{ih})$ is independent of $(\pi_{jk}, \boldsymbol{x}_{jk}, \boldsymbol{z}_{jk}, y_{jk})$ for $i \neq j$ or $h \neq k$.

(A3) For all $i = 1, \ldots, m$, there exist two constants $0 < C_1 < C_2 < \infty$ such that $C_1 < N_i \bar{n}_i^{-1} \pi_{ih} < C_2$ for $h = 1, \ldots, N_i$ almost surely with respect to the super-population model, where $\bar{n}_i = \sum_{h=1}^{N_i} \pi_{ih}$ is the expected sample size in unit i satisfying $\bar{n}_i = o_p(N_i)$, $\max\{\bar{n}_i : i = 1, \ldots, m\} / \min\{\bar{n}_i : i = 1, \ldots, m\} < C_3$, and C_3 is a positive constant. In addition,

$$\max\{N_i : i = 1, \ldots, m\} / \min\{N_i : i = 1, \ldots, m\} < C_3,$$
$$\max\{\tilde{n}_k : k = 1, \ldots, K\} / \min\{\tilde{n}_k : k = 1, \ldots, K\} < C_3,$$

where $\tilde{n}_k$ is the sample size of the kth cluster.

(A4) There exists a compact set $\mathcal{B} \subset \Theta$ containing $(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0)$ as an interior point, such that $\sup_{(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}} E \| \hat{\boldsymbol{S}}_i(\boldsymbol{\eta}, \boldsymbol{\alpha}) \|^4 < \infty$, where $\|\boldsymbol{x}\|$ is the Euclidean norm of a vector $\boldsymbol{x}$.

(A5) Given any $\boldsymbol{a} \in \mathbb{R}^{q+Kp}$ satisfying $\|\boldsymbol{a}\| = 1, V\{S_{\boldsymbol{a},ih}(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0)\} > 0$, and there exists a positive definite matrix $\boldsymbol{\Sigma}_a$ such that

$$\sum_{i=1}^{m} W_i^2 \bar{n}_i N_i^{-2} \sum_{h=1}^{N_i} \pi_{ih}^{-1} \left(1, S_{\boldsymbol{a},ih}\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)\right)^{\mathrm{T}} \left(1, S_{\boldsymbol{a},ih}\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)\right) \to \boldsymbol{\Sigma}_a$$

in probability with respect to the super-population model, where $V\left\{S_{a,ih}\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)\right\}$ is the variance of $S_{\boldsymbol{a},ih}(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0)$ with respect to the super-population model, and $S_{\boldsymbol{a},ih}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{S}_{ih}(\boldsymbol{\eta}, \boldsymbol{\alpha})$.

(A6) For $(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}$, there exists a positive definitive and non-stochastic matrix $\boldsymbol{\Sigma}_S(\boldsymbol{\eta}, \boldsymbol{\alpha})$ such that $\bar{n}_0 \boldsymbol{V}\left\{\hat{\boldsymbol{S}}_w(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\right\} \to \Sigma_S(\boldsymbol{\eta}, \boldsymbol{\alpha})$ in probability with respect to the super-population model uniformly over $\mathcal{B}$, where $\bar{n}_0 = \sum_{i=1}^m \bar{n}_i$, $\boldsymbol{x}^{\otimes 2} = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$ for a vector $\boldsymbol{x}$, and

$$\boldsymbol{V}\left\{\hat{\boldsymbol{S}}_w(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\right\} = \sum_{i=1}^m W_i^2 N_i^{-2} \sum_{h=1}^{N_i} \pi_{ih}^{-1} (1 - \pi_{ih}) \boldsymbol{S}_{ih}(\boldsymbol{\eta}, \boldsymbol{\alpha})^{\otimes 2}$$

is the design variance of $\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}, \boldsymbol{\alpha})$.

(A7) $\sup_{(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}} E\left\{\|\boldsymbol{I}_i(\boldsymbol{\eta}, \boldsymbol{\alpha})\|^2\right\} < \infty$, and $\mathcal{I}\left(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0\right)$ is invertible, where $\boldsymbol{\mathcal{I}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$ is the limit of $\sum_{i=1}^m W_i \boldsymbol{\mathcal{I}}_i(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \sum_{i=1}^m W_i E\{\boldsymbol{I}_i(\boldsymbol{\eta}, \boldsymbol{\alpha})\}$ as $m \to \infty$.

(A8) The number of clusters K, and the dimensions of covariates p and q are fixed. Conditions (C7)–(C9) hold.

Condition (A1) is similar to Condition (C2), and its second part is used to validate Condition (C3). Condition (A2) guarantees the central limit theorem by assuming independence between different elements in the finite population, and the independence is with respect to the model. Condition (A3) regulates inclusion probabilities as well as population and sample sizes, and it is required to show the convergence rate in Condition (C4). Condition (A4) is used to show the central limit theorem for the estimating function. Conditions (A5)–(A6) assume the probability limit for the scaled estimating function with respect to Poisson sampling conditional on the finite population, and it is also used to verify the central limit theorem for the estimating function. Condition (A7) is required to show the convergence of the information matrix, and Condition (A8) contains regularity conditions on the penalty term as well as others for the model.

C.1.1. Lemma 1

Under Poisson sampling and Conditions (A1)–(A3), Condition (C3) holds.

Proof.

$$\begin{aligned}
\boldsymbol{V}\{L_\omega(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\} &= \boldsymbol{V}\left\{\sum_{i=1}^m W_i N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih} \pi_{ih}^{-1} L_{ih}(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\right\} \\
&= \sum_{i=1}^m W_i^2 N_i^{-2} \sum_{h=1}^{N_i} \pi_{ih}^{-1} (1 - \pi_{ih}) L_{ih}^2(\boldsymbol{\eta}, \boldsymbol{\alpha}) \\
&\le \sum_{i=1}^m W_i^2 (C_1 \bar{n}_i)^{-1} N_i^{-1} \sum_{h=1}^{N_i} L_{ih}^2(\boldsymbol{\eta}, \boldsymbol{\alpha}) \\
&= \sum_{i=1}^m W_i^2 O_p(\bar{n}_i^{-1}),
\end{aligned}$$

where the third inequality is based on Condition (A3), and the last equality holds by Conditions (A1)–(A2) and weak law of large numbers. Since $\sum_{i=1}^{m} W_i = 1$ and $W_i \asymp m^{-1}$ by Condition (A3), $V\{L_\omega(\boldsymbol{\eta}, \boldsymbol{\alpha}) | \mathcal{F}\} \to 0$ in probability with respect to the super-population model as $n_i \to \infty$ under Poisson sampling.

This completes the proof of Lemma 1. □

C.1.2. Lemma 2

Under Poisson sampling and Conditions(A2)–(A4), we have

$$\sup_{(\boldsymbol{\eta},\boldsymbol{\alpha})\in\mathcal{B}} \hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} \asymp \bar{n}_0^{-1}$$

in probability with respect to the super-population model and the sampling mechanism, where

$$\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} = \sum_{i=1}^{m} W_i^2 N_i^{-2} \sum_{h=1}^{N_i} \delta_{ih}{\pi_{ih}}^{-2}(1-\pi_{ih})\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})\}^2,$$

where $\boldsymbol{a} \in \mathbb{R}^{q+Kp}$ satisfying $\|\boldsymbol{a}\| = 1$.

Proof. First we consider $E\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\}\right]$ with respect to the super-population model and the sampling mechanism.

$$\begin{aligned}
&E\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\}\right] \\
&= E\left(E\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} \mid \mathcal{F}\right]\right) \\
&= E\left[\sum_{i=1}^{m} W_i^2 N_i^{-2} \sum_{h=1}^{N_i} \delta_{ih}{\pi_{ih}}^{-2}(1-\pi_{ih})\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^2 \mid \mathcal{F}\right] \\
&= \sum_{i=1}^{m} W_i^2 N_i^{-2} \sum_{h=1}^{N_i} {\pi_{ih}}^{-1}(1-\pi_{ih})\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^2.
\end{aligned}$$

By Conditions (A3)–(A4), we have

$$E\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\}\right] \asymp \bar{n}_0^{-1}$$

uniformly over $\mathcal{B}$, where the expectation is taken with respect to super-population model.

Next, consider the variance of $\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\}$ with respect to the super-population model and the sampling mechanism:

$$\begin{aligned}
&V\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\}\right] \\
&= E\left(V\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} \mid \mathcal{F}\right]\right) \\
&+ V\left(E\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} \mid \mathcal{F}\right]\right).
\end{aligned}$$

Under Poisson sampling, we have

$$
\begin{aligned}
&E\left(V\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_{\omega}(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} \mid \mathcal{F}\right]\right)\\
&= E\left[V\left\{\sum_{i=1}^{m} W_i^2 N_i^{-2} \sum_{h=1}^{N_i} \delta_{ih}{\pi_{ih}}^{-2}(1-\pi_{ih})\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^2 \mid \mathcal{F}\right\}\right]\\
&= \sum_{i=1}^{m} E\left[V\left\{W_i^2 N_i^{-2} \sum_{h=1}^{N_i} \delta_{ih}{\pi_{ih}}^{-2}(1-\pi_{ih})\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^2 \mid \mathcal{F}\right\}\right]\\
&= \sum_{i=1}^{m} W_i^4 N_i^{-4} \sum_{h=1}^{N_i} \pi_{ih}(1-\pi_{ih}){\pi_{ih}}^{-4}(1-\pi_{ih})^2 E\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^4\right\}\\
&= \sum_{i=1}^{m} W_i^4 N_i^{-4} \sum_{h=1}^{N_i} {\pi_{ih}}^{-3}(1-\pi_{ih})^3 E\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^4\right\}\\
&\le \sum_{i=1}^{m} W_i^4 (C_1\bar{n}_i)^{-3} N_i^{-1} \sum_{h=1}^{N_i} E\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^4\right\}\\
&\asymp \bar{n}_0^{-3},
\end{aligned}
$$

and

$$
\begin{aligned}
&V\left(E\left[\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_{\omega}(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} \mid \mathcal{F}\right]\right)\\
&= V\left\{\sum_{i=1}^{m} W_i^2 N_i^{-2} \sum_{h=1}^{N_i} {\pi_{ih}}^{-1}(1-\pi_{ih})\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^2\right\}\\
&= \sum_{i=1}^{m} W_i^4 N_i^{-4} \sum_{h=1}^{N_i} V\left\{{\pi_{ih}}^{-1}(1-\pi_{ih})\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})^2\right\}\\
&\le \sum_{i=1}^{m} W_i^4 N_i^{-4} \sum_{h=1}^{N_i} E\left[{\pi_{ih}}^{-2}(1-\pi_{ih})^2\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})\right\}^4\right]\\
&\le \sum_{i=1}^{m} W_i^4 (C_1\bar{n}_i)^{-2} N_i^{-2} \sum_{h=1}^{N_i} E\left[(1-\pi_{ih})^2\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})\right\}^4\right]\\
&= o(1)
\end{aligned}
$$

uniformly over $\mathcal{B}$ by Condition (A4). This completes the proof of Lemma 2. □

C.1.3. Lemma 3

Under Poisson sampling and Conditions (A2)–(A6), Condition (C4) holds.

Proof. Given every $\boldsymbol{a} \in \mathbb{R}^{q+Kp}$ satisfying $\|\boldsymbol{a}\| = 1$, by Conditions (A2)–(A5), Theorem 1.3.5 of [13] shows

$$
\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_{\omega}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) \mid \mathcal{F}\right\}^{-1/2} \boldsymbol{a}^{\mathrm{T}}\left\{\hat{\boldsymbol{S}}_{\omega}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) - \boldsymbol{S}_{\omega}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\right\} \mid\!\rightarrow N(0,1)
$$

in distribution with respect to the sampling mechanism conditional on finite population in probability, where $\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) = \sum_{i=1}^m W_i N_i^{-1} \sum_{h=1}^{N_i} \delta_{ih}\pi_{ih}^{-1}\boldsymbol{S}_{ih}(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)$, $\boldsymbol{S}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) = E\{\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) \mid \mathcal{F}\}$.

Since the finite population is a random sample generated from a super-population model satisfying Conditions (A4)–(A5), by the Chebychev's inequality [2, Corollary 3.1.3], we have

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) = \boldsymbol{a}^{\mathrm{T}}\left[\boldsymbol{S}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) - E\left\{\boldsymbol{S}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\right\}\right] = o_p(\bar{n}_0^{-1/2})$$

with respect to the super-population model, where the second equality holds since $E\left\{\boldsymbol{S}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0)\right\} = 0$, and the last equality holds by Condition (A3).

Thus, by Condition (A6), Lemma 3 and Theorem 5.1 of [39] we have

$$\hat{V}\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) \mid \mathcal{F}\right\}^{-1/2}\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{S}}_\omega(\boldsymbol{\eta}^0,\boldsymbol{\alpha}^0) \to N(0,1)$$

in distribution with respect to the super-population model and the sampling mechanism, which validates Condition (C4) by the Cramér-Wold device. □

C.1.4. Lemma 4

Under Poisson sampling, Conditions (A2)–(A3) and Condition (A7), Condition (C5) holds.

Proof. Consider

$$E\left\{\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\} = \sum_{i=1}^m W_i N_i^{-1}\sum_{h=1}^{N_i}\boldsymbol{I}_{ih}(\boldsymbol{\eta},\boldsymbol{\alpha})$$
$$E\left\{\sum_{i=1}^m \boldsymbol{I}_i(\boldsymbol{\eta},\boldsymbol{\alpha})\right\} = \boldsymbol{\mathcal{I}}(\boldsymbol{\eta},\boldsymbol{\alpha}).$$

Then we have

$$\begin{aligned} E\left\{\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})\right\} &= E\left[E\left\{\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha}) \mid \mathcal{F}\right\}\right] \\ &= \boldsymbol{\mathcal{I}}(\boldsymbol{\eta},\boldsymbol{\alpha}) \end{aligned}$$

with respect to the super-population model and the sampling mechanism. For $\boldsymbol{a} \in \mathbb{R}^{q+Kp}$ satisfying $\|\boldsymbol{a}\| = 1$, consider the variance of $\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})\boldsymbol{a}$ with respect to the super-population model and the sampling mechanism:

$$V\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})\boldsymbol{a}\right\} = E\left[V\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})\boldsymbol{a} \mid \mathcal{F}\right\}\right] + V\left[E\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})\boldsymbol{a} \mid \mathcal{F}\right\}\right].$$

First consider

$$\begin{aligned} V\left\{\boldsymbol{a}^{\mathrm{T}}\hat{\boldsymbol{I}}_\omega(\boldsymbol{\eta},\boldsymbol{\alpha})\boldsymbol{a} \mid \mathcal{F}\right\} &= \sum_{i=1}^m W_i^2 N_i^{-2}\sum_{h=1}^{N_i}\frac{1-\pi_{ih}}{\pi_{ih}}Z_{\boldsymbol{a},ih}^2(\boldsymbol{\eta},\boldsymbol{\alpha}) \\ &\le \sum_{i=1}^m W_i^2(C_1 n_i)^{-1}\frac{1}{N_i}\sum_{h=1}^{N_i}Z_{\boldsymbol{a},ih}^2(\boldsymbol{\eta},\boldsymbol{\alpha}), \end{aligned}$$

where $Z_{\boldsymbol{a},ih} = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{I}_{ih}(\boldsymbol{\eta}, \boldsymbol{\alpha})\, \boldsymbol{a}$, the second inequality holds by Conditions (A2)–(A3). Thus we have

$$E\left[V\left\{\boldsymbol{a}^{\mathrm{T}} \hat{\boldsymbol{I}}_{\omega}(\boldsymbol{\eta}, \boldsymbol{\alpha})\boldsymbol{a} \mid \mathcal{F}\right\}\right] = O(\bar{n}_0^{-1})$$

in probability with respect to the super-population model uniformly over $\mathcal{B}$ by Condition (A7).

Next consider

$$\begin{aligned} V\left\{\sum_{i=1}^{m} W_i N_i^{-1} \sum_{h=1}^{N_i} \boldsymbol{a}^{\mathrm{T}} \boldsymbol{I}_{ih}(\boldsymbol{\eta}, \boldsymbol{\alpha})\, \boldsymbol{a}\right\} &= \sum_{i=1}^{m} W_i^2 N_i^{-1} V\left\{\boldsymbol{a}^{\mathrm{T}} \boldsymbol{I}_{ih}(\boldsymbol{\eta}, \boldsymbol{\alpha})\, \boldsymbol{a}\right\} \\ &\le \sum_{i=1}^{m} W_i^2 N_i^{-1} E\left\{|\boldsymbol{a}^{\mathrm{T}} \boldsymbol{I}_{ih}(\boldsymbol{\eta}, \boldsymbol{\alpha})\, \boldsymbol{a}|^2\right\} \\ &= o(1) \end{aligned}$$

in probability with respect to the super-population model uniformly for $(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}$ by Condition (A7), where the last equality holds by Condition (A3).

By Markov's inequality, we can show that

$$\hat{\boldsymbol{I}}_{\omega}(\boldsymbol{\eta}, \boldsymbol{\alpha}) \to \boldsymbol{\mathcal{I}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$$

in probability with respect to the super-population model and the sampling mechanism uniformly for $(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}$, which completes the proof of Lemma 4. □

C.2. Stratified multistage cluster sampling

In this section, we consider a popular stratified multistage cluster sampling as in Example 1 of [20]; see Example 1 of [40] for a similar treatment. We assume that elements within each stratum share common regression coefficients, and we estimate the cluster structure by regression coefficients on the stratum level.

Assume that there exist m strata in the finite population $\mathcal{F}$. Let N_i be the number of clusters in the ith stratum, M_{ih} be the number of elements in the hth cluster of the ith stratum, and $M_i = \sum_{h=1}^{N_i} M_{ih}$ be the size of the ith stratum, where N_i is the number of clusters in the ith stratum. Let n_i be the number of clusters selected from the ith stratum with selection probability p_{ih} satisfying $\sum_{h=1}^{N_i} p_{ih} = 1$ for $i = 1, \ldots, m$, where p_{ih} is proportional to a size measure, for example, $p_{ih} \propto M_{ih}$ for $h = 1, \ldots, N_i$. Within each selected cluster, denote $\hat{L}_{ih}(\boldsymbol{\eta}, \boldsymbol{\beta})$ and $\hat{S}_{w,ih}(\boldsymbol{\eta}, \boldsymbol{\beta})$ as design-unbiased estimators of the associated cluster means of the objective function and score function under sampling at the second and subsequent stages. Then, the probability-weighted score function is $\hat{S}_w(\boldsymbol{\eta}, \boldsymbol{\beta}) = \sum_{i=1}^{m} W_i \hat{S}_{w,i}(\boldsymbol{\eta}, \boldsymbol{\beta})$, where $W_i \propto M_i$ and $\sum_{i=1}^{m} W_i = 1$, $\hat{S}_{w,i}(\boldsymbol{\eta}, \boldsymbol{\beta}) = n_i^{-1} \sum_{h=1}^{n_i} \tilde{S}_{w,ih}(\boldsymbol{\eta}, \boldsymbol{\beta})$, $\tilde{S}_{w,ih}(\boldsymbol{\eta}, \boldsymbol{\beta}) = d_{a(ih)} p^{-1} a(ih) \hat{S}_{w,a(ih)}(\boldsymbol{\eta}, \boldsymbol{\beta})$, $d_{a(ih)} = M_{a(ih)}/M_i$, $a(ih)$ is the index of the hth selected cluster. We implicitly assume that $m \to \infty$, and the number of stages are fixed for the sampling design. For this sampling design, we make the following assumptions.

(B1) There exists constants $0 < C_1 < C_2 < \infty$, such that $C_1 < M_i p_{ih} < C_2$ almost surely with respect to the model for $i = 1, \ldots, m$ and $h = 1, \ldots, N_i$.

(B2) $\max_{1 \le i \le m} n_i = O(1)$.

(B3) There exists $\delta > 0$ such that

$$\sum_{i=1}^{m} W_i E\left\{\|\tilde{S}_{w,a(i1)}(\boldsymbol{\eta}, \boldsymbol{\alpha}) - S_i(\boldsymbol{\eta}, \boldsymbol{\alpha})\|^{2+\delta} \mid \mathcal{F}\right\} = O_p(1)$$

with respect to the model uniformly for $(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}$, where $\mathcal{B}$ is a compact set containing $(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0)$ as an interior point, and $S_i(\boldsymbol{\eta}, \boldsymbol{\alpha}) = E\{\tilde{S}_{w,a(i1)}(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\}$ is the average score function associated with the ith stratum.

(B4) $\max_{1 \le i \le m} \sup_{(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}} E\left\{\|\hat{I}_{w,a(i1)}(\boldsymbol{\eta}, \boldsymbol{\alpha})\|^2 \mid \mathcal{F}\right\} = O_p(1)$ with respect to the model, where $\hat{I}_{w,ih}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \partial \hat{S}_{w,ih}(\boldsymbol{\eta}, \boldsymbol{\alpha}) / \partial(\boldsymbol{\eta}, \boldsymbol{\alpha})$.

(B5) There exists a non-stochastic function $\mathcal{I}(\boldsymbol{\eta}, \boldsymbol{\alpha})$ such that

$$\sup_{(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}} \|I(\boldsymbol{\eta}, \boldsymbol{\alpha}) - \mathcal{I}(\boldsymbol{\eta}, \boldsymbol{\alpha})\| \to 0$$

in probability with respect to the model, where

$$I(\boldsymbol{\eta}, \boldsymbol{\alpha}) = E\{\hat{I}_w(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\},\ \hat{I}_w(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} W_i \hat{I}_{w,i}(\boldsymbol{\eta}, \boldsymbol{\beta})$$

$$\hat{I}_{w,i}(\boldsymbol{\eta}, \boldsymbol{\beta}) = n_i^{-1} \sum_{h=1}^{n_i} \tilde{I}_{w,ih}(\boldsymbol{\eta}, \boldsymbol{\beta}),\ \tilde{I}_{w,ih}(\boldsymbol{\eta}, \boldsymbol{\beta}) = d_{a(ih)} p^{-1} a(ih) \hat{I}_{w,a(ih)}(\boldsymbol{\eta}, \boldsymbol{\beta}),$$

and $\mathcal{I}(\boldsymbol{\eta}^0, \boldsymbol{\alpha}^0)$ is invertible.

(B6) $\sup_{(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}} \|n_0 \sum_{i=1}^{m} W_h^2 V\{\tilde{S}_{w,a(i1)}(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\} - \Sigma_S(\boldsymbol{\eta}, \boldsymbol{\alpha})\| \to 0$ in probability with respect to the model, where $\Sigma_S(\boldsymbol{\eta}, \boldsymbol{\alpha})$ is a non-stochastic and positive definitive variance matrix for $(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}$, and $n_0 = \sum_{i=1}^{m} n_i$.

(B7) $\sup_{(\boldsymbol{\eta}, \boldsymbol{\alpha}) \in \mathcal{B}} V\{S(\boldsymbol{\eta}, \boldsymbol{\alpha})\} = o(n_0^{-1})$ with respect to the model, where $S(\boldsymbol{\eta}, \boldsymbol{\alpha}) = E\{\hat{S}_w(\boldsymbol{\eta}, \boldsymbol{\alpha}) \mid \mathcal{F}\}$.

(B8) Conditions (C1)–(C3), (C6)–(C9) hold.

Conditions (B1)–(B2) are commonly assumed for stratified multistage cluster sampling [22]. That is, we consider a sampling design, where the number of selected clusters is bounded but the number of strata diverges. Condition (B3) is used to show the central limit theorem in Condition (C4), and Conditions (B4)–(B5) are mainly used to show the convergence result in Condition (C5). Conditions (B6)–(B7) shows the convergence of the sample variance conditional on the finite population, and the variance $V\{S(\boldsymbol{\eta}, \boldsymbol{\beta})\}$ is asymptotically negligible, which is required by the second part of Condition (C7). Condition (B8) regulates the objective function as well as the penalty, and the variance of the probability-weighted objective function is also assumed to converge in probability. Specifically, by Condition (B8), we can show $\max_{1 \le i \le m} W_i = O_p(m^{-1})$, which is essential to establish convergence rates under stratified multi-stage cluster sampling.

Different from Poisson sampling, the sampling indicators may be correlated due to sampling at the second and subsequent stages. Under the above conditions, Conditions (C4)–(C5) can be verified in exactly the same manner as Lemmas S5–S6 of [20], so the detailed proofs are omitted.

Appendix D: More simulation results

In this section, we present the results of our proposed estimator when the data sets are simulated from settings that are different from the assumed models.

D.1. Example 1

In this example, we generate finite populations from linear mixed models, that is

$$y_{ih} = u_i + \beta_{0,i} + \beta_{1,i} x_{ih} + \epsilon_{ih}, \tag{23}$$

where u_i is a random effect with $u_i \overset{iid}{\sim} N(0, \epsilon_u^2)$. This model used a lot in small area estimation in survey sampling, see examples in [36]. Other settings are the same as those in Section 4.1. Note that, [24] did a similar comparison without employing a survey sampling setup. In their comparison, they simulated datasets from a linear mixed model with common regression coefficients and explored the performance of the clustered intercept model for capturing the random effects. In our model, the values of $\beta'_{0,i}s$ can contain the information of random effects. We explore and compare the results for different values of $\epsilon_u = 0.1, 0.2, 0.4, 0.5$ for CC and PCC. Table 6 shows the estimated $\hat{K}$ and ARI for different setups. It can be seen that when ϵ_u and n_i are larger, the estimated $\hat{K}$ is much larger than the true number of clusters. This is because that the estimated values of $\beta'_{0,i}s$ have the information of u_i. In Figure 3, we calculate RMSEs defined as $\text{RMSE}_2 = \frac{1}{m}\sum_{i=1}^{m}(\beta_{0,i} + u_i - \hat{\beta}_{0,i})^2$, which measures the accuracy of using fixed clustered effects to estimate the fixed intercept and random effects together. From these figures, we observe that when $n_i = 30, 50$ and $\epsilon_u = 0.4, 0.5$, PCC has much smaller RMSE for estimating $\beta_{i,0} + u_i$. These are consistent with the estimated number of clusters in Table 6, which implies that the $\hat{\beta}_{i,0}$ has information of u_i. CC can capture this information only when $\epsilon_u = 0.5$ and $n_i = 50$.

D.2. Example 2

In this example, we generate a finite population from the following model.

$$y_{ih} = \beta_{0,i} + \beta_{1,i} x_{ih} + b \log(|x_{ih}|) + \epsilon_{ih}. \tag{24}$$

Table 7 summarizes the results of average $\hat{K}$ and average ARI based on 1 000 simulations when $b = 0.3$ and 1. When $b = 0.3$, the fitted model is not really different from the true model. PCC can identify the cluster structure well and

TABLE 6
Summary of average $\hat{K}$ and average ARI of CC and PCC for the linear mixed model based on 1 000 simulations, along with standard deviations in the parentheses.

		$\hat{K}$		ARI	
		CC	PCC	CC	PCC
$\epsilon_u = 0.1$	$n_i = 10$	5.01(1.089)	3.41(0.578)	0.94(0.042)	0.99(0.019)
	$n_i = 30$	4.94(0.739)	3.05(0.232)	0.95(0.042)	1(0.01)
	$n_i = 50$	4.96(0.617)	3.04(0.206)	0.96(0.036)	1(0.005)
$\epsilon_u = 0.2$	$n_i = 10$	5.00(1.085)	3.41(0.587)	0.94(0.04)	0.99(0.019)
	$n_i = 30$	4.85(0.66)	3.04(0.204)	0.96(0.033)	1(0.006)
	$n_i = 50$	4.72(0.542)	3.21(0.976)	0.97(0.021)	0.99(0.061)
$\epsilon_u = 0.4$	$n_i = 10$	5.03(1.088)	3.54(0.932)	0.93(0.042)	0.98(0.049)
	$n_i = 30$	4.70(0.527)	10.4(1.288)	0.97(0.017)	0.56(0.051)
	$n_i = 50$	5.30(2.468)	11.06(1.668)	0.92(0.142)	0.5(0.064)
$\epsilon_u = 0.5$	$n_i = 10$	5.14(1.134)	5.47(3.153)	0.92(0.049)	0.86(0.18)
	$n_i = 30$	5.00(1.299)	10.82(1.611)	0.95(0.076)	0.51(0.064)
	$n_i = 50$	12.18(1.511)	12.61(1.585)	0.56(0.069)	0.43(0.056)

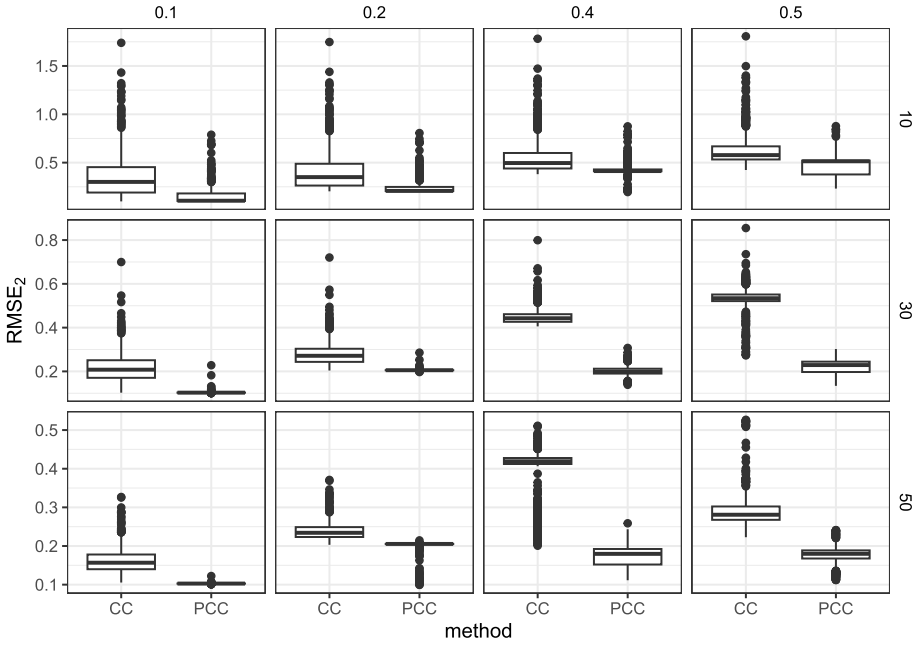

FIG 3. *$RMSE_2$ based on 1 000 simulations with expected sample size $n_i = 10$ (top), $n_i = 30$ (middle) and $n_i = 50$ (bottom).*

performs better than CC. When $b = 1$, both approaches have worse performance compared to the case in Section 4.1 when the sample size is small ($n_i = 10$). The performances are similar to the case in Section 4.1 when the sample size is large.

TABLE 7
Summary of average $\hat{K}$ and average ARI of CC and PCC for model (24) based on 1 000 simulations, along with standard deviations in the parentheses.

		$\hat{K}$		ARI	
		CC	PCC	CC	PCC
$b = 0.3$	$n_i = 10$	4.86(1.034)	3.41(0.565)	0.94(0.039)	0.98(0.024)
	$n_i = 30$	4.86(0.771)	3.06(0.234)	0.95(0.043)	1(0.015)
	$n_i = 50$	4.89(0.637)	3.05(0.233)	0.96(0.037)	1(0.013)
$b = 1$	$n_i = 10$	5.83(1.379)	4.87(0.975)	0.84(0.096)	0.86(0.065)
	$n_i = 30$	4.51(0.590)	3.17(0.378)	0.96(0.022)	0.99(0.024)
	$n_i = 50$	3.98(0.333)	3.05(0.221)	0.98(0.011)	1(0.019)

D.3. Example 3

In this example, we generate a finite population from the following model. In this example, the cluster structure is applied for a nonlinear term of x_{ih}.

$$y_{ih} = \beta_{0,i} + \beta_{1,i}(x_{ih} + \log(|x_{ih}|) + \epsilon_{ih}. \tag{25}$$

Table 8 summarizes the results of average $\hat{K}$ and average ARI. We observe that CC has a very small ARI when $n_i = 10$, which indicates that it cannot recover the cluster structure. Our proposed PCC tends to estimate the number of clusters larger than the true value, but the ARI is still relatively large.

TABLE 8
Summary of average $\hat{K}$ and average ARI of CC and PCC for model (25) based on 1 000 simulations, along with standard deviations in the parentheses.

	$\hat{K}$		ARI	
	CC	PCC	CC	PCC
$n_i = 10$	3.98(2.264)	4.73(1.054)	0.20(0.265)	0.91(0.044)
$n_i = 30$	5.54(0.787)	3.79(0.694)	0.92(0.042)	0.97(0.021)
$n_i = 50$	4.86(0.451)	4.11(0.687)	0.96(0.015)	0.96(0.019)

Acknowledgments

We would like to thank the editor, the AE, and two anonymous reviewers for constructive comments and suggestions.

Funding

This research is partially supported by National Key R&D Program of China 2022YFA1003800, NSF SES 2316353, Open Research Fund of Key Laboratory of Analytical Mathematics and Applications (Fujian Normal University), Ministry of Education, P. R. China, Humanities and Social Sciences Foundation of the Ministry of Education of China Grant (No. 23YJA910005), NSSFC (No. 23CMZ005), NSFC (No.: 72033002, 12231011, 71988101, 12471265). Wang, Z. and Zhong, W. also thank the supports of Fujian Key Lab of Statistics, Fujian Key lab of Digital Finance.

References

[1] Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.

[2] Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer, New York. MR2247694

[3] Azka Ubaidillah, A. K. Khairil Anwar Notodiputro and Mangku, I. W. (2019). Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia. *Journal of Applied Statistics* **46** 2845–2861. MR4007698

[4] Berg, E. J. and Fuller, W. A. (2014). Small area prediction of proportions with applications to the Canadian Labour Force Survey. *Journal of Survey Statistics and Methodology* **2** 227–256.

[5] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.

[6] Chen, K., Huang, R., Chan, N. H. and Yau, C. Y. (2019). Subgroup analysis of zero-inflated Poisson regression model with applications to insurance data. *Insurance: Mathematics and Economics* **86** 8–18. MR3914513

[7] Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* 613–627. MR1769758

[8] Dumitrescu, L., Qian, W. and Rao, J. (2021). Variable selection for longitudinal survey data. *arXiv preprint arXiv:2105.00504*. MR4331759

[9] Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D. and Pérez, A. (2020). Small area estimation of proportions under area-level compositional mixed models. *Test* **29** 793–818. MR4140784

[10] Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D. and Pérez, A. (2022). Empirical best prediction of small area bivariate parameters. *Scandinavian Journal of Statistics* **49** 1699–1727. MR4544817

[11] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[12] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57** 5467–5484. MR2849368

[13] Fuller, W. A. (2011). *Sampling Statistics*. Wiley, New Jersey.

[14] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663–685. MR0053460

[15] Hu, X., Huang, J., Liu, L., Sun, D. and Zhao, X. (2021). Subgroup analysis in the heterogeneous Cox model. *Statistics in Medicine* **40** 739–757. MR4198442

[16] Innocent Ngaruye, D. v. R. Joseph Nzabanita and Singull, M. (2017). Small area estimation under a multivariate linear model for repeated measures data. *Communications in Statistics – Theory and Methods* **46** 10835–10850. MR3740840

[17] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics* **40** 633–643. MR0238419

[18] Jiang, J. and Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics* **53** 217–243. MR1841133

[19] Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test* **15** 1–96. MR2252522

[20] Kim, J. K., Rao, J. N. K. and Wang, Z. (2023). Hypotheses testing from complex survey data using bootstrap weights: a unified approach. *Journal of the American Statistical Association* **Accepted** 1–11. MR4765983

[21] Kim, J. K., Wang, Z., Zhu, Z. and Cruze, N. B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level model. *Journal of Agricultural, Biological and Environmental Statistics* **23** 175–189. MR3805267

[22] Krewski, D. and Rao, J. N. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* 1010–1019. MR0628756

[23] Lahiri, P. and Salvati, N. (2023). A nested error regression model with high-dimensional parameter for small area estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85** 212–239. MR4726966

[24] Liu, L., Gordon, M., Miller, J. P., Kass, M., Lin, L., Ma, S. and Liu, L. (2021). Capturing heterogeneity in repeated measures data by fusion penalty. *Statistics in Medicine* **40** 1901–1916. MR4229828

[25] Liu, M., Yang, J., Liu, Y., Jia, B., Chen, Y.-F., Sun, L. and Ma, S. (2022). A fusion learning method to subgroup analysis of Alzheimer's disease. *Journal of Applied Statistics* 1–23. MR4593807

[26] Lohr, S. L. and Liu, J. (1994). A comparison of weighted and unweighted analyses in the National Crime Victimization Survey. *Journal of Quantitative Criminology* **10** 343–360.

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

[28] Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* **112** 410–423. MR3646581

[29] Ma, S., Huang, J., Zhang, Z. and Liu, M. (2020). Exploration of heterogeneous treatment effects via concave fusion. *International Journal of Biostatistics* **16**.

[30] Marhuenda, Y., Molina, I., Morales, D. and Rao, J. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society Series A: Statistics in Society* **180** 1111–1136. MR3723789

[31] Molina, I. and Rao, J. N. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics* **38** 369–385. MR2730115
[32] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4** 2111–2245. MR1315971
[33] Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review* **61** 317–337.
[34] Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association* **102** 1427–1439. MR2412558
[35] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66** 846–850.
[36] Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken. MR3380626
[37] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press. MR0274683
[38] Rojas-Perilla, N., Pannier, S., Schmid, T. and Tzavidis, N. (2019). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society Series A: Statistics in Society* **183** 121–148. MR4049657
[39] Rubin-Bleuer, S. and Kratina, I. S. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics* **33** 2789–2810. MR2253102
[40] Rubin-Bleuer, S. and Kratina, I. S. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics* **33** 2789–2810. MR2253102
[41] Sun, H., Berg, E. and Zhu, Z. (2022). Bivariate small-area estimation for binary and gaussian variables based on a conditionally specified model. *Biometrics* **78** 1555–1565. MR4534378
[42] Sun, H., Berg, E. and Zhu, Z. (2024). Multivariate small-area estimation for mixed-type response variables with item nonresponse. *Journal of Survey Statistics and Methodology* **12** 320–342.
[43] Tang, L. and Song, P. X. (2016). Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research* **17** 3915–3937. MR3543519
[44] Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008
[45] Wang, L., Wang, S. and Wang, G. (2014). Variable selection and estimation for longitudinal survey data. *Journal of Multivariate Analysis* **130** 409–424. MR3229546
[46] Wang, X. (2024). Clustering of longitudinal curves via a penalized method and EM algorithm. *Computational Statistics* **39** 1485–1512. MR4730670
[47] Wang, X., Zhang, X. and Zhu, Z. (2023). Clustered coefficient regression models for Poisson process with an application to seasonal warranty claim data. *Technometrics* **65** 514–523. MR4662685
[48] Wang, X. and Zhu, Z. (2019). Small area estimation with subgroup anal-

ysis. *Statistical Theory and Related Fields* **3** 129–135. MR4028311

[49] Wang, X., Zhu, Z. and Zhang, H. H. (2023). Spatial heterogeneity automatic detection and estimation. *Computational Statistics & Data Analysis* **180**. MR4519305

[50] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942. MR2604701

[51] Zhang, L.-C. and Chambers, R. L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **66** 479–496. MR2062389

[52] Zhang, X., Zhang, Q., Ma, S. and Fang, K. (2022). Subgroup analysis for high-dimensional functional regression. *Journal of Multivariate Analysis* **192** 105100. MR4484841

[53] Zhao, P., Haziza, D. and Wu, C. (2022). Sample empirical likelihood and the design-based oracle variable selection theory. *Statistica Sinica* **32** 435–457. MR4359640

[54] Zhu, X. and Qu, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics* **12** 171–193. MR3756096

[55] Zhu, X., Tang, X. and Qu, A. (2021). Longitudinal clustering for heterogeneous binary data. *Statistica Sinica* **31** 603–624. MR4286187