



Libraries and Learning Services

# University of Auckland Research Repository, ResearchSpace

## Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

## General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the [Library Thesis Consent Form](#) and [Deposit Licence](#).

# Mixed models for complex survey data

**Xudong Huang**

**A thesis submitted in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy in Statistics**



**THE UNIVERSITY  
OF AUCKLAND**  
NEW ZEALAND

**Department of Statistics**

**2019**



# Mixed models for complex survey data

Xudong Huang

March 3, 2019

## Abstract

The generalised linear mixed models (GLMM) is one of the most important tools for analysing clustered data. One of the main feature of clustered data is observational units within the same cluster are correlated, though observational units from different clusters may be independent. The random effects in the GLMM are used to model this correlation.

The random effects in the GLMM are unobservable. Writing down an exact expression for the marginal likelihood from the GLMM involves a high dimensional integral and so is intractable when the dimension of the random effects is large. There are two different approaches to handle this problem in the literature. First, approximate the integral directly by the Laplace's method (Breslow and Clayton, 1993; Pinheiro and Chao, 2006). Secondly, approximate the integrand or joint density by the lower dimensional object such as the product of marginal density or conditional density. This is also called the pseudo-likelihood estimation (Besag, 1974). Typically, one cannot even write down the marginal likelihood explicitly. So the Laplace's method doesn't apply here. But one can still use the pseudo-likelihood.

Under various regularity conditions, the consistency and asymptotic normality of the pseudo-likelihood estimator have been established using generalised estimating

---

\*Submitted to the Department of Statistics in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Auckland.

equations (GEE). There are many ways to construct the pseudo-likelihood (Lindsay, 1988; Varin et al., 2011). In this thesis, I work exclusively with the pairwise composite likelihood as it is the simplest pseudo-likelihood construction that still captures the pairwise correlation structure.

I am interested in the weighted pairwise composite likelihood under complex sampling. Complex sampling is typically informative (Pfeffermann, 1996). One has to add weights in the pairwise likelihood to account for informative sampling, usually chosen to be the inverse sampling inclusion probability. Rao et al. (2013); Yi et al. (2016) considered the weighted pairwise likelihood for two-stage samples in the special case when the sampling clusters are the model clusters. They established consistency of the weighted pairwise composite likelihood estimator and suggested a variance estimator.

In this thesis, I continue the study of the weighted pairwise composite likelihood estimator in complex sampling initiated in Rao et al. (2013); Yi et al. (2016). More precisely, my goal is to extend the asymptotic results of the weighted pairwise likelihood estimators to the case when the sampling clusters are not the same as the model clusters. In particular, the consistency and asymptotic normality of the weighted pairwise likelihood estimator are established. Furthermore, I show the empirical variance estimator is consistent. This is surprisingly more difficult than it first seems. It is complicated by the structure of the sampling design, where pairs in the same model clusters might not be in the same sampling clusters. I present simulation results examining the performance of the weighted pairwise likelihood estimators for a random intercept model and a random slope model under various two-stage sampling designs.

Finally, the random effects in the mixed model could potentially be correlated as in spatial statistics. My goal in here is to keep extending the asymptotic properties of the weighted pairwise composite likelihood estimator under the Matérn spatial random intercept model. More precisely, I establish consistency and asymptotic normality of the weighted pairwise likelihood estimator under that setting.

Thesis Supervisor: Professor Thomas Lumley

## **Acknowledgments**

It is a pleasure to thank my supervisor Prof. Thomas Lumley for answering my numerous questions. His insight and input were invaluable.

The financial support of the Marsden Fund of New Zealand Royal Society and University of Auckland Doctoral Scholarship Extension is gratefully acknowledged.

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
1.1	Example: Hispanic Community Health Study/Study of Latinos . . . . .	8
1.2	Example: Diabetes . . . . .	10
1.3	Example: School and area . . . . .	12
1.4	Thesis outline . . . . .	13
<b>2</b>	<b>An introduction to complex sampling</b>	<b>15</b>
2.1	Design-based approach . . . . .	15
2.1.1	Basic notation . . . . .	15
2.1.2	Poisson sampling . . . . .	22
2.1.3	Simple random sample without replacement (SRSWOR) . . . . .	25
2.1.4	Stratified sampling . . . . .	28
2.1.5	Two-stage sampling . . . . .	31
2.2	Model-based approach . . . . .	39
2.2.1	Conditional model . . . . .	40
2.2.2	Linear mixed model . . . . .	43
2.2.3	Marginal model . . . . .	44
2.2.3.1	Quasi-likelihood . . . . .	44
2.2.3.2	Generalized Estimating Equation (GEE) . . . . .	48
2.2.4	Relationship between marginal and conditional model . . . . .	50
2.3	Model-design-based approach . . . . .	51
2.3.1	Asymptotic setting . . . . .	51
2.3.2	Informative sampling . . . . .	52
2.3.3	Full-likelihood approach . . . . .	54
2.3.4	Weighted estimating equation . . . . .	55
<b>3</b>	<b>Maximum pseudo-likelihood</b>	<b>60</b>
3.1	Motivation . . . . .	60
3.2	Setting . . . . .	61

3.3	Pairwise composite likelihood estimation without complex sampling . . . .	61
3.4	Weighted pairwise composite likelihood estimation with complex sampling	63
3.4.1	Construction . . . . .	63
3.4.2	Consistency . . . . .	66
3.4.3	Variance estimation . . . . .	75
3.4.4	Yi's approach . . . . .	80
3.4.5	Jackknife variance estimation . . . . .	82
<b>4</b>	<b>When sampling and model clusters are not the same</b>	<b>83</b>
4.1	Setting: design . . . . .	83
4.2	Setting: model . . . . .	88
4.3	Consistency . . . . .	91
4.4	Asymptotic normality . . . . .	93
4.5	Variance estimation . . . . .	99
4.5.1	Empirical variance estimation . . . . .	99
4.6	Consistency of empirical variance estimation: the sampling clusters are the model clusters . . . . .	104
4.6.1	Example: Poisson sampling design . . . . .	105
4.6.2	Example: SRSWOR sampling design . . . . .	107
4.7	Consistency of empirical variance estimation: the sampling clusters are not the same as the model clusters . . . . .	112
4.7.1	Example: Poisson sampling design . . . . .	112
4.7.2	Example: SRSWOR sampling design . . . . .	114
4.8	Model-based variance estimation . . . . .	124
4.9	Simulation: Random intercept model . . . . .	125
4.9.1	Model . . . . .	125
4.9.2	Basic notation . . . . .	128
4.9.3	Design: stratified sampling . . . . .	131
4.9.4	Design: two-stage SRSWOR . . . . .	138
4.9.5	Design: two-stage Poisson . . . . .	142
4.10	Simulation: Random slope model . . . . .	146
4.10.1	Model . . . . .	146



4.10.2	Design: stratified sampling . . . . .	149
4.10.3	Design: two-stage SRSWOR . . . . .	154
4.10.4	Design: two-stage Poisson . . . . .	160
<b>5</b>	<b>Matérn spatial model</b>	<b>164</b>
5.1	Introduction . . . . .	164
5.2	Setting: design . . . . .	164
5.3	Alternative approaches on model . . . . .	164
5.3.1	Why not Gaussian Markov Random Fields (GMRF)? . . . . .	165
5.3.2	Why not marginal models under a mixing conditions? . . . . .	166
5.4	Matérn covariance function . . . . .	168
5.5	Setting: Matérn spatial random intercept model . . . . .	171
5.6	Pointwise Law of Large Numbers . . . . .	173
5.7	Central Limit Theorem . . . . .	179
5.8	Consistency . . . . .	184
5.9	Asymptotic normality . . . . .	186
<b>6</b>	<b>Future work</b>	<b>188</b>
<b>7</b>	<b>Appendix I: elementary result</b>	<b>189</b>
7.1	Basic notation and definition . . . . .	189
7.2	Basic results . . . . .	191
<b>8</b>	<b>References</b>	<b>193</b>

# 1 Introduction

I want to fit a mixed model to a population distribution, but I only have data from a complex (multi-stage) sample. The sampling is informative, that is, the model holding for the (biased) sample is different from the model holding for the population (Pfeffermann, 1996). Ignoring the sampling design and just fitting the mixed model to the sample distribution will lead to biased inference. Although both the sampling and model involve “clusters”, the sample clusters and model clusters need not be the same in general. In addition, the random effects could potentially be correlated over space or time.

I would like to comment on the difference between the sampling clusters and model clusters. The sampling clusters can be thought of the primary sampling units (PSU) in the sampling design and the model clusters can be thought of a fibre (or inverse image) of the random effects. More precisely, the random effect is defined on the set of model clusters and two distinct model clusters are assigned different value of the random effects.

Closely related are Pfeffermann et al. (1998); Pfeffermann and Sverchkov (1999, 2003); Rabe-Hesketh and Skrondal (2006); Pfeffermann and Sverchkov (2009), who proposed various sample weighted likelihoods, and the more recent papers by Rao et al. (2013); Yi et al. (2016), who proposed the weighted pairwise likelihood estimation to such problem under the assumption the sampling clusters are the model clusters and random effects are independent. My main interest is to extend the weighted pairwise likelihood to the case:

- (i) when the sampling clusters are not the same as the model clusters.
- (ii) when the random effects could potentially be correlated.

and works out what can we say about variance estimation under those setting.

I start by considering some examples where these conditions apply.

## 1.1 Example: Hispanic Community Health Study/Study of Latinos

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a genetic cohort study of the US Hispanics/Latinos conducted by the National Heart, Lung, and

Blood Institute ([National Heart, Lung, and Blood Institute, 2010](#); [Conomos et al., 2016](#)). The purpose of the HCHS/SOL is to identify the risk factors associated with cardiovascular disease for Hispanic/Latinos in the US. The risk factors they measured included approximately  $10^6$  genetic variants (SNP).

I want to fit a linear mixed model for the effects of ancestry  $\mathbf{A}$  and a genetic variant  $\mathbf{X}$  on the trait  $\mathbf{Y}$  with genetic relatedness  $\mathbf{b}$  as a random effect. More precisely, let  $\mathbf{Y} = (Y_1, \dots, Y_N)^T$  be the vector of trait (say blood pressure) for the finite population  $U = \{1, \dots, N\}$ . Let  $\mathbf{A} = (A_1, \dots, A_N)^T$  be the vector of ancestry data and let  $\mathbf{X} = (X_1, \dots, X_N)^T$  be the vector of genotypes. Let  $\epsilon = (\epsilon_1, \dots, \epsilon_N)^T$  be a random error with  $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ .

Consider a random intercept mixed model

$$\begin{aligned} Y_k &= \beta_0 + \beta_1 A_k + \beta_2 X_k + b_k + \epsilon_k, \\ \mathbf{b} &\sim N(\mathbf{0}, \tau^2 \mathbf{D}), \end{aligned} \tag{1.1}$$

where  $\mathbf{b} = (b_1, \dots, b_N)^T$  is an unobserved genetic random effect. Some of  $b_k$  in  $\mathbf{b}$  are correlated because of genetic relatedness. More precisely,  $\tau^2$  is a genetic variance due to genetic effect and  $\mathbf{D}$  is a matrix of empirical measure of pairwise kinship relatedness, i.e.,  $\mathbf{D}_{kl} = 2$  if  $k$  and  $l$  are self-identical twins,  $\mathbf{D}_{kl} = 1$  if  $k$  and  $l$  are parent and child, or brother,  $\mathbf{D}_{kl} = \frac{1}{2}$  for grandparent/grandchild, anti/uncle, and  $\mathbf{D}_{kl} = 0$  otherwise. Observe the variance of the blood pressure can be decomposed into two parts: genetic variance component  $\tau^2$  and non-genetic variance  $\sigma^2$ . The parameters I want to estimate are  $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \tau^2)^T$ .

The sampling in HCHS/SOL is based on geographical region. More precisely, sampling is given by a three-stage design: census block group, household and people ([Sorlie et al., 2010](#)). In particular, the sampling clusters are the objects corresponding to the census block group and household.

The sample data one has arisen from a two phase process:

- (i) First-phase: the population data (i.e., all Hispanic/Latino people in the US) are regarded as a realisation of the model 1.1.
- (ii) Second-phase: the sample data (i.e., 16415 observational units ([LaVange et al.,](#)

2010)) are drawn from the population data by complex sampling (i.e., three-stage sampling design).

Observe the sampling clusters are geographic region (i.e., census block group and household) and the model clusters are genetic relatedness (i.e., kinship relatedness). They are not the same. More specifically, some people in the same household are unrelated (e.g., spouses); some people in different households are related. There is no literature work to handle this setting. In fact, I will not solve this problem completely in this thesis, as this is the three-stage sampling design. For what I have to do in this thesis, I only consider the two-stage sampling design when the sampling clusters are not the same as the model clusters. But the three-stage sampling design should be a straightforward extension of my approach.

**Remark.** One could fit a ordinary linear regression using sampling weights

$$Y_k = \beta_0 + \beta_1 A_k + \beta_2 X_k + \epsilon_k,$$

even though the errors  $\epsilon_k$  are correlated. The least squares estimator for  $\beta_0$ ,  $\beta_1$  and  $\beta_2$  is consistent and asymptotically unbiased. However, the least squares estimator is an inefficient estimator. In fact, the loss of information is substantial in the HCHS/SOL. This result is due to HCHS/SOL Genetics Coordinating Centre (Thomas Lumley's personal communication).

## 1.2 Example: Diabetes

The map of estimates of the Percentage of Adults with Diagnosed Diabetes in the Figure 1.1 is based on a survey conducted by the Behavioural Risk Factor Surveillance System (BRFSS) in Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention (CDC), 2007). The sample and model clusters in here are counties.

Let  $Y_{ik}$  be the diabetes status for person  $k$  in county  $i$ , i.e.,  $Y_{ik} = 1$  if the person  $k$  in county  $i$  has diabetes and 0 otherwise. Define  $p_{ik} = \mathbb{E}_Y[Y_{ik}]$ . Let  $X_{ik}$  be the age group variable and let  $b_i$  be an age-adjusted county level unobserved random effect.

Consider a logistic random intercept mixed model

$$\begin{aligned} Y_{ik}|p_{ik} &\sim \text{Bernoulli}(p_{ik}), \\ \text{logit}(p_{ik}) &= \beta_0 + \beta_1 X_{ik} + b_i, \end{aligned} \tag{1.2}$$

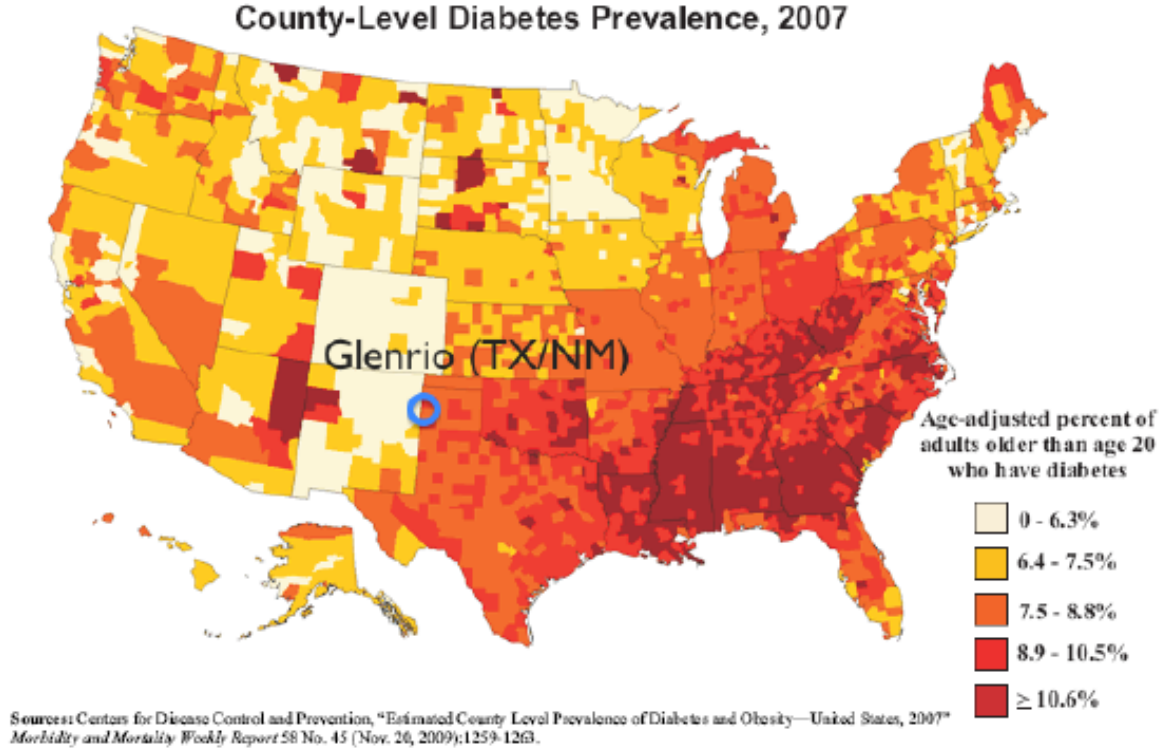


Figure 1.1: Estimates of Percentage of Adults with Diagnosed Diabetes (Centers for Disease Control and Prevention (CDC), 2007).

where the random effects  $b_i \sim N(0, \tau^2)$ . Note the random effects  $b_i$  for different counties are independent. This is a standard setting considered in the literature.

However, the problem with the logistic random intercept model 1.2 is that there is a sudden jump of the percentage of diabetes  $p_{ik}$  across the boundary between different counties as shown in the Figure 1.1. More explicitly, pick a point on the boundary and draw an  $\epsilon$ -ball around that point, no matter how small  $\epsilon$  is,  $p_{ik}$  differs significantly on that  $\epsilon$ -ball. One needs to modify the model assumption that the random effects  $b_i$  are independent.

What I want to do in here is to model the the spatial covariance structure, which could be useful for smoothing the random effect  $b_i$ . In chapter 5, I will study the Matérn

spatial model.

### 1.3 Example: School and area

This is an example raised by J.N.K. Rao in the discussion section of [Pfeffermann et al. \(1998\)](#) which still remains open. Let me start by introducing the setting.

I want to fit a linear mixed model for an individual variable such as the effects of attendance rate  $X_{ik}$  and the school effects  $b_i$  on student's math score  $Y_{ik}$  for student  $k$  in school  $i$ . The model clusters are schools, i.e., students in the same school should be positive correlated and students from different schools should be independent.

Consider a linear mixed model

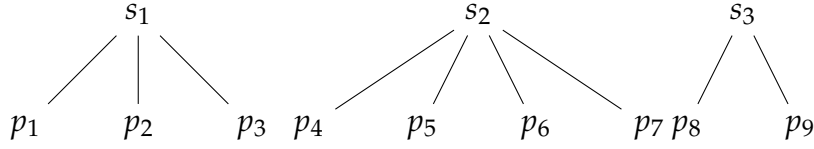
$$\begin{aligned} Y_{ik}|b_i &\sim N(\beta_0 + \beta_1 X_{ik} + b_i, \sigma^2), \\ b_i &\sim N(0, \tau^2). \end{aligned} \tag{1.3}$$

The parameter one wants to estimate is  $\theta = (\beta_0, \beta_1, \sigma^2, \tau^2)^T$ . Note the population data are regarded as a realisation of the model [1.3](#).

The sample data are drawn from the population by a two-stage sampling design based on geographic region. Consider the following two-stage sampling design, area is the primary sampling units (PSU) and students within that areas are the secondary sampling units (SSU). Observe the sampling clusters (i.e., area) are not the same as the model clusters (i.e., school). Some people in the same area can go to different schools; some people in the same school can go to different area, i.e., each student could be nested in a different sampling clusters and model clusters.

The problem one has in here is much simple compared with the problems in the HCHS/SOL, since the sampling design is just a two-stage sampling design. In [chapter 4](#), I will study the weighted pairwise likelihood approach to such problem in detail and analysis the properties of the weighted pairwise likelihood estimator.

Two-stage model for the population model



Two-stage model for the sample

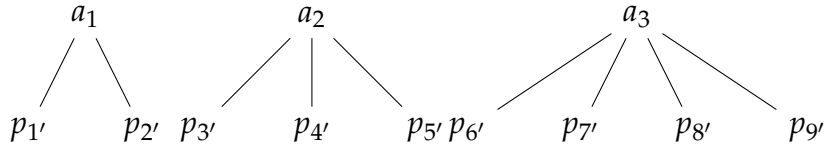


Figure 1.2: School and area.

## 1.4 Thesis outline

The purpose of this thesis is to analyse the asymptotic properties of the weighted pairwise composite likelihood estimators for a general two-stage sampling. My approach is drawn from [Yi et al. \(2016\)](#); [Rao et al. \(2013\)](#). There are five chapters in this thesis.

In chapter 2, I briefly give an introduction to complex sampling from both design and model perspective. Various sampling design such as Poisson, simple random sampling without replacement (SRSWOR), stratified sampling, two-stage sampling will be considered. In addition, one needs to introduce the mixed model to account for correlation in the population. I will also review some well-known results and techniques that will be used in this thesis.

In chapter 3, I review the key results from [Rao et al. \(2013\)](#); [Yi et al. \(2016\)](#) in the setting of the two-stage sampling where the sampling clusters are the model clusters. More precisely, I am going to present proof done by [Yi et al. \(2016\)](#) in detail, which is a starting point in the literature of the weighted pairwise likelihood estimation under complex sampling. In particular, the consistency and variance estimator of the weighted pairwise likelihood estimator will be established.

Chapter 4 is the main contribution of this thesis. In chapter 4, my main interest is to extend the previous results to the case when the sampling clusters are not the same as the model clusters. Essentially, the same proof with minor modification will yield the weighted pairwise likelihood estimator is a consistent estimator, but proving consistency of the empirical variance estimator is surprisingly more difficult than it first seems and requires a new approach. It is complicated by the structure of the sampling design, where pairs in the same model clusters might not be in the same sampling clusters. I also conduct a simulation study examining the performance of the naive maximum likelihood, pairwise likelihood and weighted pairwise likelihood estimator as well as the empirical variance estimator.

In chapter 5, my interest is to keep extending the asymptotic results of the weighted pairwise likelihood estimator to the case when the random effects could be correlated. More precisely, I establish consistency and asymptotic normality of the weighted pairwise likelihood estimator under the Matérn spatial random intercept model.



## 2 An introduction to complex sampling

In this chapter, I outline some basic techniques or terminologies used in complex sampling: design-based estimation, Horvitz-Thompson estimator, super-population model, conditional model, marginal model, informative sampling and asymptotic limit. Essentially, all the terms are standard in the literature and material in here can be found in many textbooks (Fuller, 2011; Cochran, 2007; Cassel et al., 1977; Särndal et al., 1978, 2003; Demidenko, 2013; Song, 2007; Jiang, 2007; Ardilly and Tillé, 2006; Hall, 2005; Grace, 2016; Kim and Shao, 2013).

Roughly speaking, complex sampling involves multi-stage sampling, unequal sampling probability and stratification. There are three approaches for analysing data generated from complex sampling:

- (i) Design-based approach.
- (ii) Model-based approach.
- (iii) Model-design-based approach.

In all of the following, I will take the model-design-based approach. The goal in the model-design-based approach is to construct estimator which incorporates the feature of the sampling design to estimate unknown parameters  $\theta$  from the super-population.

### 2.1 Design-based approach

In this section, I will introduce Horvitz-Thompson (HT) estimator and various sampling design such as Poisson, simple random sampling without replacement (SRSWOW), stratified sampling, two-stage sampling.

#### 2.1.1 Basic notation

In the design-based setting, the sample data are drawn from a fixed finite population of size  $N$ . For notational convenience and without loss of generality, I will identify the elements of population  $U$  by the numbers  $1, \dots, N$ , i.e.,  $U = \{1, 2, \dots, N\}$ . An element  $k$  in the population  $U$  is called an observational unit.

Let  $y_k$  be a response variable. It is assumed to be a fixed unknown constant unless observational unit  $k$  is already in the sample  $S$ . Now I want to introduce probability model to describe how the sample data are drawn from the population. Note the population itself is not considered as random.

**Definition 2.1.** Let  $\mathcal{F}$  be the set of all subsets of  $U$  and let  $\mathbb{P}$  be a probability on measure space  $(U, \mathcal{F})$ , i.e.,

$$\begin{aligned}\mathbb{P} : \mathcal{F} &\longrightarrow [0, 1] \\ A &\longmapsto \mathbb{P}(A).\end{aligned}$$

Then  $\mathbb{P}$  is called the design measure.

The first-order, second-order, fourth-order and eighth-order sampling inclusion probability are important characteristics for design measure  $\mathbb{P}$ . More precisely,

**Definition 2.2.** Given a design measure  $\mathbb{P}$ , the first-order, second-order, fourth-order and eighth-order sampling inclusion probability are defined to be

$$\begin{aligned}\pi_k &= \sum_{\{A \in \mathcal{F} : k \in A\}} \mathbb{P}(A), \\ \pi_{kl} &= \sum_{\{A \in \mathcal{F} : k, l \in A\}} \mathbb{P}(A), \\ \pi_{klk'l'} &= \sum_{\{A \in \mathcal{F} : k, l, k', l' \in A\}} \mathbb{P}(A), \\ \pi_{klk'l'k''l''k'''l'''} &= \sum_{\{A \in \mathcal{F} : k, l, k', l', k'', l'', k''', l''' \in A\}} \mathbb{P}(A),\end{aligned}$$

where  $k, l, k', l', k'', l'', k''', l''' \in U$ .

**Remark.** Observe  $\pi_{kl} = \pi_{lk}$  and  $\pi_{kk} = \pi_k$ . Similarly,  $\pi_{klk'l'}$  is invariant under permutation of  $k, l, k', l'$ .

**Remark.** Note one needs the fourth-order sampling inclusion probabilities  $\pi_{klk'l'}$  to be known explicitly for the empirical variance estimator, and needs assumptions on the eighth-order ones (more about this in chapter 3 and 4).

It is more convenient to work with the sample indicator function  $1_k$  instead of design measure  $\mathbb{P}$ . In all of the following, I denote the sample by  $S$ , which is an element of  $\mathcal{F}$  generated from the design measure  $\mathbb{P}$ .

**Definition 2.3.** Let  $k \in U$ , the sample indicator function is defined to be  $1_k = 1_{\{k \in S\}}$ .

**Remark.** Note the sample indicator function is random from design perspective.

**Remark.** I will use the following notation:

$$\begin{aligned} 1_{kl} &= 1_k 1_l, \\ 1_{klk'l'} &= 1_k 1_l 1_{k'} 1_{l'}, \\ 1_{klk'l'k''l''k'''l'''} &= 1_k 1_l 1_{k'} 1_{l'} 1_{k''} 1_{l''} 1_{k'''} 1_{l'''}. \end{aligned}$$

Henceforth, I denote the expectation with respect to the sampling design by  $\mathbb{E}_\pi$ , variance by  $\text{Var}_\pi$  and covariance by  $\text{Cov}_\pi$ . Define

$$\Delta_{kl} = \text{Cov}_\pi(1_k, 1_l)$$

for all  $k, l \in U$ .

**Lemma 2.4.** *The sample indicator function  $1_k$  has the following properties:*

- (i)  $\mathbb{E}_\pi(1_k) = \pi_k$ .
- (ii)  $\mathbb{E}_\pi(1_{kl}) = \pi_{kl}$ .
- (iii)  $\text{Var}_\pi(1_k) = \pi_k(1 - \pi_k)$ .
- (iv)  $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ .

*Proof.* Obvious. □

**Definition 2.5.** Define the sample size to be  $n = \sum_{k \in U} 1_k$ .

**Remark.** For some design, the sample size  $n$  will be a random variable.

**Lemma 2.6.** *The sample size  $n$  has the following properties:*

$$(i) \mathbb{E}_\pi(n) = \sum_{k \in U} \mathbb{E}_\pi(1_k).$$

$$(ii) \text{Var}_\pi(n) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl}.$$

*Proof.* (i) follows directly from the definition of  $n$ , i.e.,  $n = \sum_{k \in U} 1_k$ . To prove (ii), observe

$$\begin{aligned} \text{Var}_\pi(n) &= \text{Var}_\pi \left( \sum_{k \in U} 1_k \right) \\ &= \mathbb{E}_\pi \left[ \sum_{k \in U} 1_k - \mathbb{E}_\pi \left[ \sum_{k \in U} 1_k \right] \right]^2 \\ &= \mathbb{E}_\pi \left[ \sum_{k \in U} (1_k - \mathbb{E}_\pi 1_k) \right]^2 \\ &= \mathbb{E}_\pi \left[ \sum_{k \in U} \sum_{l \in U} (1_k - \mathbb{E}_\pi 1_k)(1_l - \mathbb{E}_\pi 1_l) \right] \\ &= \sum_{k \in U} \sum_{l \in U} \mathbb{E}_\pi [1_k 1_l - 1_k \mathbb{E}_\pi [1_l] - \mathbb{E}_\pi [1_k] 1_l + \mathbb{E}_\pi [1_k] \mathbb{E}_\pi [1_l]] \\ &= \sum_{k \in U} \sum_{l \in U} [\mathbb{E}_\pi [1_{kl}] - \mathbb{E}_\pi [1_k] \mathbb{E}_\pi [1_l]] \\ &= \sum_{k \in U} \sum_{l \in U} \Delta_{kl}. \end{aligned}$$

□

**Lemma 2.7.** *For a fixed sample size design, then*

$$(i) \sum_{k \in U} \sum_{l \in U} \mathbb{E}_\pi(1_{kl}) = n^2.$$

$$(ii) \sum_{k \in U} \Delta_{kl} = \sum_{l \in U} \Delta_{kl} = 0.$$

*Proof.* To prove (i), observe

$$\sum_{k \in U} \sum_{l \in U} \mathbb{E}_\pi(1_{kl}) = \sum_{k \in U} \sum_{l \in U} \mathbb{E}_\pi(1_k 1_l) = \mathbb{E}_\pi \left( \sum_{k \in U} 1_k \sum_{l \in U} 1_l \right) = n^2.$$

To prove (ii), observe

$$\begin{aligned}
\sum_{k \in U} \Delta_{kl} &= \sum_{k \in U} \mathbb{E}_\pi(1_{kl}) - \sum_{k \in U} \mathbb{E}_\pi(1_k) \mathbb{E}_\pi(1_l) \\
&= \mathbb{E}_\pi \left( \sum_{k \in U} 1_k 1_l \right) - \mathbb{E}_\pi \left( \sum_{k \in U} 1_k \right) \mathbb{E}_\pi(1_l) \\
&= n \mathbb{E}_\pi(1_l) - n \mathbb{E}_\pi(1_l) \\
&= 0.
\end{aligned}$$

Similarly, one can show  $\sum_{l \in U} \Delta_{kl} = 0$ . □

**Definition 2.8.** Define

$$\begin{aligned}
\Delta_{klk'l'} &= \text{Cov}_\pi(1_{kl}, 1_{k'l'}), \\
\Delta_{klk'l'k''l''k'''l'''} &= \text{Cov}_\pi(1_{klk'l'}, 1_{k''l''k'''l'''}).
\end{aligned}$$

**Remark.** Observe

$$\begin{aligned}
\pi_{klk'l'} &= \mathbb{E}_\pi[1_{klk'l'}], \\
\pi_{klk'l'k''l''k'''l'''} &= \mathbb{E}_\pi[1_{klk'l'k''l''k'''l'''}], \\
\Delta_{klk'l'} &= \pi_{klk'l'} - \pi_{kl} \pi_{k'l'}, \\
\Delta_{klk'l'k''l''k'''l'''} &= \pi_{klk'l'k''l''k'''l'''} - \pi_{klk'l'} \pi_{k''l''k'''l'''} .
\end{aligned}$$

**Remark.** In all of the following, I assume there exists  $\epsilon \geq 0$  such that  $\pi_k > \epsilon, \pi_{kl} > \epsilon, \pi_{klk'l'} > \epsilon, \pi_{klk'l'k''l''k'''l'''} > \epsilon$  for all  $k, l, k', l', k'', l'', k''', l''' \in U$ , namely each unit, pair, quadruple and octuple in the population have a positive probability to be in the sample  $S$ .

**Definition 2.9.** Define the first-order, second-order, fourth-order and eighth-order weight

to be

$$\begin{aligned}\omega_k &= \frac{1}{\pi_k}, \\ \omega_{kl} &= \frac{1}{\pi_{kl}}, \\ \omega_{klk'l'} &= \frac{1}{\pi_{klk'l'}}, \\ \omega_{klk'l'k''l''k'''l'''} &= \frac{1}{\pi_{klk'l'k''l''k'''l'''}}.\end{aligned}$$

**Remark.** The first-order weight  $\omega_k$  is interpreted as the number of observational units in the population represented by unit  $k$  in the sample.

The goal of the design-based approach is to estimate some function of  $\mathbf{y} = (y_1, \dots, y_N)^T$  such as the population total  $t = \sum_{k \in U} y_k$ . Let us talk about the Horvitz-Thompson estimator.

**Definition 2.10.** The Horvitz-Thompson (HT) estimator for the population total  $t = \sum_{k \in U} y_k$  is defined to be

$$\hat{t} = \sum_{k \in U} 1_k \omega_k y_k.$$

**Definition 2.11.** Let  $\theta_0$  be the true value. Estimator  $\hat{\theta}_n$  is called the design-unbiased with respect to design measure  $\pi$  if  $\mathbb{E}_\pi(\hat{\theta}_n) = \theta_0$ .

**Lemma 2.12.** The HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:

- (i) The HT estimator  $\hat{t}$  is an unbiased estimator for the population total  $t$ , i.e.,  $\mathbb{E}_\pi(\hat{t}) = t$ .
- (ii) The variance of  $\hat{t}$  is given by

$$\text{Var}_\pi(\hat{t}) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k \omega_l y_l).$$

- (iii) An unbiased estimator for  $\text{Var}_\pi(\hat{t})$  is

$$\widehat{\text{Var}}_\pi(\hat{t}) = \sum_{k \in U} \sum_{l \in U} 1_{kl} \omega_{kl} \Delta_{kl} (\omega_k y_k \omega_l y_l).$$

(iv) The variance of the variance estimator  $\widehat{\text{Var}}_\pi(\hat{t})$  is

$$\text{Var}_\pi \left( \widehat{\text{Var}}_\pi(\hat{t}) \right) = \sum_{k \in U} \sum_{l \in U} \sum_{k' \in U} \sum_{l' \in U} \Delta_{klk'l'} \omega_{kl} \Delta_{kl} \omega_{k'l'} \Delta_{k'l'} (\omega_k y_k \omega_l y_l) (\omega_{k'} y_{k'} \omega_{l'} y_{l'}).$$

**Remark.** The key proofs in Chapters 3 and 4 will have a similar basic structure to Lemma 2.12.

*Proof.* To see this, observe

$$\begin{aligned} \mathbb{E}_\pi(\hat{t}) &= \mathbb{E}_\pi \left[ \sum_{k \in U} 1_k \omega_k y_k \right] = \sum_{k \in U} \mathbb{E}_\pi[1_k] \omega_k y_k = \sum_{k \in U} y_k, \\ \text{Var}_\pi(\hat{t}) &= \text{Var}_\pi \left[ \sum_{k \in U} 1_k \omega_k y_k \right] = \sum_{k \in U} \sum_{l \in U} \text{Cov}_\pi(1_k, 1_l) (\omega_k y_k \omega_l y_l) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k \omega_l y_l), \\ \mathbb{E}_\pi \left[ \widehat{\text{Var}}_\pi(\hat{t}) \right] &= \mathbb{E}_\pi \left[ \sum_{k \in U} \sum_{l \in U} 1_{kl} \omega_{kl} \Delta_{kl} (\omega_k y_k \omega_l y_l) \right] \\ &= \sum_{k \in U} \sum_{l \in U} \mathbb{E}_\pi[1_{kl}] \omega_{kl} \Delta_{kl} (\omega_k y_k \omega_l y_l) \\ &= \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k \omega_l y_l), \\ \text{Var}_\pi \left[ \widehat{\text{Var}}_\pi(\hat{t}) \right] &= \text{Var}_\pi \left[ \sum_{k \in U} \sum_{l \in U} 1_{kl} \omega_{kl} \Delta_{kl} (\omega_k y_k \omega_l y_l) \right] \\ &= \sum_{k \in U} \sum_{l \in U} \sum_{k' \in U} \sum_{l' \in U} \text{Cov}_\pi(1_{kl}, 1_{k'l'}) \omega_{kl} \Delta_{kl} \omega_{k'l'} \Delta_{k'l'} (\omega_k y_k \omega_l y_l) (\omega_{k'} y_{k'} \omega_{l'} y_{l'}) \\ &= \sum_{k \in U} \sum_{l \in U} \sum_{k' \in U} \sum_{l' \in U} \Delta_{klk'l'} \omega_{kl} \Delta_{kl} \omega_{k'l'} \Delta_{k'l'} (\omega_k y_k \omega_l y_l) (\omega_{k'} y_{k'} \omega_{l'} y_{l'}). \end{aligned}$$

□

**Lemma 2.13.** For a fixed sample size design, the HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:

(i) The variance of  $\hat{t}$  is given by

$$\text{Var}_\pi(\hat{t}) = -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k - \omega_l y_l)^2.$$

(ii) An unbiased estimator for  $\text{Var}_\pi(\hat{t})$  is given by

$$\widehat{\text{Var}}_\pi(\hat{t}) = -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} 1_{kl} \omega_{kl} \Delta_{kl} (\omega_k y_k - \omega_l y_l)^2.$$

(iii) The variance of the variance estimator  $\widehat{\text{Var}}_\pi(\hat{t})$  is

$$\text{Var}_\pi \left( \widehat{\text{Var}}_\pi(\hat{t}) \right) = \frac{1}{4} \sum_{k \in U} \sum_{l \in U} \sum_{k' \in U} \sum_{l' \in U} \Delta_{klk'l'} \omega_{kl} \Delta_{kl} \omega_{k'l'} \Delta_{k'l'} (\omega_k y_k - \omega_l y_l)^2 (\omega_{k'} y_{k'} - \omega_{l'} y_{l'})^2.$$

*Proof.* To prove (i), observe

$$\begin{aligned} \text{Var}_\pi(\hat{t}) &= \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k \omega_l y_l) \\ &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k^2 y_k^2 + \omega_l^2 y_l^2 - 2\omega_k y_k \omega_l y_l) \\ &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k - \omega_l y_l)^2. \end{aligned}$$

The first equality follows from Lemma 2.12. The second equality follows from Lemma 2.7, i.e.,

$$\begin{aligned} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \omega_k^2 y_k^2 &= \sum_{k \in U} \omega_k^2 y_k^2 \sum_{l \in U} \Delta_{kl} = 0, \\ \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \omega_l^2 y_l^2 &= \sum_{l \in U} \omega_l^2 y_l^2 \sum_{k \in U} \Delta_{kl} = 0. \end{aligned}$$

The rest is clear. I omit the detail. □

### 2.1.2 Poisson sampling

In a Poisson sampling design, the observational units are sampled independently, but with unequal probability  $\pi_k$ .

**Definition 2.14.** Let  $\{\pi_k\}_{k \in U}$  be a sequence of real number in  $[0, 1]$  and  $\{\epsilon_k\}_{k \in U}$  be a sequence of independent uniform random variable on  $[0, 1]$ . Observational unit  $k$  is in the sample  $S$  if  $\epsilon_k \leq \pi_k$ .



**Remark.** If  $\pi_k = \pi$  for all  $k \in U$ , then it is called the Bernoulli sampling design.

**Lemma 2.15.** *Under a Poisson sample design, the sample indicator function  $1_k$  has the following properties:*

- (i)  $\mathbb{E}_\pi(1_k) = \pi_k$ .
- (ii)  $\mathbb{E}_\pi(1_{kl}) = \pi_k \pi_l$ , if  $k \neq l$ .
- (iii)  $\text{Var}_\pi(1_k) = \pi_k(1 - \pi_k)$ .
- (iv)  $\Delta_{kl} = 0$ , if  $k \neq l$ .

*Proof.* By definition,  $1_k$  is independent of  $1_l$  under design measure if  $k \neq l$ . □

Note the Poisson sampling design is not a fixed sample size design. More precisely,

**Lemma 2.16.** *Under a Poisson sampling design,*

- (i)  $\mathbb{E}_\pi(n) = \sum_{k \in U} \pi_k$ .
- (ii)  $\text{Var}_\pi(n) = \sum_{k \in U} \pi_k(1 - \pi_k)$ .

*Proof.* By Lemma 2.6, one has

$$\begin{aligned} \mathbb{E}_\pi(n) &= \sum_{k \in U} \mathbb{E}_\pi(1_k) = \sum_{k \in U} \pi_k, \\ \text{Var}_\pi(n) &= \sum_{k \in U} \sum_{l \in U} \Delta_{kl} = \sum_{\substack{k=l \\ k, l \in U}} \Delta_{kl} + \sum_{\substack{k \neq l \\ k, l \in U}} \Delta_{kl} = \pi_k(1 - \pi_k) + 0 = \pi_k(1 - \pi_k). \end{aligned}$$

□

**Lemma 2.17.** *Under a Poisson sampling design, the HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:*

- (i) *The HT estimator  $\hat{t}$  for the population total  $t$  is*

$$\hat{t} = \sum_{k \in U} 1_k \omega_k y_k.$$

(ii) The variance of  $\hat{t}$  is given by

$$\text{Var}_{\pi}(\hat{t}) = \sum_{k \in U} (\omega_k - 1) y_k^2.$$

(iii) An unbiased estimator for  $\text{Var}_{\pi}(\hat{t})$  is

$$\widehat{\text{Var}}_{\pi}(\hat{t}) = \sum_{k \in U} 1_k \omega_k (\omega_k - 1) y_k^2.$$

(iv) The variance of the variance estimator  $\widehat{\text{Var}}_{\pi}(\hat{t})$  is

$$\text{Var}_{\pi}(\widehat{\text{Var}}_{\pi}(\hat{t})) = \sum_{k \in U} (\omega_k - 1)^3 y_k^4.$$

*Proof.* (i) is obvious.

To prove (ii), observe

$$\begin{aligned} \text{Var}_{\pi}(\hat{t}) &= \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k \omega_l y_l) \\ &= \sum_{\substack{k=l \\ k, l \in U}} \Delta_{kl} (\omega_k y_k \omega_l y_l) + \sum_{\substack{k \neq l \\ k, l \in U}} \Delta_{kl} (\omega_k y_k \omega_l y_l) \\ &= \sum_{k \in U} \pi_k (1 - \pi_k) \omega_k^2 y_k^2 + 0 \\ &= \sum_{k \in U} (\omega_k - 1) y_k^2. \end{aligned}$$

The first equality follows from Lemma 2.12. The third equality follows from Lemma 2.15.

(iii) is clear.

To prove (iv), observe

$$\begin{aligned}
\text{Var}_{\pi} \left( \widehat{\text{Var}}_{\pi}(\hat{t}) \right) &= \text{Var}_{\pi} \left( \sum_{k \in U} 1_k \omega_k (\omega_k - 1) y_k^2 \right) \\
&= \sum_{k \in U} \omega_k^2 (\omega_k - 1)^2 y_k^4 \text{Var}_{\pi}(1_k) \\
&= \sum_{k \in U} \omega_k^2 (\omega_k - 1)^2 y_k^4 \pi_k (1 - \pi_k) \\
&= \sum_{k \in U} (\omega_k - 1)^3 y_k^4.
\end{aligned}$$

The second equality follows from the fact  $1_k$  is independent of  $1_l$  under design measure if  $k \neq l$ . The third equality follows from Lemma 2.15.  $\square$

**Remark.** Observe  $\text{Var}_{\pi} \left( \widehat{\text{Var}}_{\pi}(\hat{t}) \right) = O(N)$ .

### 2.1.3 Simple random sample without replacement (SRSWOR)

**Definition 2.18.** A simple random sample without replacement (SRSWOR) of fixed size  $n$  is a sampling design  $\mathbb{P}$  such that

$$\begin{aligned}
\mathbb{P} : \mathcal{F} &\longrightarrow [0, 1] \\
A &\longmapsto \mathbb{P}(A) = \begin{cases} \binom{N}{n}^{-1}, & \text{if } |A| = n. \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

**Lemma 2.19.** Under a SRSWOR, the sample indicator function  $1_k$  has the following properties:

- (i)  $\mathbb{E}_{\pi}(1_k) = \frac{n}{N}$ .
- (ii)  $\mathbb{E}_{\pi}(1_{kl}) = \frac{n}{N} \frac{n-1}{N-1}$ , if  $k \neq l$ .
- (iii)  $\text{Var}_{\pi}(1_k) = \frac{n}{N} \left(1 - \frac{n}{N}\right)$ .
- (iv)  $\Delta_{kl} = -\frac{1}{N-1} \frac{n}{N} \left(1 - \frac{n}{N}\right)$ , if  $k \neq l$ .

*Proof.* To see this, observe

$$\begin{aligned}\mathbb{E}_\pi(1_k) &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \\ \mathbb{E}_\pi(1_{kl}) &= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n}{N} \frac{n-1}{N-1}, \\ \text{Var}_\pi(1_k) &= \mathbb{E}_\pi[1_k^2] - [\mathbb{E}_\pi 1_k]^2 = \frac{n}{N} \left(1 - \frac{n}{N}\right), \\ \Delta_{kl} &= \mathbb{E}_\pi[1_{kl}] - \mathbb{E}_\pi[1_k] \mathbb{E}_\pi[1_l] = \frac{n}{N} \left(\frac{n-1}{N-1} - \frac{n}{N}\right) = -\frac{1}{N-1} \frac{n}{N} \left(1 - \frac{n}{N}\right).\end{aligned}$$

□

**Lemma 2.20.** *Under a SRSWOR design, the HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:*

(i) *The HT estimator  $\hat{t}$  for the population total  $t$  is*

$$\hat{t} = \frac{N}{n} \sum_{k \in U} 1_k y_k.$$

(ii) *The variance of  $\hat{t}$  is given by*

$$\text{Var}_\pi(\hat{t}) = \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \sigma^2,$$

where

$$\sigma^2 = \frac{1}{N-1} \sum_{k \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N}\right)^2.$$

(iii) *An unbiased estimator for  $\text{Var}_\pi(\hat{t})$  is*

$$\widehat{\text{Var}}_\pi(\hat{t}) = \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \hat{\sigma}^2,$$

where

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{k \in U} \left(1_k y_k - \frac{\sum_{k \in U} 1_k y_k}{n}\right)^2.$$

*Proof.* (i) is obvious. To prove (ii), one has

$$\begin{aligned}
\text{Var}_{\pi}(\hat{t}) &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (\omega_k y_k - \omega_l y_l)^2 \\
&= \frac{1}{2} \frac{1}{N-1} \frac{n}{N} \left(1 - \frac{n}{N}\right) \frac{N^2}{n^2} \sum_{k \in U} \sum_{l \in U} (y_k - y_l)^2 \\
&= \frac{1}{2} \frac{1}{N-1} \left(1 - \frac{n}{N}\right) \frac{N}{n} \sum_{k \in U} \sum_{l \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N} + \frac{\sum_{k \in U} y_k}{N} - y_l\right)^2 \\
&= \frac{1}{2} \frac{1}{N-1} \left(1 - \frac{n}{N}\right) \frac{N}{n} 2N \sum_{k \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N}\right)^2 \\
&= \frac{1}{N-1} \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \sum_{k \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N}\right)^2 \\
&= \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \sigma^2.
\end{aligned}$$

The first equality follows from Lemma 2.13. The second equality follows from Lemma 2.19. The fourth equality follows from the following fact:

$$\begin{aligned}
\sum_{k \in U} \sum_{l \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N}\right)^2 &= N \sum_{k \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N}\right)^2, \\
\sum_{k \in U} \sum_{l \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N}\right) \left(y_l - \frac{\sum_{k \in U} y_k}{N}\right) &= \sum_{k \in U} \left(y_k - \frac{\sum_{k \in U} y_k}{N}\right) \sum_{l \in U} \left(y_l - \frac{\sum_{k \in U} y_k}{N}\right) = 0.
\end{aligned}$$

To prove (iii), observe

$$\begin{aligned}
\widehat{\text{Var}}_{\pi}(\hat{t}) &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} 1_{kl} \omega_{kl} \Delta_{kl} (\omega_k y_k - \omega_l y_l)^2 \\
&= \frac{1}{2} \frac{N(N-1)}{n(n-1)} \frac{1}{N-1} \frac{n}{N} \left(1 - \frac{n}{N}\right) \frac{N^2}{n^2} \sum_{k \in U} \sum_{l \in U} 1_{kl} (y_k - y_l)^2 \\
&= \frac{1}{2} \frac{1}{n-1} \left(1 - \frac{n}{N}\right) \frac{N^2}{n^2} \sum_{k \in U} \sum_{l \in U} 1_k 1_l \left(y_k - \frac{\sum_{k \in U} 1_k y_k}{n} + \frac{\sum_{k \in U} 1_k y_k}{n} - y_l\right)^2 \\
&= \frac{1}{2} \frac{1}{n-1} \left(1 - \frac{n}{N}\right) \frac{N^2}{n^2} 2n \sum_{k \in U} 1_k \left(y_k - \frac{\sum_{k \in U} 1_k y_k}{n}\right)^2 \\
&= \frac{1}{n-1} \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \sum_{k \in U} 1_k \left(y_k - \frac{\sum_{k \in U} 1_k y_k}{n}\right)^2 \\
&= \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \hat{\sigma}^2.
\end{aligned}$$

The first equality follows from Lemma 2.13. The second equality follows from Lemma 2.19. The fourth equality follows from the following fact:

$$\begin{aligned}
\sum_{k \in U} \sum_{l \in U} 1_k 1_l \left(y_k - \frac{\sum_{k \in U} 1_k y_k}{n}\right)^2 &= n \sum_{k \in U} 1_k \left(y_k - \frac{\sum_{k \in U} 1_k y_k}{n}\right)^2, \\
\sum_{k \in U} \sum_{l \in U} 1_k 1_l \left(y_k - \frac{\sum_{k \in U} 1_k y_k}{n}\right) \left(y_l - \frac{\sum_{k \in U} 1_k y_k}{n}\right) &= \sum_{k \in U} 1_k \left(y_k - \frac{\sum_{k \in U} 1_k y_k}{n}\right) \sum_{l \in U} 1_l \left(y_l - \frac{\sum_{k \in U} 1_k y_k}{n}\right) \\
&= 0.
\end{aligned}$$

□

## 2.1.4 Stratified sampling

**Definition 2.21.** Let  $U_1^c = \{U_1^c, \dots, U_{N_1}^c\}$  be a partition of population  $U$  of size  $N_1, \dots, N_{N_1}$ . For each strata  $U_i^c$ , I select a sample  $U_i^s \subset U_i^c$  of size  $n_i$  according to a sample design  $\mathbb{P}_i(\cdot)$ . The sample  $S$  is given by

$$S = \bigcup_i U_i^s.$$

Assume the selection in each strata is made independently, i.e.,

$$\mathbb{P}_{ii'}(kl) = \mathbb{P}_i(k)\mathbb{P}_{i'}(l)$$

for all  $k \in U_i^c, l \in U_{i'}^c$  and  $i \neq i'$ .

Let  $y_{ik}$  be the value of a response variable  $y$  at the  $k$ -th element in the strata  $i$  and let  $t$  be the population total, i.e.,  $t = \sum_{i=1}^{N_I} t_i = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} y_{ik}$ .

**Lemma 2.22.** *Under a stratified sampling, the HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:*

(i) *The HT estimator  $\hat{t}$  for the population total  $t$  is given by*

$$\hat{t} = \sum_{i=1}^{N_I} \hat{t}_i,$$

where  $\hat{t}_i = \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} y_{ik}$ .

(ii) *The variance of  $\hat{t}$  is given by*

$$\text{Var}_{\pi}(\hat{t}) = \sum_{i=1}^{N_I} \text{Var}_{\pi}(\hat{t}_i),$$

where  $\text{Var}_{\pi}(\hat{t}_i) = \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il})$ .

(iii) *An unbiased estimator for  $\text{Var}_{\pi}(\hat{t})$  is*

$$\widehat{\text{Var}}_{\pi}(\hat{t}) = \sum_{i=1}^{N_I} \widehat{\text{Var}}_{\pi}(\hat{t}_i),$$

where  $\widehat{\text{Var}}_{\pi}(\hat{t}_i) = \sum_{k \in U_i^c} \sum_{l \in U_i^c} 1_{kl|i} \omega_{kl|i} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il})$ .

(iv) *The variance of the variance estimator  $\widehat{\text{Var}}_{\pi}(\hat{t})$  is*

$$\text{Var}_{\pi}(\widehat{\text{Var}}_{\pi}(\hat{t})) = \sum_{i=1}^{N_I} \text{Var}_{\pi}(\widehat{\text{Var}}_{\pi}(\hat{t}_i)),$$

where

$$\begin{aligned} \text{Var}_{\pi} \left( \widehat{\text{Var}}_{\pi}(\hat{t}_i) \right) &= \sum_{k \in U_i^c} \sum_{l \in U_i^c} \sum_{k' \in U_i^c} \sum_{l' \in U_i^c} \Delta_{klk'l'|i} \omega_{kl|i} \Delta_{kl|i} \omega_{k'l'|i} \Delta_{k'l'|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\ &\quad (\omega_{k'|i} y_{ik'} \omega_{l'|i} y_{il'}). \end{aligned}$$

*Proof.* These are simply sum over strata and follow directly from Lemma 2.12. □

**Corollary 2.23** (Stratified Poisson). *Under a stratified Poisson sampling, the HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:*

(i) *The HT estimator  $\hat{t}$  for the population total  $t$  is given by*

$$\hat{t} = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} y_{ik}.$$

(ii) *The variance of  $\hat{t}$  is given by*

$$\text{Var}_{\pi}(\hat{t}) = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} (\omega_{k|i} - 1) y_{ik}^2.$$

(iii) *An unbiased estimator for  $\text{Var}_{\pi}(\hat{t})$  is*

$$\widehat{\text{Var}}_{\pi}(\hat{t}) = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} (\omega_{k|i} - 1) y_{ik}^2.$$

(iv) *The variance of the variance estimator  $\widehat{\text{Var}}_{\pi}(\hat{t})$  is*

$$\text{Var}_{\pi} \left( \widehat{\text{Var}}_{\pi}(\hat{t}) \right) = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} (\omega_{k|i} - 1)^3 y_{ik}^4.$$

**Corollary 2.24** (Stratified SRSWOR). *Under a stratified SRSWOR sampling, the HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:*



(i) The HT estimator  $\hat{t}$  for the population total  $t$  is given by

$$\hat{t} = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} 1_{k|i} \frac{N_i}{n_i} y_{ik}.$$

(ii) The variance of  $\hat{t}$  is given by

$$\text{Var}_{\pi}(\hat{t}) = \sum_{i=1}^{N_I} \left(1 - \frac{n_i}{N_i}\right) \frac{N_i^2}{n_i} \sigma_i^2,$$

where

$$\sigma_i^2 = \frac{1}{N_i - 1} \sum_{k \in U_i^c} \left( y_{ik} - \frac{\sum_{k \in U_i^c} y_{ik}}{N_i} \right)^2.$$

(iii) An unbiased estimator for  $\text{Var}_{\pi}(\hat{t})$  is

$$\widehat{\text{Var}}_{\pi}(\hat{t}) = \sum_{i=1}^{N_I} \left(1 - \frac{n_i}{N_i}\right) \frac{N_i^2}{n_i} \hat{\sigma}_i^2,$$

where

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{k \in U_i^c} \left( 1_{k|i} y_{ik} - \frac{\sum_{k \in U_i^c} 1_{k|i} y_{ik}}{n_i} \right)^2.$$

### 2.1.5 Two-stage sampling

Let  $U = \{1, \dots, N\}$  be the population and consider a partition of  $U$  by the sampling clusters, i.e., let  $U_I^c = \{U_1^c, \dots, U_{N_I}^c\}$  be the sampling clusters. In practice, the sampling clusters are the objects corresponding to the first-stage sampling units or primary sampling units (PSU).

Define  $\mathcal{F}$  to be the set of all subsets of  $U_I^c$ , i.e.,

$$\mathcal{F} = \left\{ \bigcup_{i \in I} U_i^c : I \subset \{1, 2, \dots, N_I\} \right\}.$$

Let  $\mathbb{P}_I$  be the first-stage design probability on measure space  $(U_I^c, \mathcal{F})$

$$\begin{aligned}\mathbb{P}_I : \mathcal{F} &\longrightarrow [0, 1] \\ A &\longmapsto \mathbb{P}_I(A).\end{aligned}$$

Define  $\mathcal{F}_i$  to be the set of all subsets of  $U_i^c$ , where  $i = 1, \dots, N_I$ . Let  $\mathbb{P}_i$  be the second-stage design probability on measure space  $(U_i^c, \mathcal{F}_i)$

$$\begin{aligned}\mathbb{P}_i : \mathcal{F}_i &\longrightarrow [0, 1] \\ A &\longmapsto \mathbb{P}_i(A).\end{aligned}$$

Let  $U_I^s \in \mathcal{F}$  be the first-stage sample. If  $U_i^c \in U_I^s$ , a second-stage sample  $U_i^s$  is selected from  $\mathcal{F}_i$  by sampling design  $\mathbb{P}_i$ . Let  $S = \bigcup_i U_i^s$  and  $n = |S|$ .

**Remark.** In all of the following, I will denote  $i, i', i'', i'''$  for the clusters and  $k, l, k', l', k'', l'', k''', l'''$  for elements in the clusters.

**Definition 2.25.** Define the first-order, second-order and fourth-order sampling inclusion probability to be

$$\begin{aligned}\pi_i &= \sum_{A \in \mathcal{F} : U_i^c \in A} \mathbb{P}_I(A), \\ \pi_{ii'} &= \sum_{A \in \mathcal{F} : U_i^c, U_{i'}^c \in A} \mathbb{P}_I(A), \\ \pi_{ii'i''i'''} &= \sum_{A \in \mathcal{F} : U_i^c, U_{i'}^c, U_{i''}^c, U_{i'''}^c \in A} \mathbb{P}_I(A), \\ \pi_{k|i} &= \sum_{A \in \mathcal{F}_i : k \in A} \mathbb{P}_i(A), \\ \pi_{kl|i} &= \sum_{A \in \mathcal{F}_i : k, l \in A} \mathbb{P}_i(A), \\ \pi_{klk'l'|i} &= \sum_{A \in \mathcal{F}_i : k, l, k', l' \in A} \mathbb{P}_i(A).\end{aligned}$$

Assuming the invariance and independent properties, i.e.,

- (i) The second-stage sampling design is invariant of the first-stage  $U_1^s$ , i.e.,  $\mathbb{P}_i(k|U_1^s) = \mathbb{P}_i(k)$  for all  $k \in U_i^c$  and  $i = 1, \dots, N_1$ .
- (ii) The second-stage sampling design is independent of the first-stage  $U_1^s$ , i.e.,  $\mathbb{P}_{ii'}(kl) = \mathbb{P}_i(k)\mathbb{P}_{i'}(l)$  for all  $k \in U_i^c, l \in U_{i'}^c$  and  $i \neq i'$ .

In other words, given that cluster  $i$  is sampled, the probability of sampling element  $k$  in cluster  $i$  does not depend on

- (i) which other clusters were sampled.
- (ii) which elements are sampled in those clusters.

Then one can deduce the final inclusion probability  $\pi_k$  and  $\pi_{kl}$  to be

$$\pi_k = \pi_i \pi_{k|i}, \quad \text{if } k \in U_i^c.$$

$$\pi_{kl} = \begin{cases} \pi_i \pi_{kl|i}, & \text{if } k, l \in U_i^c. \\ \pi_{ii'} \pi_{k|i} \pi_{l|i'}, & \text{if } k \in U_i^c, l \in U_{i'}^c \text{ and } i \neq i'. \end{cases}$$

**Definition 2.26.** Define the first and second-stage sample weight to be

$$\begin{aligned} \omega_i &= \frac{1}{\pi_i}, \\ \omega_{ii'} &= \frac{1}{\pi_{ii'}}, \\ \omega_{k|i} &= \frac{1}{\pi_{k|i}}, \\ \omega_{kl|i} &= \frac{1}{\pi_{kl|i}}, \\ \omega_k &= \frac{1}{\pi_k}, \\ \omega_{kl} &= \frac{1}{\pi_{kl}}. \end{aligned}$$

It is more convenient to work with the sample indicator function  $1_i, 1_{k|i}$  instead of design measure.

**Definition 2.27.** Define the sample indicator function to be

$$1_i = 1_{U_i^c \in U_1^s},$$

$$1_{k|i} = 1_{k \in U_i^s | U_i^c \in U_1^s}.$$

**Remark.** I will use the following notation:

$$1_{ii'} = 1_i 1_{i'},$$

$$1_{ii' i'' i'''} = 1_i 1_{i'} 1_{i''} 1_{i'''},$$

$$1_{kl|i} = 1_{k|i} 1_{l|i},$$

$$1_{klk'l'|i} = 1_{k|i} 1_{l|i} 1_{k'|i} 1_{l'|i}.$$

**Remark.** Observe

$$\mathbb{E}_\pi(1_i) = \pi_i,$$

$$\mathbb{E}_\pi(1_{ii'}) = \pi_{ii'},$$

$$\mathbb{E}_\pi(1_{ii' i'' i'''}) = \pi_{ii' i'' i'''},$$

$$\mathbb{E}_\pi(1_{k|i}) = \pi_{k|i},$$

$$\mathbb{E}_\pi(1_{kl|i}) = \pi_{kl|i},$$

$$\mathbb{E}_\pi(1_{klk'l'|i}) = \pi_{klk'l'|i}.$$

**Definition 2.28.** Define

$$\Delta_{ii'} = \text{Cov}_\pi(1_i, 1_{i'}),$$

$$\Delta_{ii' i'' i'''} = \text{Cov}_\pi(1_{ii'}, 1_{i'' i'''}),$$

$$\Delta_{kl|i} = \text{Cov}_\pi(1_{k|i}, 1_{l|i}),$$

$$\Delta_{klk'l'|i} = \text{Cov}_\pi(1_{kl|i}, 1_{k'l'|i}).$$

**Remark.** Observe

$$\Delta_{ii'} = \pi_{ii'} - \pi_i \pi_{i'},$$

$$\Delta_{ii' i'' i'''} = \pi_{ii' i'' i'''} - \pi_{ii'} \pi_{i'' i'''},$$

$$\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i} \pi_{l|i},$$

$$\Delta_{klk'l'|i} = \pi_{klk'l'|i} - \pi_{kl|i} \pi_{k'l'|i}.$$

**Definition 2.29.** Let  $N_I$  be the number of clusters in the population  $U$  and let  $n_I$  be the number of clusters in the sample  $S$ , i.e.,  $n_I = \sum_{i=1}^{N_I} 1_i$ .

**Definition 2.30.** Let  $N_i$  be the number of elements in the population cluster  $U_i^c$  and let  $n_i$  be the number of elements in the sample clusters  $U_i^s$ , i.e.,  $n_i = \sum_{k \in U_i^c} 1_{k|i}$ .

**Remark.** Observe

$$N = \sum_{i=1}^{N_I} N_i,$$

$$n = \sum_{i=1}^{N_I} 1_i n_i.$$

Let  $y_{ik}$  be the value of a response variable  $y$  at the  $k$ -th element in the cluster  $i$ . Define the population total to be  $t = \sum_{i=1}^{N_I} t_i$ , where  $t_i = \sum_{k \in U_i^c} y_{ik}$ .

In all of the following, let  $\pi_1$  be the first-stage sampling probability and let  $\pi_2$  be the second-stage sampling probability, conditional on the first-stage sampling clusters  $U_1^s$ .

**Lemma 2.31.** *Under a two-stage sampling, the HT estimator  $\hat{t}$  for the population total  $t$  has the following properties:*

(i) *The HT estimator  $\hat{t}$  for the population total  $t$  is given by*

$$\hat{t} = \sum_{i=1}^{N_I} 1_i \omega_i \hat{t}_i,$$

*where  $\hat{t}_i = \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} y_{ik}$ .*

(ii) *The HT estimator  $\hat{t}$  is an unbiased estimator for population total  $t$ .*

(iii) *The variance of  $\hat{t}$  is given by*

$$\text{Var}_{\pi}(\hat{t}) = V_1 + V_2,$$

where

$$V_1 = \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} (\omega_i t_i \omega_{i'} t_{i'}) ,$$

$$V_2 = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \omega_i \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) .$$

(iv) An unbiased estimator for  $\text{Var}_\pi(\hat{t})$  is

$$\widehat{\text{Var}}_\pi(\hat{t}) = \hat{V}_1 + \hat{V}_2,$$

where

$$\hat{V}_1 = \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} 1_{ii'} \omega_{ii'} \Delta_{ii'} (\omega_i \hat{t}_i \omega_{i'} \hat{t}_{i'}) ,$$

$$\hat{V}_2 = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} 1_i \omega_i 1_{kl|i} \omega_{kl|i} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) .$$

*Proof.* (i) is obvious. To prove (ii), note

$$\begin{aligned} \mathbb{E}_\pi(\hat{t}) &= \mathbb{E}_{\pi_1, \pi_2} \left[ \sum_{i=1}^{N_I} 1_i \omega_i \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} y_{ik} \right] \\ &= \mathbb{E}_{\pi_1} \left[ \sum_{i=1}^{N_I} 1_i \omega_i \sum_{k \in U_i^c} \mathbb{E}_{\pi_2} [1_{k|i}] \omega_{k|i} y_{ik} \right] \\ &= \mathbb{E}_{\pi_1} \left[ \sum_{i=1}^{N_I} 1_i \omega_i t_i \right] \\ &= t. \end{aligned}$$

To prove (iii), observe

$$\begin{aligned}
\text{Var}_{\pi}(\hat{t}) &= \text{Var}_{\pi_1} \left( \mathbb{E}_{\pi_2} \left[ \sum_{i=1}^{N_I} \sum_{k \in U_i^c} 1_i \omega_i 1_{k|i} \omega_{k|i} y_{ik} \right] \right) + \mathbb{E}_{\pi_1} \left[ \text{Var}_{\pi_2} \left( \sum_{i=1}^{N_I} \sum_{k \in U_i^c} 1_i \omega_i 1_{k|i} \omega_{k|i} y_{ik} \right) \right] \\
&= \text{Var}_{\pi_1} \left( \sum_{i=1}^{N_I} 1_i \omega_i \mathbb{E}_{\pi_2} \left[ \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} y_{ik} \right] \right) + \mathbb{E}_{\pi_1} \left[ \sum_{i=1}^{N_I} 1_i \omega_i^2 \text{Var}_{\pi_2} \left( \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} y_{ik} \right) \right] \\
&= \text{Var}_{\pi_1} \left( \sum_{i=1}^{N_I} 1_i \omega_i t_i \right) + \mathbb{E}_{\pi_1} \left[ \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} 1_i \omega_i^2 \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \right] \\
&= \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} (\omega_i t_i \omega_{i'} t_{i'}) + \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \omega_i \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\
&= V_1 + V_2.
\end{aligned}$$

In the second equality, I used the fact that  $1_i$  is independent of  $1_{i'}$  under design measure  $\pi_2$ .

To prove (iv), note

$$\begin{aligned}
& \mathbb{E}_\pi \left[ \widehat{\text{Var}}_\pi(\hat{t}) \right] \\
&= \mathbb{E}_{\pi_1, \pi_2} \left[ \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} 1_{ii'} \omega_{ii'} \Delta_{ii'} \omega_i \hat{t}_i \omega_{i'} \hat{t}_{i'} + \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} 1_i \omega_i 1_{kl|i} \omega_{kl|i} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \right] \\
&= \mathbb{E}_{\pi_1} \left[ \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} 1_{ii'} \omega_{ii'} \Delta_{ii'} \omega_i \omega_{i'} \mathbb{E}_{\pi_2} [\hat{t}_i \hat{t}_{i'}] \right] + \mathbb{E}_{\pi_1} \left[ \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} 1_i \omega_i \mathbb{E}_{\pi_2} [1_{kl|i}] \omega_{kl|i} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \right] \\
&= \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} \omega_i \omega_{i'} \mathbb{E}_{\pi_2} [\hat{t}_i \hat{t}_{i'}] + \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\
&= \sum_{\substack{i, i'=1 \\ i=i'}}^{N_I} \Delta_{ii'} \omega_i \omega_{i'} \mathbb{E}_{\pi_2} [\hat{t}_i \hat{t}_{i'}] + \sum_{\substack{i, i'=1 \\ i \neq i'}}^{N_I} \Delta_{ii'} \omega_i \omega_{i'} \mathbb{E}_{\pi_2} [\hat{t}_i \hat{t}_{i'}] + \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\
&= \sum_{\substack{i, i'=1 \\ i=i'}}^{N_I} \Delta_{ii'} \omega_i \omega_{i'} \left[ \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) + t_i^2 \right] + \sum_{\substack{i, i'=1 \\ i \neq i'}}^{N_I} \Delta_{ii'} \omega_i \omega_{i'} t_i t_{i'} + \\
&\quad \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\
&= \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} \omega_i \omega_{i'} t_i t_{i'} + \sum_{i=1}^{N_I} \pi_i (1 - \pi_i) \omega_i^2 \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) + \\
&\quad \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\
&= \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} (\omega_i t_i \omega_{i'} t_{i'}) + \sum_{i=1}^{N_I} (\omega_i - 1) \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) + \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\
&= \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} (\omega_i t_i \omega_{i'} t_{i'}) + \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \sum_{l \in U_i^c} \omega_i \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) \\
&= V_1 + V_2.
\end{aligned}$$



The fifth equality follows from the following facts:

$$\begin{aligned}\mathbb{E}_{\pi_2}[\widehat{t}_i \widehat{t}_{i'}] &= \text{Cov}_{\pi_2}(\widehat{t}_i, \widehat{t}_{i'}) + \mathbb{E}_{\pi_2}[\widehat{t}_i] \mathbb{E}_{\pi_2}[\widehat{t}_{i'}] \\ &= \begin{cases} \sum_{k \in U_i^c} \sum_{l \in U_{i'}^c} \Delta_{kl|i} (\omega_{k|i} y_{ik} \omega_{l|i} y_{il}) + t_i^2, & \text{if } i = i'. \\ t_i t_{i'}, & \text{if } i \neq i'. \end{cases}\end{aligned}$$

□

## 2.2 Model-based approach

The model-based approach assumes a model for the population and sampling design is typically assumed to be ignorable given the model covariates, i.e., the sample data follow the population model when the sampling is ignorable. In other words,  $\mathbf{y} = (y_1, \dots, y_N)^T$  is considered to be an realisation of a random vector  $\mathbf{Y} = (Y_1, \dots, Y_N)^T$  from some parametric model. The goal in here is to estimate the model parameters. For what I have to do in this thesis, the observational units are not homogenous, i.e., the conditional distribution  $y_i|x_i$  is not independent identically distributed (IID). More explicitly, consider a partition of  $U$  by the structure of the model, i.e., let  $M_1^c = \{M_1^c, \dots, M_{T_c}^c\}$  be the model clusters. What I want to do is to assume the observational units within the same model cluster are correlated, though observational units from different model clusters may be independent. In other words, the model clusters are the objects corresponding to the fibre (inverse image) of random effects. More precisely, the random effects are defined on the set of model clusters and two distinct model clusters are assigned different value of the random effects. In this section, I assume the sampling clusters are the model clusters, i.e.,  $U_1^c = M_1^c$ .

Roughly speaking, there are two approaches to model the correlated data, namely the conditional models and marginal models (Neuhaus et al., 1991). A marginal (conditional) model is one in which the marginal (conditional) mean and covariance are modelled directly. The main difference between two modelling approaches is on the interpretation of the parameters. In the marginal (conditional) model, the parameters are interpreted as population-average (cluster-specific effect). Typically, estimation in the conditional model is done by the maximum likelihood estimation (MLE) and estimation

in the marginal model is done by generalised estimating equation (GEE). I will take a conditional approach in this thesis.

Henceforth, I denote the expectation with respect to model by  $\mathbb{E}_Y$ , covariance by  $\text{Cov}_Y$  and variance by  $\text{Var}_Y$ .

### 2.2.1 Conditional model

The history of the generalised linear mixed model (GLMM) is reviewed by [Breslow and Clayton \(1993\)](#); [Diggle \(2002\)](#). The GLMM is an extension of generalized linear model ([McCullagh and Nelder, 1989](#)). More precisely,

**Definition 2.32.** Let  $Y_{ik}$  be a  $k$ th random variable for cluster  $i$  for  $i = 1, \dots, N_I, k = 1, \dots, N_i$  and  $\mathbf{b}_i$  be the random effects for cluster  $i$ . Conditional distribution  $Y_{ik}|\mathbf{b}_i$  is said to be GLMM if the following conditions are satisfied:

- (1)  $Y_{ik}|\mathbf{b}_i$  is independent and follows a distribution from the exponential family, i.e.,

$$f(y_{ik}|\mathbf{b}_i) = \exp \left[ \frac{y_{ik}\theta_{ik} - d(\theta_{ik})}{\phi} + c(y_{ik}, \phi) \right],$$

where  $\phi$  is some scale parameter and  $d, c$  is some known functions.

- (2) The conditional mean  $\mathbb{E}_Y[Y_{ik}|\mathbf{b}_i]$  can be modelled through a smooth monotone link function  $g$ , i.e.,

$$g(\mathbb{E}_Y[Y_{ik}|\mathbf{b}_i]) = \mathbf{X}_{ik}^T \boldsymbol{\beta} + \mathbf{Z}_{ik}^T \mathbf{b}_i,$$

where

- (i)  $\mathbf{X}_{ik}$  is a  $p_1 \times 1$  fixed matrix and  $\boldsymbol{\beta}$  is a  $p_1 \times 1$  fixed unknown parameter.
- (ii)  $\mathbf{Z}_{ik}$  is a  $p_2 \times 1$  fixed matrix,  $\mathbf{b}_i$  is a  $p_2 \times 1$  unobserved random effect and follows a normal distribution, i.e.,  $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\eta}))$ .

The parameters one wants to estimate are  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})^T$ .

**Remark.** The joint density of  $\mathbf{Y}$  and  $\mathbf{b}$  is given by

$$f(\mathbf{Y}, \mathbf{b}) = \prod_{i=1}^{N_I} f(\mathbf{Y}_i, \mathbf{b}_i) = \prod_{i=1}^{N_I} f(\mathbf{Y}_i|\mathbf{b}_i)f(\mathbf{b}_i) = \prod_{i=1}^{N_I} \prod_{k=1}^{N_i} f(Y_{ik}|\mathbf{b}_i)f(\mathbf{b}_i).$$

Since the random effects  $\mathbf{b}_i$  are unobserved, then the marginal density  $f(\mathbf{Y})$  is given by

$$f(\mathbf{Y}) = \int \prod_{i=1}^{N_I} \prod_{k=1}^{N_i} f(Y_{ik}|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i = \prod_{i=1}^{N_I} \int \prod_{k=1}^{N_i} f(Y_{ik}|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i.$$

**Definition 2.33.** The census full log-likelihood is given by

$$\ell^c(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} \ell_i^c(\boldsymbol{\theta}),$$

where

$$\ell_i^c(\boldsymbol{\theta}) = \log \left[ \int \prod_{k=1}^{N_i} f(y_{ik}|\mathbf{b}_i) |\mathbf{D}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i \right) d\mathbf{b}_i \right].$$

**Definition 2.34.** The naive sample log-likelihood is given by

$$\ell^s(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} 1_i \ell_i^s(\boldsymbol{\theta}),$$

where

$$\ell_i^s(\boldsymbol{\theta}) = \log \left[ \int \prod_{k \in \mathcal{U}_i^s} f(y_{ik}|\mathbf{b}_i) |\mathbf{D}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i \right) d\mathbf{b}_i \right]. \quad (2.1)$$

**Remark.** The integral in 2.1 involves a high-dimensional integral and so is intractable when the dimension of random effects  $\mathbf{b}_i$  is large. Approximation of integration based on the Laplace approximation is discussed in [Breslow and Clayton \(1993\)](#); [Pinheiro and Chao \(2006\)](#). For the linear mixed model, one can compute the marginal distribution of  $\mathbf{Y}_i$  directly (more about this in section 2.2.2).

**Example 1.** Let  $p_{ik} \in [0, 1]$ . The conditional logistic mixed model is given by

$$\begin{aligned} Y_{ik} | p_{ik} &\sim \text{Bernoulli}(p_{ik}), \\ \text{logit}(p_{ik}) &= \mathbf{X}_{ik}^T \boldsymbol{\beta} + \mathbf{Z}_{ik}^T \mathbf{b}_i, \end{aligned}$$

where the random effects  $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\eta}))$ . The census full log-likelihood is given by

$$\ell^c(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} \ell_i^c(\boldsymbol{\theta}),$$

where

$$\ell_i^c(\boldsymbol{\theta}) = \log \left[ \int \prod_{k=1}^{N_i} p_{ik}^{y_{ik}} (1 - p_{ik})^{1-y_{ik}} |\mathbf{D}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i \right) d\mathbf{b}_i \right].$$

The naive sample log-likelihood is given by

$$\ell^s(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} 1_i \ell_i^s(\boldsymbol{\theta}),$$

where

$$\ell_i^s(\boldsymbol{\theta}) = \log \left[ \int \prod_{k \in U_i^s} p_{ik}^{y_{ik}} (1 - p_{ik})^{1-y_{ik}} |\mathbf{D}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i \right) d\mathbf{b}_i \right].$$

**Example 2.** Let  $\lambda_{ik} \in \mathbb{N}$ . The conditional Poisson mixed model is given by

$$\begin{aligned} Y_{ik} | \lambda_{ik} &\sim \text{Poisson}(\lambda_{ik}), \\ \log(\lambda_{ik}) &= \mathbf{x}_{ik}^T \boldsymbol{\beta} + \mathbf{z}_{ik}^T \mathbf{b}_i, \end{aligned}$$

where the random effects  $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\eta}))$ . The census full log-likelihood is given by

$$\ell^c(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} \ell_i^c(\boldsymbol{\theta}),$$

where

$$\ell_i^c(\boldsymbol{\theta}) = \log \left[ \int \prod_{k=1}^{N_i} (\exp \lambda_{ik}) \frac{\lambda_{ik}^{y_{ik}}}{y_{ik}!} |\mathbf{D}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i \right) d\mathbf{b}_i \right].$$

The naive sample log-likelihood is given by

$$\ell^s(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} 1_i \ell_i^s(\boldsymbol{\theta}),$$

where

$$\ell_i^s(\boldsymbol{\theta}) = \log \left[ \int \prod_{k \in U_i^s} (\exp \lambda_{ik}) \frac{\lambda_{ik}^{y_{ik}}}{y_{ik}!} |\mathbf{D}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i \right) d\mathbf{b}_i \right].$$

### 2.2.2 Linear mixed model

The modern formulation of the linear mixed model was proposed by [Laird and Ware \(1982\)](#). More precisely,

**Definition 2.35.** A linear mixed model is given by

$$\begin{aligned} Y_i | X_i, b_i &\sim N(X_i \beta + Z_i b_i, \sigma^2 I), \\ b_i &\sim N(0, D(\eta)), \end{aligned}$$

where

- (i)  $Y_i$  is a  $N_i \times 1$  random vector for  $i = 1, \dots, N_I$ .
- (ii)  $X_i$  is a  $N_i \times p_1$  fixed matrix and  $\beta$  is a  $p_1 \times 1$  fixed unknown parameter.
- (iii)  $Z_i$  is a  $N_i \times p_2$  fixed matrix and  $b_i$  is a  $p_2 \times 1$  dimensional random effects.

The parameters I want to estimate are  $\theta = (\beta, \sigma^2, \eta)^T$ . One can show the census full log-likelihood is given by

$$Y_i | X_i \sim N(X_i \beta, V_i(\sigma^2, \eta)),$$

where  $V_i(\sigma^2, \eta) = \sigma^2 I + Z_i D(\eta) Z_i^T$ . In other words, the census full log-likelihood is given by

$$\ell^c(\theta) = -\frac{1}{2} \sum_{i=1}^{N_I} \log |V_i| - \frac{1}{2} \sum_{i=1}^{N_I} (y_i - x_i \beta)^T V_i^{-1} (y_i - x_i \beta).$$

Let  $y_i^s, x_i^s, b_i^s, V_i^s$  be the sample data corresponding to  $y_i, x_i, b_i, V_i$ . Then the naive sample log-likelihood is given by

$$\ell^s(\theta) = -\frac{1}{2} \sum_{i=1}^{N_I} 1_i \log |V_i^s| - \frac{1}{2} \sum_{i=1}^{N_I} 1_i (y_i^s - x_i^s \beta)^T (V_i^s)^{-1} (y_i^s - x_i^s \beta).$$

The naive maximum likelihood estimator for  $\beta$  is given by

$$\hat{\beta}(\sigma^2, \eta) = \left( \sum_{i=1}^{N_I} 1_i (x_i^s)^T (V_i^s)^{-1} x_i^s \right)^{-1} \left( \sum_{i=1}^{N_I} 1_i x_i^s (V_i^s)^{-1} y_i^s \right).$$

For the variance component  $(\sigma^2, \boldsymbol{\eta})^T$ , this reduces to maximise the profile log-likelihood

$$\ell_p^s(\sigma^2, \boldsymbol{\eta}) = \ell^s(\hat{\boldsymbol{\beta}}(\sigma^2, \boldsymbol{\eta}), \sigma^2, \boldsymbol{\eta}).$$

In general, there is no closed form formula for  $(\hat{\sigma}^2, \hat{\boldsymbol{\eta}})^T$ , but numerical method can be used to obtain  $\boldsymbol{\theta}$  (Demidenko, 2013).

### 2.2.3 Marginal model

In the marginal model, I only make assumptions about the first and second moment of a response variable  $Y$ , without specifying the exact distribution of  $Y$ . Hence one cannot use the MLE, since the full-likelihood can not be uniquely determined from the first two moments. In general, one has to use GEE techniques to handle the marginal model with correlated data.

**2.2.3.1 Quasi-likelihood** In this section, I briefly mention how to construct the census quasi-likelihood from sum of the single quasi-likelihood. The quasi-likelihood estimation method was first proposed by Wedderburn (1974). I start by introducing the setting.

- (i) Marginal mean  $\mu_{ik} = \mathbb{E}_Y(Y_{ik})$  depends on the model variables through a smooth monotone link function  $g$ , i.e.,  $g(\mu_{ik}) = \mathbf{X}_{ik}^T \boldsymbol{\beta}$ .
- (ii) Marginal variance depends on the marginal mean, i.e.,  $\text{Var}_Y(Y_{ik}) = \phi V(\mu_{ik})$ , where  $\phi$  is some scaling parameter and  $V$  is some known function of  $\mu_{ik}$ .

The parameters I want to estimate are  $\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_1, \dots, \beta_{p_1})^T$ . One can construct the quasi-likelihood. More precisely,

**Definition 2.36.** The census quasi-likelihood function is defined to be

$$Q^c(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} Q_{ik}(\mu_{ik}, y_{ik}),$$

where

$$Q_{ik}(\mu_{ik}, y_{ik}) = \int_{y_{ik}}^{\mu_{ik}} \frac{y_{ik} - t}{\phi V(t)} dt.$$

**Remark.** Note there is no assumption that the data are generated from this quasi-likelihood function.

**Remark.** Observe

$$\begin{aligned}\mathbb{E}_Y [\nabla_{\mu_{ik}} Q_{ik}(\mu_{ik}, Y_{ik})] &= 0, \\ \text{Var}_Y [\nabla_{\mu_{ik}} Q_{ik}(\mu_{ik}, Y_{ik})] &= \frac{1}{\phi V(\mu_{ik})}, \\ \mathbb{E}_Y [\nabla_{\mu_{ik}\mu_{ik}}^2 Q_{ik}(\mu_{ik}, Y_{ik})] &= -\frac{1}{\phi V(\mu_{ik})}.\end{aligned}$$

**Lemma 2.37.** *Each quasi-score function  $\nabla_{\beta_q} Q_{ik}(\mu_{ik}, Y_{ik})$  is an unbiased estimating equation with respect to model measure  $Y$ , i.e.,*

$$\mathbb{E}_Y [\nabla_{\beta_q} Q_{ik}(\mu_{ik}, Y_{ik})] = 0 \quad \text{for all } q = 1, \dots, p_1.$$

*Proof.* To see this, note

$$\begin{aligned}\mathbb{E}_Y [\nabla_{\beta_q} Q_{ik}(\mu_{ik}, Y_{ik})] &= \mathbb{E}_Y \left[ \nabla_{\beta_q} \int_{Y_{ik}}^{\mu_{ik}} \frac{Y_{ik} - t}{\phi V(t)} dt \right] \\ &= \mathbb{E}_Y \left[ \nabla_{\mu_{ik}} \int_{Y_{ik}}^{\mu_{ik}} \frac{Y_{ik} - t}{\phi V(t)} dt \nabla_{\beta_q} \mu_{ik} \right] \\ &= \mathbb{E}_Y \left[ \frac{Y_{ik} - \mu_{ik}}{\phi V(\mu_{ik})} \right] \nabla_{\beta_q} \mu_{ik} \\ &= \frac{\mathbb{E}_Y [Y_{ik}] - \mu_{ik}}{\phi V(\mu_{ik})} \nabla_{\beta_q} \mu_{ik} \\ &= 0, \quad \text{for all } q = 1, \dots, p_1.\end{aligned}$$

□

**Proposition 2.38.** *The census quasi-score function is given by*

$$\nabla_{\beta} Q^c(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{N_I} \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i), \quad (2.2)$$

where

$$\begin{aligned} \mathbf{D}_i^T &= (\nabla_{\boldsymbol{\beta}} \boldsymbol{\mu}_i)^T = \begin{pmatrix} (\nabla_{\beta_1} \boldsymbol{\mu}_i)^T \\ \vdots \\ (\nabla_{\beta_{p_1}} \boldsymbol{\mu}_i)^T \end{pmatrix} = \begin{pmatrix} \nabla_{\beta_1} \mu_{i1} & \cdots & \nabla_{\beta_1} \mu_{iN_i} \\ \vdots & & \vdots \\ \nabla_{\beta_{p_1}} \mu_{i1} & \cdots & \nabla_{\beta_{p_1}} \mu_{iN_i} \end{pmatrix}, \\ \mathbf{V}_i &= \phi \operatorname{diag} \{V(\mu_{i1}), \dots, V(\mu_{iN_i})\}, \\ \mathbf{Y}_i - \boldsymbol{\mu}_i &= \begin{pmatrix} Y_{i1} - \mu_{i1} \\ \vdots \\ Y_{iN_i} - \mu_{iN_i} \end{pmatrix}. \end{aligned}$$

*Proof.* Observe

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} Q^c(\boldsymbol{\mu}, \mathbf{y}) &= \begin{pmatrix} \nabla_{\beta_1} Q^c(\boldsymbol{\mu}, \mathbf{y}) \\ \vdots \\ \nabla_{\beta_{p_1}} Q^c(\boldsymbol{\mu}, \mathbf{y}) \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \nabla_{\beta_1} Q_{ik}(\mu_{ik}, y_{ik}) \\ \vdots \\ \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \nabla_{\beta_{p_1}} Q_{ik}(\mu_{ik}, y_{ik}) \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \frac{Y_{ik} - \mu_{ik}}{\phi V(\mu_{ik})} \nabla_{\beta_1} \mu_{ik} \\ \vdots \\ \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \frac{Y_{ik} - \mu_{ik}}{\phi V(\mu_{ik})} \nabla_{\beta_{p_1}} \mu_{ik} \end{pmatrix} \\ &= \sum_{i=1}^{N_I} (\nabla_{\boldsymbol{\beta}} \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{N_I} \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i). \end{aligned}$$

□

**Definition 2.39.** The sample quasi-likelihood function is defined to be

$$Q^s(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{N_I} 1_i \sum_{k \in U_i^c} 1_{k|i} Q_{ik}(\mu_{ik}, y_{ik}),$$



where

$$Q_{ik}(\mu_{ik}, y_{ik}) = \int_{y_{ik}}^{\mu_{ik}} \frac{y_{ik} - t}{\phi V(t)} dt.$$

**Lemma 2.40.** *The sample quasi-score function is an unbiased estimating equation with respect to model measure  $Y$ , i.e.,*

$$\mathbb{E}_Y [\nabla_{\beta} Q^s(\boldsymbol{\mu}, \mathbf{Y})] = \mathbf{0}.$$

*Proof.* Note the sample quasi-score function is a sum of single unbiased estimating equation, i.e.,

$$\mathbb{E}_Y [\nabla_{\beta} Q^s(\boldsymbol{\mu}, \mathbf{Y})] = \sum_{i=1}^{N_I} 1_i \sum_{k \in \mathcal{U}_i^c} 1_{k|i} \mathbb{E}_Y [\nabla_{\beta} Q_{ik}(\mu_{ik}, Y_{ik})] = \mathbf{0}.$$

The last equality follows from Lemma 2.37. □

Let  $\mathbf{y}_i^s, \mathbf{x}_i^s, \boldsymbol{\mu}_i^s, \mathbf{b}_i^s, \mathbf{V}_i^s, \mathbf{D}_i^s$  be the sample data corresponding to  $\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\mu}_i, \mathbf{b}_i, \mathbf{V}_i, \mathbf{D}_i$ .

**Definition 2.41.** Define the sample quasi-likelihood estimator  $\hat{\boldsymbol{\beta}}$  for  $\boldsymbol{\beta}$  to be the solution of the following sample quasi-score function

$$\nabla_{\beta} Q^s(\boldsymbol{\mu}, \mathbf{Y}) = \sum_{i=1}^{N_I} 1_i (\mathbf{D}_i^s)^T (\mathbf{V}_i^s)^{-1} (\mathbf{Y}_i^s - \boldsymbol{\mu}_i^s) = \mathbf{0}.$$

**Remark.** If  $\boldsymbol{\mu}_i^s = \mathbf{X}_i^s \boldsymbol{\beta}$ , then the sample quasi-likelihood estimator  $\hat{\boldsymbol{\beta}}$  is given by

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^{N_I} 1_i (\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \mathbf{X}_i^s \right)^{-1} \left( \sum_{i=1}^{N_I} 1_i (\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \mathbf{Y}_i^s \right).$$

This is also the weighted least square estimator. The variance estimator of  $\hat{\boldsymbol{\beta}}$  is given by

$$\begin{aligned} \text{Var}_Y(\hat{\boldsymbol{\beta}}) &= \left( \sum_{i=1}^{N_I} 1_i (\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \mathbf{X}_i^s \right)^{-1} \left( \sum_{i=1}^{N_I} 1_i (\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \text{Var}_Y(\mathbf{Y}_i^s) (\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \right) \\ &\quad \left( \sum_{i=1}^{N_I} 1_i (\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \mathbf{X}_i^s \right)^{-1}. \end{aligned}$$

The true variance  $\text{Var}_Y(\mathbf{Y}_i^s)$  is typically unknown, but one can estimate  $\text{Var}_Y(\mathbf{Y}_i^s)$  by the empirical variance  $\widehat{\text{Var}}_Y(\mathbf{Y}_i^s)$ . More precisely,

$$\widehat{\text{Var}}_Y(\hat{\beta}) = \left( \sum_{i=1}^{N_I} 1_i(\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \mathbf{X}_i^s \right)^{-1} \left( \sum_{i=1}^{N_I} 1_i(\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \widehat{\text{Var}}_Y(\mathbf{Y}_i^s) (\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \right) \left( \sum_{i=1}^{N_I} 1_i(\mathbf{X}_i^s)^T (\mathbf{V}_i^s)^{-1} \mathbf{X}_i^s \right)^{-1},$$

where  $\widehat{\text{Var}}_Y(\mathbf{Y}_i^s) = (\mathbf{Y}_i^s - \hat{\mu}^s)^T (\mathbf{Y}_i^s - \hat{\mu}^s)$ .

**Remark.** In general, there is no closed form formula for  $\hat{\beta}$  and numerical methods has to be used (Demidenko, 2013). More explicitly, let  $\nabla_{\beta} Q^s(\beta)$  be the sample quasi-score function, i.e.,

$$\nabla_{\beta} Q^s(\beta) = \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s)^{-1} (\mathbf{Y}_i^s - \mu_i^s).$$

By Fisher's iterated method, one has

$$\hat{\beta}^{(r+1)} = \hat{\beta}^{(r)} + \left[ -\mathbb{E}_Y \left[ \nabla_{\beta\beta}^2 Q^s(\hat{\beta}^{(r)}) \right] \right]^{-1} \nabla_{\beta} Q^s(\hat{\beta}^{(r)}),$$

where

$$\mathbb{E}_Y \left[ \nabla_{\beta\beta}^2 Q^s(\beta) \right] = - \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s)^{-1} \mathbf{D}_i^s.$$

**2.2.3.2 Generalized Estimating Equation (GEE)** Generalized estimating equation (GEE) is an extension of quasi-score function. Note  $\mathbf{V}_i = \phi \text{diag} \{V(\mu_{i1}), \dots, V(\mu_{iN_i})\}$  in 2.2 is diagonal in the quasi-score setting. Even if covariance matrix  $\mathbf{V}_i$  is misspecified, the quasi-likelihood estimator is still consistent, but it does affect the variance. To improve efficiency, one could consider modelling correlation structure by a working covariance matrix  $\mathbf{V}_i(\boldsymbol{\eta})$ . In other words, the working covariance matrix  $\mathbf{V}_i(\boldsymbol{\eta})$  is user-defined and non-diagonal. The parameter one wants to estimate is  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})^T$ . This is also known as the generalized estimating equation. More precisely,

**Definition 2.42.** The census GEE estimator  $\tilde{\theta}_N$  is defined as a solution of

$$\sum_{i=1}^{N_I} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\boldsymbol{\eta})(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where  $\mathbf{V}_i(\boldsymbol{\eta})$  is a user-defined working covariance matrix.

**Definition 2.43.** The sample GEE estimator  $\hat{\theta}_n$  is defined as a solution of

$$\sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s)^{-1}(\boldsymbol{\eta})(\mathbf{Y}_i^s - \boldsymbol{\mu}_i^s) = \mathbf{0},$$

where  $\mathbf{V}_i^s(\boldsymbol{\eta})$  is a user-defined working covariance matrix.

**Theorem 2.44.** Under certain regularity conditions, then

$$N_I^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{G}(\theta_0)),$$

where

$$\begin{aligned} \mathbf{G}(\theta_0) = \lim_{N_I \rightarrow \infty} N_I \left( \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s(\theta_0))^{-1} \mathbf{D}_i^s \right)^{-1} & \left[ \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s(\theta_0))^{-1} \text{Var}_Y(\mathbf{Y}_i^s) (\mathbf{V}_i^s(\theta_0))^{-1} (\mathbf{D}_i^s) \right] \\ & \left( \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s(\theta_0))^{-1} \mathbf{D}_i^s \right)^{-1}. \end{aligned}$$

*Proof.* This is proved in [Liang and Zeger \(1986\)](#). □

**Remark.** One can estimate  $\mathbf{G}(\theta_0)$  by the empirical variance

$$\begin{aligned} \hat{\mathbf{G}}(\hat{\theta}_n) = \lim_{N_I \rightarrow \infty} N_I \left( \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s(\hat{\theta}_n))^{-1} \mathbf{D}_i^s \right)^{-1} & \left[ \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s(\hat{\theta}_n))^{-1} \widehat{\text{Var}}_Y(\mathbf{Y}_i^s) (\mathbf{V}_i^s(\hat{\theta}_n))^{-1} (\mathbf{D}_i^s) \right] \\ & \left( \sum_{i=1}^{N_I} 1_i(\mathbf{D}_i^s)^T (\mathbf{V}_i^s(\hat{\theta}_n))^{-1} \mathbf{D}_i^s \right)^{-1}, \end{aligned} \quad (2.3)$$

where  $\widehat{\text{Var}}_Y(\mathbf{Y}_i^s) = (\mathbf{Y}_i^s - \widehat{\boldsymbol{\mu}}^s)^T (\mathbf{Y}_i^s - \widehat{\boldsymbol{\mu}}^s)$ .

**Remark.** Variance estimation in 2.3 is given by a sandwich form, which is very similar to variance estimation for the weighted pairwise likelihood estimator under a two-stage sampling design (more about this in chapter 3 and 4).

## 2.2.4 Relationship between marginal and conditional model

The relationship between the marginal and conditional model has been discussed extensively in the literature (Zeger et al., 1988; Neuhaus et al., 1991). I would like to comment on the difference between the marginal and conditional model for the logistic regression model. In particular, I want to show they are not equivalent. Let  $Y_{ik}$  be a binary random variable.

(i) Consider a conditional logistic model

$$\begin{aligned} Y_{ik}|p_{ik} &\sim \text{Bernoulli}(p_{ik}), \\ \text{logit}(p_{ik}) &= \beta_0 + \beta_1 x_{ik} + b_i, \end{aligned}$$

then the conditional mean is given by

$$\mathbb{E}_Y[Y_{ik}|b_i] = \frac{\exp(\beta_0 + \beta_1 x_{ik} + b_i)}{1 + \exp(\beta_0 + \beta_1 x_{ik} + b_i)}.$$

Hence, the marginal mean is given by

$$\mathbb{E}_Y[Y_{ik}] = \mathbb{E}_Y[\mathbb{E}_Y[Y_{ik}|b_i]] = \mathbb{E}_Y \left[ \frac{\exp(\beta_0 + \beta_1 x_{ik} + b_i)}{1 + \exp(\beta_0 + \beta_1 x_{ik} + b_i)} \right]. \quad (2.4)$$

(ii) Consider a marginal logistic model

$$\text{logit}(p_{ik}) = \gamma_0 + \gamma_1 x_{ik},$$

then the marginal mean is given by

$$\mathbb{E}_Y[Y_{ik}] = p_{ik} = \frac{\exp(\gamma_0 + \gamma_1 x_{ik})}{1 + \exp(\gamma_0 + \gamma_1 x_{ik})}. \quad (2.5)$$

Observe 2.5 is not equivalent to 2.4 as the expectation is a linear operator and integrand is nonlinear in  $b_i$ .

**Remark.** It can be shown that estimation from marginal and conditional approach are the same for linear and Poisson model. For more detail, see Demidenko (2013); Grömping (1996). But this is not true for logistic model.

## 2.3 Model–design–based approach

The goal in the model–design–based approach is to construct estimator which incorporates the feature of the sampling design to estimate unknown parameters  $\theta$  from the super-population. Roughly speaking, this can be decomposed into two steps:

- (i) Inference from the sample data to the finite population.
- (ii) Inference from the finite population to the model.

There are two probability measures in this setting: the design measure  $\mathbb{P}_\pi$  (the values of the sample inclusion indicators 1 for each observational unit is random) and the model measure  $\mathbb{P}_Y$  (the values of the model variable  $Y$  for each observational unit is random). I will use  $\mathbb{E}_{Y\pi}$  for the expectation,  $\text{Var}_{Y\pi}$  for the variance and  $\text{Cov}_{Y\pi}$  for the covariance under the model–design–based measure.

I would like to discuss informally how to think about the framework. The design-based framework works conditional with  $1|Y$  and model-based framework works conditional with  $Y|1$ . The model-design-based framework mean working jointly in  $(Y, 1)$  rather than conditionally. Since design-expectations are always taken first, the basic approach is to first conditional on  $Y$  and then remove the conditioning by law of iterated expectation, i.e.,  $\mathbb{E}_{Y\pi} = \mathbb{E}_Y \mathbb{E}_{\pi|Y}$ . The model-design variance is given by conditional on  $Y$  first, then remove the conditioning by law of conditional variance, i.e.,  $\text{Var}_{Y\pi} = \text{Var}_Y \mathbb{E}_{\pi|Y} + \mathbb{E}_Y \text{Var}_{\pi|Y}$ .

In this section, I assume the sampling clusters are the model clusters, i.e.,  $U_1^c = M_1^c$ .

### 2.3.1 Asymptotic setting

I want to take the limit of an estimator. In particular, I want to talk about the notion of convergence in probability and convergence in distribution. With the structure I introduced so far, I cannot define those notions. This is because the sample size  $n$  is bounded by population size  $N$ , which is assumed to be finite. To formally define those notions, one needs to construct infinite produce spaces and extend probability on it. But for what I have to do in this thesis, this level of mathematical formality is unnecessary. I shall

not elaborate the details of this construction, as this has been discussed in many papers (Isaki and Fuller, 1982; Rubin-Bleuer and Kratina, 2005).

In all of the following, I assume both the sample and population size need to diverge and the sampling fraction should converge, i.e.,  $n \rightarrow \infty$ ,  $N \rightarrow \infty$  and  $\frac{n}{N} \rightarrow c$ , where  $c \in [0, 1)$ . Let  $\theta_0$  be the true value.

**Definition 2.45.** Given a superpopulation model, an estimator  $\hat{\theta}_n$  is design consistent if for almost all model realisation  $\omega$  from the superpopulation and for all  $\epsilon > 0$ , there exists a  $n_0 \in \mathbb{N}$  such that for all  $n > n_0$ , one has

$$\mathbb{P}_\pi \left( \|\hat{\theta}_n(\omega) - \theta_0\| > \epsilon \right) < \epsilon.$$

**Remark.** It can be shown that HT estimator are design consistent (Isaki and Fuller, 1982).

**Definition 2.46.** Estimator  $\hat{\theta}_n$  is model consistent if for all  $\epsilon > 0$ , there exists a  $n_0 \in \mathbb{N}$  such that for all  $n > n_0$ , one has

$$\mathbb{P}_Y(\|\hat{\theta}_n - \theta_0\| > \epsilon) < \epsilon.$$

**Remark.** Finite-dimensional MLE estimator are model consistent under certain regularity conditions. For more details, see chapter 5 in Van der Vaart (2000).

### 2.3.2 Informative sampling

There are several possible definitions of the informative sampling in the literature. I will use the following one in this thesis (Pfeffermann et al., 1998; Rubin-Bleuer and Kratina, 2005).

**Definition 2.47.** Sampling design is uninformative if design variable is independent of the response variable after conditioning on the model variable, i.e.,

$$1_U \perp Y_U | X_U.$$

If sampling design is not uninformative, then one shall call it informative.

**Remark.** The definition of the informative sampling can be easily extended to the two-stage sampling (Pfeffermann et al., 1998).

**Example 3.** Suppose sampling inclusion probability is given by

$$\begin{aligned}\pi : U &\longrightarrow [0, 1] \\ k &\longmapsto \pi(k) = \frac{1}{1 + e^{-y_k}},\end{aligned}$$

then  $\pi(k)$  is an increasing function of  $y_k$ . Hence, observational units with a larger value of  $y$  are oversampled.

What is significant in here is ignoring the sampling design and just fitting the mixed model to the sampling distribution will lead to biased inference (Pfeffermann and Sverchkov, 2009). To see this, let  $f_s(y, \mathbf{x})$  be the sampling distribution and  $f_p(y, \mathbf{x})$  be the population distribution.

**Lemma 2.48.** Under the informative sampling design, the model holding for the sample is different from the model holding for the population, i.e.,

$$f_s(y_k, \mathbf{x}_k) \neq f_p(y_k, \mathbf{x}_k).$$

*Proof.* Observe

$$\begin{aligned}f_s(y_k, \mathbf{x}_k) &= f_p(y_k, \mathbf{x}_k | 1_k = 1) \\ &= \frac{f_p(y_k, \mathbf{x}_k, 1_k = 1)}{f_p(1_k = 1)} \\ &\neq \frac{f_p(y_k, \mathbf{x}_k) f_p(1_k = 1)}{f_p(1_k = 1)} \\ &= f_p(y_k, \mathbf{x}_k).\end{aligned}$$

□

There are many possible solutions in here. One could add design variables in the the model, but there are two problems with this approach. First, we don't want the sampling design (which might be chose by someone else) to affect the choice of parameters that we estimate. Hence we prefer not to directly model the design variables. Too many design variables reduce the power of the model. Secondly, design variable  $Z$  may be a mediator variable, i.e., design variable  $Z$  may be correlated with both the response variable  $Y$  and model variable  $X$ . Then one cannot distinguish the effect of model variable  $X$  on the response variable  $Y$ . See the figure below.



Figure 2.1: Mediator variable Z (left) and confounding variable Z (right) .

Alternatively, one could incorporate the sampling weights into the likelihood, producing a weighted likelihood, where the weight is given by the inverse sampling inclusion probability, i.e.,  $w_k = 1/\pi_k$ . This is the approach I will take for the rest of thesis. The rational of this construction is the weighted likelihood is an unbiased estimator for the census likelihood (more about this later).

**Remark.** In the literature, there are other ways to construct the weights such as calibration to handle more sophisticated setting such as incomplete data or missing data. For more detail, see [Gelman \(2007\)](#); [Kim and Shao \(2013\)](#); [Kim and Park \(2010\)](#); [Lumley and Scott \(2017\)](#); [Grace \(2016\)](#).

### 2.3.3 Full-likelihood approach

In this section, I will briefly mention full-likelihood approach. [Scott and Wild \(1997, 2001\)](#) show how to construct full-likelihood for the logistic regression when the sampling clusters are the model clusters.

The full-likelihood is much more complicated and is not tractable when the sampling clusters are not equal the model clusters. In the literature, no one has described how to construct full-likelihood. This is because it lacks the conditional independent, i.e., observational units in the same sampling cluster might not be in the same model cluster.

Consider a random intercept model with informative Poisson sampling at stage 1 and simple random sampling at stage 2. Assume the sampling clusters are equal to model



clusters, the sampling is independent across the clusters. Let  $g(b_i, \theta)$  be the density of  $b_i$  and  $f(y_{ik}, \theta|b_i)$  be the conditional density of  $y_{ik}$ . The population likelihood is

$$\prod_{i=1}^{N_I} \int g(b_i, \theta) \prod_{k \in U_i^c} f(y_{ik}, \theta|b_i) db_i.$$

Let  $g_R$  be the sampling likelihood at stage 1. Then the likelihood of the data is proportional to

$$\prod_{i=1}^{N_I} \int g_R(b_i, \theta) \prod_{k \in U_i^s} f(y_{ik}, \theta|b_i) db_i.$$

We have a product of one-dimensional numerical integrals for the full likelihood to maximise, which is completely feasible. The sample likelihood is also tractable.

When the sampling clusters are not the same as the model clusters, the sampling likelihood  $g_R$  for a particular sample cluster depends on the  $b_i$  and  $Y_{ij}$  for all model clusters that intersect it, so the product over  $i$  cannot simply be taken outside the integral. In the simulation setting in the thesis ( $100 \times 100$ ) with overlap of 0.6, 40  $b_i$  contribute to each sampling probability so even under Poisson sampling we have 40-dimensional numerical integrals to compute the likelihood. This is just intractable, which is why we want to use pseudo-likelihood.

### 2.3.4 Weighted estimating equation

In this section, I will briefly mention how to construct an unbiased sample weighted estimating equation when the census estimating equation can be written as a linear combination of individual estimating equation. In general, it is impossible to construct an unbiased sample weighted estimating equation when the census estimating equation is non-linear. It happens quite often and is not just remote mathematical territory. For example, normal linear mixed model. What then can we do? I will come back to this question in chapter 3, which is why I want to study the pseudo-likelihood estimation.

Let  $\psi^c(\theta)$  be a census estimating equation and suppose  $\psi^c(\theta)$  can be written as a

linear combination of individual estimating equation, i.e.,

$$\psi^c(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} \sum_{k \in U_i^c} \psi_{ik}(y_{ik}). \quad (2.6)$$

Then the Horvitz-Thompson estimator for  $\psi^c(\boldsymbol{\theta})$  is given by

$$\psi^s(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} 1_i \omega_i \sum_{k \in U_i^c} 1_{k|i} \omega_{k|i} \psi_{ik}(y_{ik}).$$

Observe  $\mathbb{E}_\pi [\psi^s(\boldsymbol{\theta})] = \psi^c(\boldsymbol{\theta})$ .

In all of the following, let  $\hat{\boldsymbol{\theta}}_n$  be the sample weighted estimator and  $\boldsymbol{\theta}_0$  be the true value of the model. More precisely,

**Definition 2.49.** The sample weighted estimator  $\hat{\boldsymbol{\theta}}_n$  of  $\boldsymbol{\theta}$  is defined as a solution of

$$\frac{1}{N_I} \psi^s(\boldsymbol{\theta}) = 0.$$

**Definition 2.50.** The census estimator  $\tilde{\boldsymbol{\theta}}_N$  of  $\boldsymbol{\theta}$  is defined as a solution of

$$\frac{1}{N_I} \psi^c(\boldsymbol{\theta}) = 0.$$

**Remark.** It can be shown that the true parameter  $\boldsymbol{\theta}_0$  of  $\boldsymbol{\theta}$  is a solution of

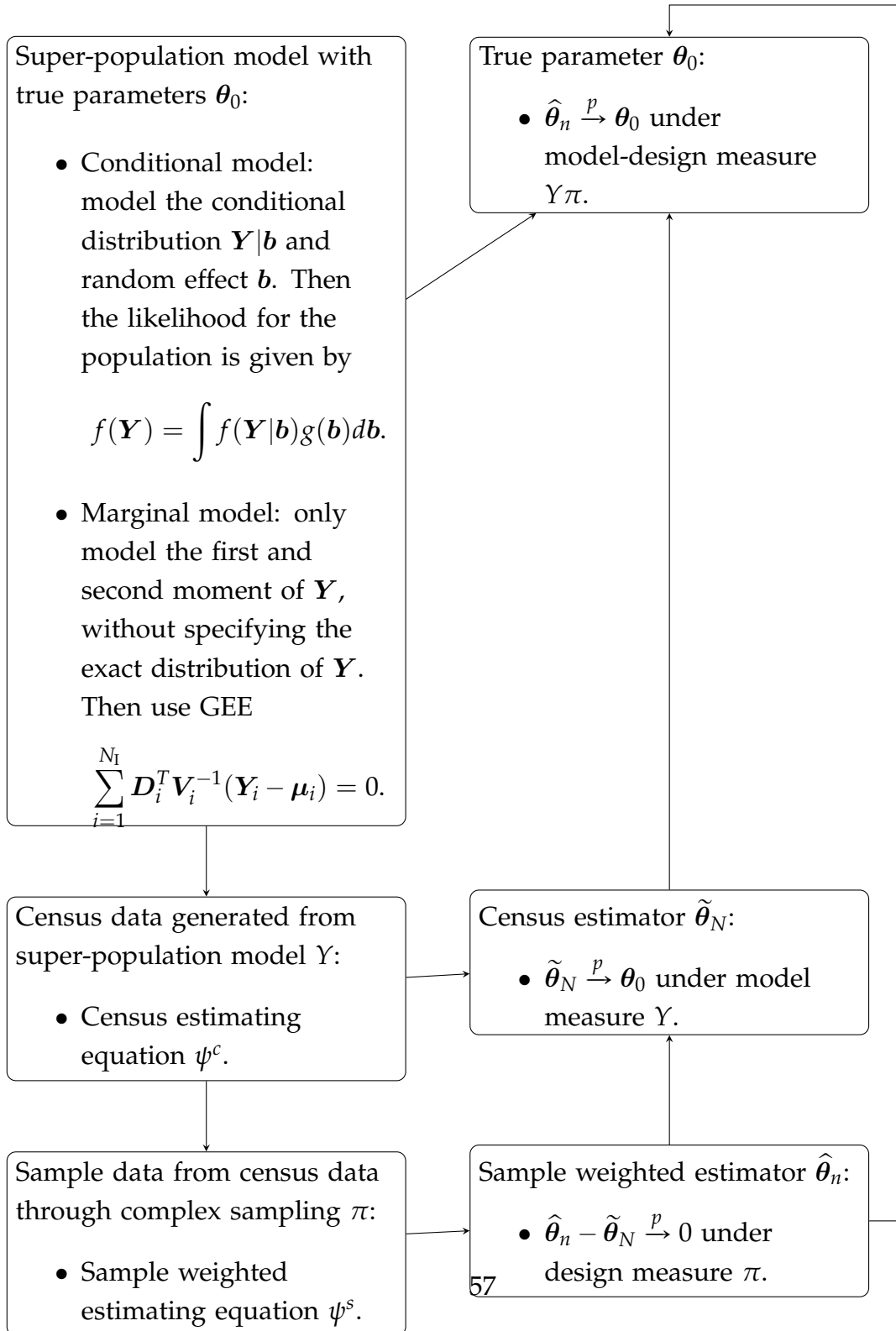
$$\mathbb{E}_Y \left[ \frac{1}{N_I} \psi^c(\boldsymbol{\theta}) \right] = 0,$$

where  $\psi^c$  is the census likelihood score. In other words,  $\frac{1}{N_I} \psi^c(\boldsymbol{\theta}_0) = 0$  is an unbiased estimating equation.

Under regularity conditions, as  $N \rightarrow \infty, n \rightarrow \infty$  and  $\frac{n}{N} \rightarrow c$ , one can show the sample weighted estimator converges to the census estimator and census estimator converges to the true value by classical smooth argument as in chapter 5 of [Van der Vaart \(2000\)](#), i.e.,

$$\begin{aligned} \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_N &\xrightarrow{p} 0 \quad \text{under design measure } \pi, \\ \tilde{\boldsymbol{\theta}}_N &\xrightarrow{p} \boldsymbol{\theta}_0 \quad \text{under model measure } Y. \end{aligned}$$

See the figure below.



Let me summarise the key step in here, as I will apply such argument without further comment.

- (i) Write down the census estimating equation as a linear combination of individual estimating equation.
- (ii) Add weight to form an unbiased sample weighted estimating equation.
- (iii) Reduce to a standard argument to show the estimator is consistent and asymptotic normality, as  $N \rightarrow \infty, n \rightarrow \infty$  and  $\frac{n}{N} \rightarrow c$ , where  $c \in [0, 1)$ .

But there are still two important questions remaining.

- (i) How can we handle the non-linear case, i.e., the census estimating equation cannot be written as a linear combination of individual estimating equation.
- (ii) How can we add the weight for the non-linear case to form an unbiased sample weighted estimating equation? Note consistency of the estimation relied on an unbiased estimating equation.

This is essentially impossible to construct an unbiased estimating equation, as the expectation is a linear operator and integrand is nonlinear. I will propose alternative solution in chapter 3. My short answer is to transform a nonlinear case into a linear case, i.e., instead of considering full-likelihood, considering pseudo-likelihood so that the weight is linear in the sample weighted estimating equation. Outside complex sampling, the rationale for using pseudo-likelihood is to avoid strong distributional assumption and reduce computational burden (Heagerty and Lele, 1998). With complex sampling, pseudo-likelihood is used for a completely different reason. Namely, we want to construct an unbiased estimating equation, so that weight is linear in the sample weighted estimating equation (more about this in chapter 3).

I want to come back to the non-linear case and show an example one could potentially encounter. In general, the census estimating equation cannot always be written as a linear combination of individual estimating equation as in 2.6. In that case, it is not clear how to construct the sample likelihood. This is not just remote mathematical territory and it happens quite often. For example, normal linear mixed model.

Consider a linear mixed model in section 2.2.2. The census full-likelihood is given by

$$\ell^c(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{N_I} \log |\mathbf{V}_i| - \frac{1}{2} \sum_{i=1}^{N_I} (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta}),$$

where  $\mathbf{V}_i(\sigma^2, \boldsymbol{\eta}) = \sigma^2 \mathbf{I} + \mathbf{Z}_i \mathbf{D}(\boldsymbol{\eta}) \mathbf{Z}_i^T$ .

**Lemma 2.51.** Suppose  $\mathbf{V}_i$  can be decomposed into a  $2 \times 2$  block matrices

$$\mathbf{V}_i = \left( \begin{array}{c|c} \mathbf{V}_i^s & \mathbf{A}_i \\ \hline \mathbf{A}_i^T & \mathbf{B}_i \end{array} \right),$$

where  $\mathbf{V}_i^s$  is the sample variance, then

$$\mathbf{V}_i^{-1} = \left( \begin{array}{c|c} (\mathbf{V}_i^{-1})_s & \cdot \\ \hline \cdot & \cdot \end{array} \right), \quad (2.7)$$

$$|\mathbf{V}_i^s| = |\mathbf{V}_i^s \mathbf{B}_i - \mathbf{A}_i^T \mathbf{A}_i|. \quad (2.8)$$

where  $(\mathbf{V}_i^{-1})_s = (\mathbf{V}_i^s - \mathbf{A}_i \mathbf{B}_i^{-1} \mathbf{A}_i^T)^{-1}$ .

*Proof.* For more detail, see [Lu and Shiou \(2002\)](#). □

With complex sampling, the elements of  $(\mathbf{V}_i^{-1})_s$  is not available, where  $(\mathbf{V}_i^{-1})_s$  is the sample block of the inverse of covariance. Hence it is not clear how to construct an unbiased sample log-pseudo-likelihood. To see this, one could try naive sample log-likelihood

$$\ell^s(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{N_I} 1_i \log |\mathbf{V}_i^s| - \frac{1}{2} \sum_{i=1}^{N_I} 1_i (\mathbf{y}_i^s - \mathbf{x}_i^s \boldsymbol{\beta})^T (\mathbf{V}_i^{-1})_s (\mathbf{y}_i^s - \mathbf{x}_i^s \boldsymbol{\beta}).$$

But it is impossible to calculate  $(\mathbf{V}_i^{-1})_s$  and  $|\mathbf{V}_i^s|$  based on the sample data as shown by Lemma 2.51. It is hopeless to apply formula 2.7 and 2.8 as block matrices  $\mathbf{A}_i, \mathbf{B}_i$  involve the census data  $\mathbf{Z}_i$ .

### 3 Maximum pseudo-likelihood

In this chapter, I want to study the pseudo-likelihood estimation with complex sampling. More precisely, I discuss a specific pseudo-likelihood construction, namely the pairwise composite likelihood. With complex sampling, one has to add weight in the pairwise likelihood. My goal in this chapter is to review the key results from Rao et al. (2013); Yi et al. (2016) in the setting of the two-stage sampling where the sampling clusters are the model clusters. More precisely, I am going to present proof done by Rao et al. (2013); Yi et al. (2016) in detail, which is a starting point in the literature of the weighted pairwise likelihood estimator under complex sampling. In particular, the consistency and variance estimator of the weighted pairwise likelihood estimator will be presented.

#### 3.1 Motivation

Roughly speaking, a pseudo-likelihood is constructed by modifying a true likelihood function to get a more tractable objective function. In particular, there is no assumption the data are generated from this pseudo-likelihood. Without complex sampling, there are two main reasons for such approach (Heagerty and Lele, 1998). One is to avoid strong distributional assumption. The other is to reduce the computational burden. With complex sampling, pseudo-likelihood is used for a completely different reason. Namely, one wants to construct an unbiased estimation equation, so that weight is linear in the sample weighted estimating equation. The pairwise likelihood is a special case of the pseudo-likelihood estimation. In the literature, the pseudo-likelihood is first developed by Besag (1974). In the modern sense, the pseudo-likelihood refers to any functions that are modification from a true likelihood. Composite likelihood is constructed from the sum of marginal likelihood or conditional likelihood (Lindsay, 1988). In particular, the pairwise composite likelihood is constructed from the sum of pairwise likelihood. Observe

$$\{\text{Pairwise likelihood}\} \in \{\text{Composite likelihood}\} \subset \{\text{Pseudo-likelihood}\}.$$

In all of the following, I will exclusively study the pairwise likelihood.

### 3.2 Setting

In this chapter, I will always work under the following setting. Suppose the two-level super-population model is given by

$$\begin{aligned} Y_{ik}|X_{ik}, \mathbf{b}_i &\sim f(y_{ik}|x_{ik}, \mathbf{b}_i, \boldsymbol{\theta}_1), \\ \mathbf{b}_i &\sim g(\mathbf{b}_i|\boldsymbol{\theta}_2), \end{aligned}$$

for  $i = 1, \dots, N_I$  and  $k = 1, \dots, N_i$ . This is more general setting than GLMM, as the conditional model does not have to be an exponential family and the random effects do not have to be normal random vector. The paper by [Rao et al. \(2013\)](#); [Yi et al. \(2016\)](#) is devoted to studying the weighted pairwise likelihood under such setting and I follow their approaches very closely in this chapter.

Let  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iN_i})^T$ ,  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_{N_I})^T$  and  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ . Define  $\boldsymbol{\theta}_0$  to be the true value for the super-population model. Assume  $\boldsymbol{\theta}_0$  is an interior point of a compact set  $\boldsymbol{\Theta}$ . In all of this chapter, I assume the sampling clusters are the model clusters, i.e.,  $U_I^c = M_I^c$ .

### 3.3 Pairwise composite likelihood estimation without complex sampling

Consider the pairwise composite likelihood without complex sampling, i.e., one observes all the census data. The idea is to replace the census log-full-likelihood  $\ell_i(y_{i1}, \dots, y_{iN_i})$  for each cluster  $i$  by the sum of all possible pairwise log-likelihood  $p\ell(y_{ik}, y_{il})$  in that cluster. Observe correlation information about observational units  $k$  and  $l$  in cluster  $i$  is captured in the pairwise log-likelihood  $p\ell(y_{ik}, y_{il})$ . More precisely,

**Definition 3.1.** Define the census pairwise log-likelihood for cluster  $i$  to be

$$p\ell_i^c(\boldsymbol{\theta}) = \sum_{\substack{k < l \\ k, l \in U_i^c}} p\ell_{kl|i}(\boldsymbol{\theta}),$$

where

$$p\ell_{kl|i}(\boldsymbol{\theta}) = \log \left[ \int f(y_{ik}|x_{ik}, \mathbf{b}_i, \boldsymbol{\theta}_1) f(y_{il}|x_{il}, \mathbf{b}_i, \boldsymbol{\theta}_1) g(\mathbf{b}_i|\boldsymbol{\theta}_2) d\mathbf{b}_i \right]. \quad (3.1)$$

Then the census pairwise log-likelihood is given by summing over all possible clusters, i.e.,

$$p^{\ell^c}(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} p^{\ell_i^c}(\boldsymbol{\theta}).$$

**Remark.** There are many ways of constructing the pairwise log-likelihood by imposing conditions on which pair one chooses in each cluster. I took all possible pairs in the same cluster as my definition. This might not always be a optimal choice. In fact, one can show that choosing all pairs can reduce efficiency, even if all the pairwise likelihood functions are independent (Xu, 2012). This is still an open question in the literature on how to construct the optimal pairs and how or whether to weight them. For what I have to do in this thesis, it suffices to consider all possible pairs in the same model clusters.

**Remark.** The integral in 3.1 involves a high-dimensional integral and so is intractable when the dimension of random effects  $\mathbf{b}_i$  is large. Approximation of integration based on the Laplace approximation is discussed in Breslow and Clayton (1993); Pinheiro and Chao (2006).

**Remark.** Observe the pairwise likelihood and full likelihood are exact the same when the cluster are of size 2. Note  $p^{\ell_i^c}$  is independent of  $p^{\ell_{i'}^c}$ , for  $i \neq i'$  under the model measure  $Y$ .

**Definition 3.2.** The census pairwise log-likelihood estimator  $\tilde{\boldsymbol{\theta}}_N$  of  $\boldsymbol{\theta}$  is defined as a solution of

$$\frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) = \mathbf{0}.$$

**Remark.** Consistency and asymptotical normality of the pairwise log-likelihood estimator without complex sampling have been established in many papers (Xu, 2012; Cox and Reid, 2004).



### 3.4 Weighted pairwise composite likelihood estimation with complex sampling

#### 3.4.1 Construction

With complex sampling, the population model does not hold for the sample (Pfeffermann, 1996). To estimate the census pairwise log-likelihood from sample data, one needs to add weight to account for informative sampling.

**Definition 3.3.** Define the sample weighted pairwise log-likelihood for cluster  $i$  to be

$$p\ell_i^s(\boldsymbol{\theta}) = \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} p\ell_{kl|i}(\boldsymbol{\theta}),$$

where

$$p\ell_{kl|i}(\boldsymbol{\theta}) = \log \left[ \int f(y_{ik}|x_{ik}, \mathbf{b}_i, \boldsymbol{\theta}_1) f(y_{il}|x_{il}, \mathbf{b}_i, \boldsymbol{\theta}_1) g(\mathbf{b}_i|\boldsymbol{\theta}_2) d\mathbf{b}_i \right].$$

The sample weighted pairwise log-likelihood is given by

$$p\ell^s(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} 1_i \omega_i p\ell_i^s(\boldsymbol{\theta}).$$

**Lemma 3.4.** Let  $h_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta}\}} \|\nabla_{\boldsymbol{\theta}} p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i, k, l\}} \mathbb{E}_Y \left[ h_{kl|i}^{1+\delta} \right] < \infty$ , then the sample weighted pairwise log-likelihood is a design-unbiased estimator for the census pairwise log-likelihood under design measure  $\pi$ , i.e.,

$$\mathbb{E}_{\pi} [p\ell^s(\boldsymbol{\theta})] = p\ell^c(\boldsymbol{\theta}).$$

*Proof.* Observe

$$\begin{aligned}
\mathbb{E}_\pi [p^{\ell^s}(\boldsymbol{\theta})] &= \mathbb{E}_{\pi_1, \pi_2} \left[ \sum_{i=1}^{N_I} 1_i \omega_i p^{\ell_i^s}(\boldsymbol{\theta}) \right] \\
&= \mathbb{E}_{\pi_1} \left[ \mathbb{E}_{\pi_2} \left[ \sum_{i=1}^{N_I} 1_i \omega_i p^{\ell_i^s}(\boldsymbol{\theta}) | U_I^s \right] \right] \\
&= \sum_{i=1}^{N_I} \omega_i \mathbb{E}_{\pi_1} [1_i \mathbb{E}_{\pi_2} [p^{\ell_i^s}(\boldsymbol{\theta}) | U_I^s]] \\
&= \sum_{i=1}^{N_I} \omega_i \mathbb{E}_{\pi_1} [1_i p^{\ell_i^c}(\boldsymbol{\theta})] \\
&= \sum_{i=1}^{N_I} p^{\ell_i^c}(\boldsymbol{\theta}) \\
&= p^{\ell^c}(\boldsymbol{\theta}).
\end{aligned}$$

□

**Corollary 3.5.** Let  $h_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta}\}} \|\nabla_{\boldsymbol{\theta}} p^{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})\|$  and  $g_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta}\}} \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y \left[ h_{kl|i}^{1+\delta} \right] < \infty$  and  $\sup_{\{i,k,l\}} \mathbb{E}_Y \left[ g_{kl|i}^{1+\delta} \right] < \infty$ , then the sample weighted pairwise score function is a design-unbiased estimator for the census pairwise score function under the design measure  $\pi$ , i.e.,

$$\begin{aligned}
\mathbb{E}_\pi [\nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta})] &= \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}), \\
\mathbb{E}_\pi [\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^s}(\boldsymbol{\theta})] &= \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}).
\end{aligned}$$

**Lemma 3.6.** Let  $h_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta}\}} \|\nabla_{\boldsymbol{\theta}} p^{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y \left[ h_{kl|i}^{1+\delta} \right] < \infty$ , then

$$\mathbb{E}_{Y\pi} [\nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0)] = \mathbf{0}.$$

*Proof.* By Corollary 3.5, one has

$$\begin{aligned}
\mathbb{E}_{Y\pi} [\nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0)] &= \mathbb{E}_Y [\nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}_0)] \\
&= \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \mathbb{E}_Y [\nabla_{\boldsymbol{\theta}} p^{\ell_{kl|i}}(\boldsymbol{\theta}_0)] \\
&= \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \int \frac{\nabla_{\boldsymbol{\theta}} f(y_{ik}, y_{il}, \boldsymbol{\theta}_0)}{f(y_{ik}, y_{il}, \boldsymbol{\theta}_0)} f(\mathbf{y}_1, \dots, \mathbf{y}_{i-1}, y_{i1}, \dots, y_{i(k-1)}, \\
&\quad y_{ik}, y_{i(k+1)}, \dots, y_{i(l-1)}, y_{il}, y_{i(l+1)}, \dots, y_{iN_I}, \mathbf{y}_{i+1}, \dots, \mathbf{y}_{N_I}, \boldsymbol{\theta}_0) dy_1 \dots dy_{i-1} \\
&\quad dy_{i1} \dots dy_{i(k-1)} dy_{ik} dy_{i(k+1)} \dots dy_{i(l-1)} dy_{il} dy_{i(l+1)} \dots dy_{iN_I} dy_{i+1} \dots dy_{N_I} \\
&= \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \int \frac{\nabla_{\boldsymbol{\theta}} f(y_{ik}, y_{il}, \boldsymbol{\theta}_0)}{f(y_{ik}, y_{il}, \boldsymbol{\theta}_0)} f(y_{ik}, y_{il}, \boldsymbol{\theta}_0) dy_{ik} dy_{il} \\
&= \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \nabla_{\boldsymbol{\theta}} \int f(y_{ik}, y_{il}, \boldsymbol{\theta}_0) dy_{ik} dy_{il} \\
&= \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \nabla_{\boldsymbol{\theta}} 1 \\
&= \mathbf{0}.
\end{aligned}$$

□

**Example 4** (Pairwise likelihood estimator for linear mixed model). Consider a linear mixed model defined in section 2.2.2. I work with the pairwise composite likelihood instead. The census pairwise likelihood is given by

$$p^{\ell^c}(\boldsymbol{\theta}) = \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \left[ -\frac{1}{2} \log |\boldsymbol{\Sigma}_{kl}| - \frac{1}{2} (\mathbf{y}_{kl} - \mathbf{x}_{kl}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{kl}^{-1} (\mathbf{y}_{kl} - \mathbf{x}_{kl}\boldsymbol{\beta}) \right],$$

where  $\boldsymbol{\Sigma}_{kl}$  are now  $2 \times 2$  variance-covariance matrices for pair  $k, l$ . Then one can show the census

pairwise likelihood estimators of  $\beta$  is given by

$$\tilde{\beta} = \left( \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \mathbf{x}_{kl}^T \Sigma_{kl}^{-1} \mathbf{x}_{kl} \right)^{-1} \left( \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} \mathbf{x}_{kl}^T \Sigma_{kl}^{-1} \mathbf{y}_{kl} \right).$$

The sample weighted pairwise likelihood is given by

$$p^{\ell^s}(\theta) = \sum_{i \in U_1^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \left[ -\frac{1}{2} \log |\Sigma_{kl}| - \frac{1}{2} (\mathbf{y}_{kl} - \mathbf{x}_{kl}\beta)^T \Sigma_{kl}^{-1} (\mathbf{y}_{kl} - \mathbf{x}_{kl}\beta) \right].$$

One can show the sample weighted pairwise likelihood estimators of  $\beta$  is given by

$$\hat{\beta} = \left( \sum_{i \in U_1^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \mathbf{x}_{kl}^T \Sigma_{kl}^{-1} \mathbf{x}_{kl} \right)^{-1} \left( \sum_{i \in U_1^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \mathbf{x}_{kl}^T \Sigma_{kl}^{-1} \mathbf{y}_{kl} \right).$$

### 3.4.2 Consistency

I now turn to the problem I stated in the beginning of this chapter: when is the pairwise likelihood estimator a consistent estimator? I am going to present proof done by [Yi et al. \(2016\)](#), which is a starting point in the literature of the weighted pairwise likelihood estimator under complex sampling, but I will include more detail. I am going to use a similar approach for the more general case (more about this in chapter 4). Let  $\hat{\theta}_n$  be the sample weighted pairwise likelihood estimator and  $\theta_0$  be the true value of the model. More precisely,

**Definition 3.7.** The sample weighted pairwise likelihood estimator  $\hat{\theta}_n$  of  $\theta$  is defined as a solution of

$$\frac{1}{N_I} \nabla_{\theta} p^{\ell^s}(\theta) = \mathbf{0}.$$

**Remark.** It can be shown that the true parameter  $\theta_0$  of  $\theta$  is a solution of

$$\mathbb{E}_Y \left[ \frac{1}{N_I} \nabla_{\theta} p^{\ell^c}(\theta) \right] = \mathbf{0}.$$

In all of the following, I assume the elements within the cluster are bounded, both the sample and population clusters need to diverge and the sampling fraction for the cluster should converge, i.e.,  $N_i \leq \lambda$  for all  $i$ ,  $n_I \rightarrow \infty$ ,  $N_I \rightarrow \infty$  and  $\frac{n_I}{N_I} \rightarrow c$ , where  $\lambda > 0$  and  $c \in [0, 1)$ .

**Theorem 3.8.** *Under the following regularity conditions,*

**A.1**  $\Theta$  is a compact subset of  $\mathbb{R}^p$  and  $\theta_0$  is an interior point of  $\Theta$ .

**A.2** Let  $h_{kl|i}(Y_{ik}, Y_{il}) = \sup_{\theta \in \Theta} \|\nabla p_{\ell_{kl|i}}(\theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y h_{kl|i}^{1+\delta} < \infty$  and  $\sup_i \mathbb{E}_Y \|\mathbf{Y}_i\|^\delta < \infty$ , where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iN_i})^\top$ .

**A.3** For any given  $c > 0$  and a given sequence  $\{\mathbf{y}_i\}$  satisfying  $\|\mathbf{y}_i\| \leq c$ , the sequence of function  $\{\nabla_{\theta} p_{\ell_{kl|i}}(\theta)\}$  is equicontinuous on any open subset  $A$  of  $\Theta$ .

**A.4** For all variables  $V_{kl|i}$  satisfy  $\frac{1}{N_I} \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} V_{kl|i}^2 = O_p(1)$ , one has

$$\frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} V_{kl|i} - \frac{1}{N_I} \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} V_{kl|i} = O_p(n_I^{-\frac{1}{2}})$$

with respect to design probability  $\pi$ .

**A.5** The number of elements within any clusters is bounded, i.e.,  $\sup_i N_i \leq \lambda$  for some  $\lambda > 0$ .

**A.6** For all  $\epsilon > 0$ , there exists a  $\delta > 0$  such that

$$\inf_{\{\theta \in \Theta: \|\theta - \theta_0\| \geq \epsilon\}} \left\| \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla_{\theta} p_{\ell^s}(\theta) \right] \right\| > \delta.$$

then

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

with respect to model-design probability  $Y\pi$ .

**Remark. A.2** is a standard  $1 + \delta$  moment assumption for the pointwise law of large number to hold. Observe  $1 + \delta$  moment bound is crucial, it is not enough to assume first moment exists. One typically needs a  $2 + \delta$  moment bound for the central limit theorem to hold (more about this in section 4.4).

**Remark.** **A.6** is a technical assumption to ensure that  $\theta_0$  is the unique zero of  $\mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla_{\theta} p \ell^s(\theta) \right]$  on  $\Theta$ . The assumptions given above are essentially the same as **Yi et al. (2016)**.

**Remark.** This result also holds when the sampling clusters are not the same as the model clusters (more about this in Chapter 4).

**Remark.** In view of Theorem 7.21, I may assume  $p = 1$ .

The key step to establish Theorem 3.8 is the uniform law of large numbers (ULLN) on a compact set which can be proved from the pointwise law of large numbers (PLLN) and equicontinuity conditions. Once this has been done, the rest is just routine. The argument is slightly modification from Lemma 5.3 from **Shao (2003)**, **Yi et al. (2016)**; **Carrillo-Garcia (2008)**. More precisely,

**Lemma 3.9.** *With the same conditions **A.1** – **A.5** as in Theorem 3.8, then one has*

$$\sup_{\theta \in \Theta} \left\| \frac{1}{N_I} \nabla p \ell^s(\theta) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla p \ell^s(\theta) \right] \right\| \xrightarrow{p} 0 \quad (3.2)$$

with respect to model-design probability  $Y\pi$ .

*Proof.* I start by introducing some notation. Let  $\theta \in \Theta$  and  $\rho > 0$ , define  $B_{\rho}(\theta) = \{\theta' \in \Theta : \|\theta - \theta'\| < \rho\}$ . Consider the upper sum and the lower sum of the weighted pairwise score function on  $B_{\rho}(\theta)$ , i.e.,

$$\begin{aligned} U_{\rho}(\theta) &= \frac{1}{N_I} \sum_{i \in U_1^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \sup_{\theta' \in B_{\rho}(\theta)} \nabla_{\theta} p \ell_{kl|i}(\theta'), \\ L_{\rho}(\theta) &= \frac{1}{N_I} \sum_{i \in U_1^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \inf_{\theta' \in B_{\rho}(\theta)} \nabla_{\theta} p \ell_{kl|i}(\theta'). \end{aligned}$$

Observe

$$\|U_{\rho}(\theta) - L_{\rho}(\theta)\| \leq \frac{2}{N_I} \sum_{i \in U_1^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} h_{kl|i}, \quad (3.3)$$

where  $h_{kl|i}(Y_{ik}, Y_{il}) = \sup_{\theta \in \Theta} \|\nabla_{\theta} p \ell_{kl|i}(\theta)\|$ .

I want to show the difference between the upper sum and lower sum of the weighted pairwise score function convergence to 0 as the radius of the ball  $\rho$  go to 0. More precisely,

**Claim:** Let  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , for all  $\epsilon > 0$ , there exists a  $\rho > 0$  and  $n_0 \in \mathbb{N}$  such that for all  $n_I \geq n_0$ , one has

$$\mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| \geq \epsilon) \leq \epsilon.$$

**Proof:** The proof is based on a truncation argument. I want to truncate the difference between the upper sum and lower sum such that  $\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\|1_{\{\|\mathbf{Y}_i\| > c\}}$  is bounded by using moment condition [A.2](#), where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iN_i})^T$ . For the remainder term  $\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\|1_{\{\|\mathbf{Y}_i\| \leq c\}}$ , I am going to show the measure of that set is small by using the equicontinuity condition [A.3](#). More precisely, let  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , for all  $\epsilon > 0$ ,  $c > 0$  and  $\rho > 0$ , one has

$$\begin{aligned} & \mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| \geq \epsilon) \\ &= \mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\|1_{\{\|\mathbf{Y}_i\| > c\}} \geq \epsilon) + \mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\|1_{\{\|\mathbf{Y}_i\| \leq c\}} \geq \epsilon). \end{aligned} \quad (3.4)$$

For the first term in [3.4](#), I can choose a sufficiently large  $c > 0$  to make it as small as I want by using the moment bound. Fixed such  $c > 0$ , I can bound the second term in [3.4](#)

by equicontinuity condition. More precisely, for the first term in 3.4, one has

$$\begin{aligned}
\mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| 1_{\{\|\mathbf{Y}_i\| > c\}} \geq \epsilon) &\leq \mathbb{P}_{Y\pi} \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} h_{kl|i} 1_{\{\|\mathbf{Y}_i\| > c\}} \geq \frac{\epsilon}{2} \right) \\
&\leq \frac{2}{\epsilon} \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} h_{kl|i} 1_{\{\|\mathbf{Y}_i\| > c\}} \right] \\
&= \frac{2}{\epsilon} \mathbb{E}_Y \left[ \frac{1}{N_I} \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} h_{kl|i} 1_{\{\|\mathbf{Y}_i\| > c\}} \right] \\
&\leq \frac{2\lambda^2}{\epsilon} \sup_{\{i,k,l\}} \mathbb{E}_Y [h_{kl|i} 1_{\{\|\mathbf{Y}_i\| > c\}}] \\
&\leq \frac{2\lambda^2}{\epsilon} \sup_{\{i,k,l\}} \left[ \mathbb{E}_Y h_{kl|i}^{1+\delta} \right]^{\frac{1}{1+\delta}} [\mathbb{P}_Y(\|\mathbf{Y}_i\| > c)]^{\frac{\delta}{1+\delta}} \\
&= \frac{2\lambda^2}{\epsilon} \sup_{\{i,k,l\}} \left[ \mathbb{E}_Y h_{kl|i}^{1+\delta} \right]^{\frac{1}{1+\delta}} \left[ \mathbb{P}_Y(\|\mathbf{Y}_i\|^\delta > c^\delta) \right]^{\frac{\delta}{1+\delta}} \\
&\leq \frac{2\lambda^2}{\epsilon} \sup_{\{i,k,l\}} \left[ \mathbb{E}_Y h_{kl|i}^{1+\delta} \right]^{\frac{1}{1+\delta}} \left[ \mathbb{E}_Y(\|\mathbf{Y}_i\|^\delta) \right]^{\frac{\delta}{1+\delta}} c^{-\frac{\delta^2}{1+\delta}}.
\end{aligned}$$

In the first inequality, I used 3.3. In the second inequality, I used Markov's Inequality Lemma 7.16. In the third inequality, I used A.5, i.e.,  $\sup_i N_i \leq \lambda$ . In the fourth, I used Hölder's Inequality Lemma 7.14. In the last inequality, I used Markov's Inequality Lemma 7.16 again.

From the moment bound A.2, I can find a  $c > 0$  such that

$$\frac{2\lambda^2}{\epsilon} \sup_{\{i,k,l\}} \left[ \mathbb{E}_Y h_{kl|i}^{1+\delta} \right]^{\frac{1}{1+\delta}} \left[ \mathbb{E}_Y(\|\mathbf{Y}_i\|^\delta) \right]^{\frac{\delta}{1+\delta}} c^{-\frac{\delta^2}{1+\delta}} \leq \frac{\epsilon}{2},$$

i.e.,

$$\mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| 1_{\{\|\mathbf{Y}_i\| > c\}} \geq \epsilon) \leq \frac{\epsilon}{2}. \tag{3.5}$$



Fix such  $c > 0$ , for the second term in 3.4, I use the equicontinuity condition A.3, i.e., for all  $\epsilon > 0$ , there exists a  $\rho > 0$  such that

$$\sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in A} \left\| \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}_1) 1_{\{\|\mathbf{Y}_i\| \leq c\}} - \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}_2) 1_{\{\|\mathbf{Y}_i\| \leq c\}} \right\| \leq \frac{\epsilon}{\lambda^2}$$

for all open sets  $A$  with  $\text{diam}(A) \leq \rho$  and for all  $k, l, i$ . In particular, for all  $\epsilon > 0$ , there exists a  $\rho > 0$  such that

$$\sup_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}') 1_{\{\|\mathbf{Y}_i\| \leq c\}} - \inf_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}') 1_{\{\|\mathbf{Y}_i\| \leq c\}} \leq \frac{\epsilon}{\lambda^2}$$

for all  $k, l, i$ . Summing over all possible pairs and normalising by  $\frac{1}{N_I}$ , one has

$$\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| 1_{\{\|\mathbf{Y}_i\| \leq c\}} \leq \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \frac{\epsilon}{\lambda^2}.$$

Hence,

$$\begin{aligned} \mathbb{P}_\pi \left( \|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| 1_{\{\|\mathbf{Y}_i\| \leq c\}} \geq \epsilon \right) &\leq \mathbb{P}_\pi \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \frac{\epsilon}{\lambda^2} \geq \epsilon \right) \\ &= \mathbb{P}_\pi \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \geq \lambda^2 \right) \\ &\leq \frac{\epsilon}{2}. \end{aligned} \tag{3.6}$$

In the last inequality, I used the following fact: by A.4 and A.5, one has

$$\begin{aligned} \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} - \frac{1}{N_I} \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in U_i^c}} 1 &= O_p(n_I^{-\frac{1}{2}}), \\ \sup_i N_i &\leq \lambda. \end{aligned}$$

Putting all those estimates 3.5 and 3.6 together, one has

$$\begin{aligned}
& \mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| \geq \epsilon) \\
&= \mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| 1_{\{\|\mathbf{Y}_i\| > c\}} \geq \epsilon) + \mathbb{P}_{Y\pi} (\|U_\rho(\boldsymbol{\theta}) - L_\rho(\boldsymbol{\theta})\| 1_{\{\|\mathbf{Y}_i\| \leq c\}} \geq \epsilon) \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&\leq \epsilon.
\end{aligned}$$

This completes the proof of the claim. ■

Now, I want to show the lower sum convergences to the expectation of the lower sum function. More precisely,

**Claim:** Let  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , for all  $\rho > 0$  and  $\epsilon > 0$ , there exists a  $n_0 \in \mathbb{N}$  such that for all  $n_I \geq n_0$ , one has

$$\mathbb{P}_{Y\pi} (\|L_\rho(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta})]\| \geq \epsilon) \leq \epsilon.$$

**Proof:** Observe

$$\begin{aligned}
& \mathbb{P}_{Y\pi} (\|L_\rho(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta})]\| \geq \epsilon) \\
&\leq \mathbb{P}_{Y\pi} \left( \|L_\rho(\boldsymbol{\theta}) - \mathbb{E}_\pi [L_\rho(\boldsymbol{\theta})]\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}_{Y\pi} \left( \|\mathbb{E}_\pi [L_\rho(\boldsymbol{\theta})] - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta})]\| \geq \frac{\epsilon}{2} \right). \quad (3.7)
\end{aligned}$$

For the first term in 3.7, using A.4, one has

$$\frac{1}{N_I} \sum_{i \in \mathcal{U}_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in \mathcal{U}_i^s}} \omega_{kl|i} \inf_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}') - \frac{1}{N_I} \sum_{i=1}^{N_I} \sum_{\substack{k < l \\ k, l \in \mathcal{U}_i^c}} \inf_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}') = O_p(n_I^{-\frac{1}{2}}),$$

i.e.,

$$\mathbb{P}_\pi \left( \|L_\rho(\boldsymbol{\theta}) - \mathbb{E}_\pi [L_\rho(\boldsymbol{\theta})]\| \geq \frac{\epsilon}{2} \right) \leq \frac{\epsilon}{2}.$$

Hence,

$$\mathbb{P}_{Y\pi} \left( \|L_\rho(\boldsymbol{\theta}) - \mathbb{E}_\pi [L_\rho(\boldsymbol{\theta})]\| \geq \frac{\epsilon}{2} \right) \leq \frac{\epsilon}{2}. \quad (3.8)$$

For the second term in 3.7, observe

$$\mathbb{E}_\pi [L_\rho(\boldsymbol{\theta})] = \frac{1}{N_I} \sum_{i=1}^{N_I} \inf_{\boldsymbol{\theta} \in B_\rho} p_{\ell_i^c}(\boldsymbol{\theta})$$

is a sum of independent random variables under model measure  $Y$ . By the pointwise law of large numbers, one has

$$\mathbb{P}_{Y\pi} \left( \|\mathbb{E}_\pi [L_\rho(\boldsymbol{\theta})] - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta})]\| \geq \frac{\epsilon}{2} \right) \leq \frac{\epsilon}{2}. \quad (3.9)$$

Putting all those estimates 3.7, 3.8 and 3.9 together, one has

$$\mathbb{P}_{Y\pi} (\|L_\rho(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta})]\| \geq \epsilon) \leq \epsilon.$$

■

Now I proceed to the proof of the uniform law of large numbers Lemma 3.9. Observe the set of open balls  $\{B_\rho(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$  forms an open cover for  $\Theta$ . Since  $\Theta$  is compact, then there exists a finite subcover, say  $\{B_\rho(\boldsymbol{\theta}_h) : \boldsymbol{\theta}_h \in \Theta, h = 1, \dots, H\}$  such that  $\Theta \subset \bigcup_{h=1}^H B_\rho(\boldsymbol{\theta}_h)$ . Observe

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in \Theta} \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}) \right] \right) \\ &= \sup_h \sup_{\boldsymbol{\theta}'_h \in B_\rho(\boldsymbol{\theta}_h)} \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}'_h) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}'_h) \right] \right) \\ &= \sup_h \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \sup_{\boldsymbol{\theta}'_h \in B_\rho(\boldsymbol{\theta}_h)} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}'_h) - \inf_{\boldsymbol{\theta}'_h \in B_\rho(\boldsymbol{\theta}_h)} \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}'_h) \right] \right) \\ &\leq \sup_h (U_\rho(\boldsymbol{\theta}_h) - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta}_h)]) \\ &= \sup_h (U_\rho(\boldsymbol{\theta}_h) - L_\rho(\boldsymbol{\theta}_h)) + \sup_h (L_\rho(\boldsymbol{\theta}_h) - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta}_h)]). \end{aligned} \quad (3.10)$$

The first equality follows from construction. The second equality follows from Lemma 7.13. The inequality follows from Fatou's Lemma, i.e.,

$$\mathbb{E}_{Y\pi} \left[ \inf_{\boldsymbol{\theta}'_h \in B_\rho(\boldsymbol{\theta}_h)} \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}'_h) \right] \leq \inf_{\boldsymbol{\theta}'_h \in B_\rho(\boldsymbol{\theta}_h)} \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}'_h) \right].$$

Hence

$$\begin{aligned}
& \mathbb{P}_{Y\pi} \left( \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}) \right] \right) \geq \epsilon \right) \\
& \leq \mathbb{P}_{Y\pi} \left( \sup_h (U_\rho(\boldsymbol{\theta}_h) - L_\rho(\boldsymbol{\theta}_h)) \geq \frac{\epsilon}{2} \right) + \mathbb{P}_{Y\pi} \left( \sup_h (L_\rho(\boldsymbol{\theta}_h) - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta}_h)]) \geq \frac{\epsilon}{2} \right) \\
& \leq \sum_{h=1}^H \mathbb{P}_{Y\pi} \left( U_\rho(\boldsymbol{\theta}_h) - L_\rho(\boldsymbol{\theta}_h) \geq \frac{\epsilon}{2} \right) + \sum_{h=1}^H \mathbb{P}_{Y\pi} \left( L_\rho(\boldsymbol{\theta}_h) - \mathbb{E}_{Y\pi} [L_\rho(\boldsymbol{\theta}_h)] \geq \frac{\epsilon}{2} \right) \\
& \leq H \frac{\epsilon}{2} + H \frac{\epsilon}{2} \\
& \leq H\epsilon.
\end{aligned}$$

The first inequality follows from 3.10. The third inequality follows from the previous two claims.

Using the same argument, one can show

$$\begin{aligned}
& \mathbb{P}_{Y\pi} \left( \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left( \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^s}} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(\boldsymbol{\theta}) \right] \right) \leq -\epsilon \right) \\
& \leq \epsilon.
\end{aligned}$$

This completes the proof of Lemma 3.9. □

I now proceed to the proof of Theorem 3.8.

*Proof.* From A.6, for all  $\epsilon > 0$ , there exists a  $\delta > 0$  such that

$$\left\{ \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \geq \epsilon \right\} \subset \left\{ \left\| \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p_{\ell^s}(\hat{\boldsymbol{\theta}}_n) \right] \right\| > \delta \right\}.$$

Fix such  $\delta > 0$ , observe

$$\begin{aligned}
\mathbb{P}_{Y\pi} \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right\| \geq \epsilon \right) &\leq \mathbb{P}_{Y\pi} \left( \left\| \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\hat{\boldsymbol{\theta}}_n) \right] \right\| > \delta \right) \\
&= \mathbb{P}_{Y\pi} \left( \left\| \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\hat{\boldsymbol{\theta}}_n) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\hat{\boldsymbol{\theta}}_n) \right] \right\| > \delta \right) \\
&\leq \mathbb{P}_{Y\pi} \left( \sup_{\boldsymbol{\theta} \in \Theta} \left\| \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\boldsymbol{\theta}) \right] \right\| > \delta \right) \xrightarrow{p} 0.
\end{aligned}$$

The last inequality follows from Lemma 3.9. This completes the proof of Theorem 3.8.  $\square$

### 3.4.3 Variance estimation

I now turn to the problem of variance estimation. Let us write this down in detail, as I will apply similar arguments without further comment. The argument given in here essentially follows the paper [Yi et al. \(2016\)](#), though with more explicit detail.

Variance is given by

$$\text{Var}_{Y\pi}(\hat{\boldsymbol{\theta}}_n) = \mathbf{J}_1(\boldsymbol{\theta}) + \mathbf{J}_2(\boldsymbol{\theta}), \quad (3.11)$$

where  $\mathbf{J}_1(\boldsymbol{\theta}) = \mathbb{E}_Y \left[ \text{Var}_{\pi}(\hat{\boldsymbol{\theta}}_n) \right]$  and  $\mathbf{J}_2(\boldsymbol{\theta}) = \text{Var}_Y \left[ \mathbb{E}_{\pi}(\hat{\boldsymbol{\theta}}_n) \right]$ . The first term  $\mathbf{J}_1(\boldsymbol{\theta})$  is the variance due to design and the second term  $\mathbf{J}_2(\boldsymbol{\theta})$  is due to model variance. If the first-stage sampling fraction  $n_I/N_I$  is small, then one can show model variance  $\mathbf{J}_2(\boldsymbol{\theta})$  can be ignored. More precisely,

**Lemma 3.10.** *Under the following regularity conditions,*

**A.1**  $\mathbf{J}_2(\boldsymbol{\theta}) = O(N_I^{-1})$ .

**A.2** *The first-stage sampling fraction  $n_I/N_I$  is small, i.e.,  $n_I/N_I = o(1)$ .*

*then one has*

$$\mathbf{J}_2(\boldsymbol{\theta}) = o(n_I^{-1}).$$

*Proof.*

$$\mathbf{J}_2(\boldsymbol{\theta}) = O(N_I^{-1}) = O(o(n_I^{-1})) = o(n_I^{-1}).$$

$\square$

**Remark.** Observe if  $\mathbb{E}_\pi [\hat{\boldsymbol{\theta}}_n] = \tilde{\boldsymbol{\theta}}_N$  and  $\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0 = O_p(N_I^{-\frac{1}{2}})$ , then  $\mathbf{J}_2(\boldsymbol{\theta}) = O(N_I^{-1})$ .

Therefore it suffices to estimate  $\mathbf{J}_1(\boldsymbol{\theta})$ . Typically  $\hat{\boldsymbol{\theta}}_n$  is a nonlinear function and  $\text{Var}_\pi$  operator does not behave nicely on the space of nonlinear functions. The standard technique in the literature is to linearise it by a Taylor series. In complex sampling setting, the argument is due to [Binder \(1983\)](#). More precisely,

**Lemma 3.11.** *With the same conditions as in Theorem 3.8 and assume  $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_p(n_I^{-\frac{1}{2}})$ , then*

$$\text{Var}_\pi(\hat{\boldsymbol{\theta}}_n) = \left( -\frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0) \right)^{-\text{T}} \text{Var}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) \right) \left( -\frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0) \right)^{-\text{T}} + o_p(n_I^{-1}).$$

*Proof.* Observe

$$\begin{aligned} \mathbf{0} &= \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\hat{\boldsymbol{\theta}}_n) \\ &= \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) + \frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^s}(\boldsymbol{\theta}_0)^{\text{T}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(n_I^{-\frac{1}{2}}) \\ &= \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) + \frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0)^{\text{T}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \frac{1}{N_I} \left( \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^s}(\boldsymbol{\theta}_0) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0) \right)^{\text{T}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ &\quad + o_p(n_I^{-\frac{1}{2}}) \\ &= \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) + \frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0)^{\text{T}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(n_I^{-\frac{1}{2}}). \end{aligned}$$

In the second equality, I used Taylor Theorem 7.18 to function  $\frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta})$  at  $\boldsymbol{\theta}_0$ . In the fourth equality, I used the fact that  $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = o_p(1)$  and [A.4](#) in Theorem 3.8. The point in here is that I replace  $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^s}(\boldsymbol{\theta}_0)$  by  $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0)$ , which is a constant with respect to design measure  $\pi$ . Therefore

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = \left( -\frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0) \right)^{-\text{T}} \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) + o_p(n_I^{-\frac{1}{2}}).$$

Since  $\boldsymbol{\theta}_0$  and  $\nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}_0)$  are constants with respect to design  $\pi$ , one has

$$\text{Var}_\pi(\hat{\boldsymbol{\theta}}_n) = \left( -\frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0) \right)^{-\text{T}} \text{Var}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) \right) \left( -\frac{1}{N_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p^{\ell^c}(\boldsymbol{\theta}_0) \right)^{-\text{T}} + o_p(n_I^{-1}).$$

□

The natural estimator for  $\frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta_0)$  is  $\frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\hat{\theta}_n)$ . More precisely,

**Proposition 3.12.** *With the same conditions as in Theorem 3.8. In addition, assume*

**B.1** *Let  $g_{kl|i}(Y_{ik}, Y_{il}) = \sup_{\{\theta \in \Theta\}} \|\nabla_{\theta\theta}^2 p_{kl|i}(\theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y g_{kl|i}^{1+\delta} < \infty$  and  $\sup_i \mathbb{E}_Y \|\mathbf{Y}_i\|^\delta < \infty$ , where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iN_i})^T$ .*

**B.2** *For any given  $c > 0$  and a given sequence  $\{\mathbf{y}_i\}$  satisfying  $\|\mathbf{y}_i\| \leq c$ , the sequence of function  $\{\nabla_{\theta\theta}^2 p_{kl|i}(\theta)\}$  is equicontinuous on any open subset  $A$  of  $\Theta$ .*

then

$$\frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\hat{\theta}_n) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta_0) = o_p(1)$$

under design measure  $\pi$ .

To prove this, one needs a lemma to start with.

**Lemma 3.13.** *With the same conditions A.1 – A.5 as in Theorem 3.8. In addition, assume*

**B.1** *Let  $g_{kl|i}(Y_{ik}, Y_{il}) = \sup_{\{\theta \in \Theta\}} \|\nabla_{\theta\theta}^2 p_{kl|i}(\theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y g_{kl|i}^{1+\delta} < \infty$  and  $\sup_i \mathbb{E}_Y \|\mathbf{Y}_i\|^\delta < \infty$ , where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iN_i})^T$ .*

**B.2** *For any given  $c > 0$  and a given sequence  $\{\mathbf{y}_i\}$  satisfying  $\|\mathbf{y}_i\| \leq c$ , the sequence of function  $\{\nabla_{\theta\theta}^2 p_{kl|i}(\theta)\}$  is equicontinuous on any open subset  $A$  of  $\Theta$ .*

then

$$\sup_{\theta \in \Theta} \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\theta) - \mathbb{E}_{Y\pi} \left[ \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\theta) \right] \right\| \xrightarrow{p} 0$$

with respect to model-design probability  $Y\pi$ .

*Proof.* This can be done using exactly the same argument as in Lemma 3.9. I omit the detail.  $\square$

Now I proceed to the proof of Proposition 3.12.

*Proof.* Observe

$$\begin{aligned}
& \mathbb{P}_\pi \left( \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\hat{\theta}_n) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta_0) \right\| \geq \epsilon \right) \\
& \leq \mathbb{P}_\pi \left( \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\hat{\theta}_n) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\hat{\theta}_n) \right\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}_\pi \left( \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\hat{\theta}_n) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta_0) \right\| \geq \frac{\epsilon}{2} \right) \\
& \leq \mathbb{P}_\pi \left( \sup_{\theta \in \Theta} \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\theta) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta) \right\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}_\pi \left( \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\hat{\theta}_n) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta_0) \right\| \geq \frac{\epsilon}{2} \right). \tag{3.12}
\end{aligned}$$

For the first term in 3.12, one has

$$\mathbb{P}_\pi \left( \sup_{\theta \in \Theta} \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\theta) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta) \right\| \geq \frac{\epsilon}{2} \right) \leq \frac{\epsilon}{2} \tag{3.13}$$

by Lemma 3.13.

For the second term in 3.12, note  $\hat{\theta}_n \xrightarrow{p} \theta_0$  by Theorem 3.8. Hence, by the Continuous Mapping Theorem 7.20, one has

$$\mathbb{P}_\pi \left( \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\hat{\theta}_n) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta_0) \right\| \geq \frac{\epsilon}{2} \right) \leq \frac{\epsilon}{2}. \tag{3.14}$$

Putting all those estimates 3.12, 3.13 and 3.14 together, one has

$$\mathbb{P}_\pi \left( \left\| \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^s}(\hat{\theta}_n) - \frac{1}{N_I} \nabla_{\theta\theta}^2 p^{\ell^c}(\theta_0) \right\| \geq \epsilon \right) \leq \epsilon.$$

□

It remains to estimate  $\text{Var}_\pi(\nabla_{\theta} p^{\ell^s}(\theta_0))$ . I will explicitly calculate  $\text{Var}_\pi(\nabla_{\theta} p^{\ell^s}(\theta_0))$ . Let  $\pi_1$  be the first-stage design measure and let  $\pi_2$  be the second-stage design measure conditional on the first-stage sampling cluster  $U_1^s$ . Recall  $\Delta_{klk'l'|i} = \pi_{klk'l'|i} - \pi_{kl|i}\pi_{k'l'|i}$ .

**Lemma 3.14.** *Observe*

$$\text{Var}_\pi \left( \frac{1}{N_I} \nabla_{\theta} p^{\ell^s}(\theta_0) \right) = V_1(\theta_0) + V_2(\theta_0),$$



where

$$V_1(\boldsymbol{\theta}_0) = \frac{1}{N_I^2} \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} (\omega_i \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \omega_{i'} \nabla_{\boldsymbol{\theta}} p \ell_{i'}^c(\boldsymbol{\theta}_0)),$$

$$V_2(\boldsymbol{\theta}_0) = \frac{1}{N_I^2} \sum_{i=1}^{N_I} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^c}} \sum_{\substack{k' < l' \\ k', l' \in U_i^c}} \Delta_{klk'l'|i} (\omega_{kl|i} \nabla_{\boldsymbol{\theta}} p \ell_{kl|i}(\boldsymbol{\theta}_0) \omega_{k'l'|i} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'|i}(\boldsymbol{\theta}_0)).$$

*Proof.* Observe

$$\begin{aligned} & \text{Var}_{\pi} \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\boldsymbol{\theta}_0) \right) \\ &= \frac{1}{N_I^2} \text{Var}_{\pi_1} [\mathbb{E}_{\pi_2} (\nabla_{\boldsymbol{\theta}} p \ell^s(\boldsymbol{\theta}_0))] + \frac{1}{N_I^2} \mathbb{E}_{\pi_1} \left[ \text{Var}_{\pi_2} (\nabla_{\boldsymbol{\theta}} p \ell^s(\boldsymbol{\theta}_0)) \right] \\ &= \frac{1}{N_I^2} \text{Var}_{\pi_1} \left[ \sum_{i=1}^{N_I} 1_i \omega_i \mathbb{E}_{\pi_2} (\nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0)) \right] + \frac{1}{N_I^2} \mathbb{E}_{\pi_1} \left[ \sum_{i=1}^{N_I} 1_i \omega_i^2 \text{Var}_{\pi_2} (\nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0)) \right] \\ &= \frac{1}{N_I^2} \text{Var}_{\pi_1} \left[ \sum_{i=1}^{N_I} 1_i \omega_i \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \right] + \frac{1}{N_I^2} \sum_{i=1}^{N_I} \omega_i \text{Var}_{\pi_2} \left( \sum_{\substack{k < l \\ k, l \in U_i^c}} 1_{kl|i} \omega_{kl|i} \nabla_{\boldsymbol{\theta}} p \ell_{kl|i}(\boldsymbol{\theta}_0) \right) \\ &= \frac{1}{N_I^2} \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \text{Cov}_{\pi_1} (1_i, 1_{i'}) (\omega_i \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \omega_{i'} \nabla_{\boldsymbol{\theta}} p \ell_{i'}^c(\boldsymbol{\theta}_0)) + \frac{1}{N_I^2} \sum_{i=1}^{N_I} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^c}} \sum_{\substack{k' < l' \\ k', l' \in U_i^c}} \text{Cov}_{\pi_2} (1_{kl|i}, 1_{k'l'|i}) \\ & \quad (\omega_{kl|i} \nabla_{\boldsymbol{\theta}} p \ell_{kl|i}(\boldsymbol{\theta}_0) \omega_{k'l'|i} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'|i}(\boldsymbol{\theta}_0)) \\ &= \frac{1}{N_I^2} \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} \Delta_{ii'} (\omega_i \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \omega_{i'} \nabla_{\boldsymbol{\theta}} p \ell_{i'}^c(\boldsymbol{\theta}_0)) + \frac{1}{N_I^2} \sum_{i=1}^{N_I} \omega_i \sum_{\substack{k < l \\ k, l \in U_i^c}} \sum_{\substack{k' < l' \\ k', l' \in U_i^c}} \Delta_{klk'l'|i} \\ & \quad (\omega_{kl|i} \nabla_{\boldsymbol{\theta}} p \ell_{kl|i}(\boldsymbol{\theta}_0) \omega_{k'l'|i} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'|i}(\boldsymbol{\theta}_0)). \end{aligned}$$

In the second equality, I used the fact  $1_i$  is independent of  $1_{i'}$  for all  $i \neq i'$  with respect to design measure  $\pi_2$ .  $\square$

**Remark.** One can estimate  $\text{Var}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) \right)$  by the empirical variance estimator  $\widehat{\text{Var}}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\hat{\boldsymbol{\theta}}_n) \right)$ , i.e.,

$$\widehat{\text{Var}}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\hat{\boldsymbol{\theta}}_n) \right) = \hat{V}_1(\hat{\boldsymbol{\theta}}_n) + \hat{V}_2(\hat{\boldsymbol{\theta}}_n),$$

where

$$\hat{V}_1(\hat{\boldsymbol{\theta}}_n) = \frac{1}{N_I^2} \sum_{i=1}^{N_I} \sum_{i'=1}^{N_I} 1_{ii'} \omega_{ii'} \Delta_{ii'} \left( \omega_i \nabla_{\boldsymbol{\theta}} p^{\ell_i^s}(\hat{\boldsymbol{\theta}}_n) \omega_{i'} \nabla_{\boldsymbol{\theta}} p^{\ell_{i'}^s}(\hat{\boldsymbol{\theta}}_n) \right),$$

$$\hat{V}_2(\hat{\boldsymbol{\theta}}_n) = \frac{1}{N_I^2} \sum_{i=1}^{N_I} 1_i \omega_i \sum_{\substack{k < l \\ k, l \in U_i^c}} \sum_{\substack{k' < l' \\ k', l' \in U_i^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} \left( \omega_{kl} \nabla_{\boldsymbol{\theta}} p^{\ell_{kl}}(\hat{\boldsymbol{\theta}}_n) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p^{\ell_{k'l'}}(\hat{\boldsymbol{\theta}}_n) \right).$$

One can show this is an unbiased estimator, i.e.,  $\mathbb{E}_\pi \left[ \widehat{\text{Var}}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}) \right) \right] = \text{Var}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}) \right)$ . The argument is similar to Lemma 2.31, so I omit the detail.

It can be shown that the empirical variance estimator  $\widehat{\text{Var}}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\hat{\boldsymbol{\theta}}_n) \right)$  is a consistent estimator for some reasonable sampling designs. For more detail, see Chapter 4.

### 3.4.4 Yi's approach

Recall exact computation for  $\text{Var}_\pi (\nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0))$  needs fourth-order sampling inclusion probability as shown in Lemma 3.14. Yi et al. (2016) approximated the design by one where the PSU  $i$  is selected with replacement with probability  $p_i$  for all  $i = 1, \dots, N_I$ . Note  $\pi_i = n_I p_i$ . Hence

$$\begin{aligned} \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}_0) &= \frac{1}{N_I} \sum_{i \in U_I^s} \omega_i \nabla_{\boldsymbol{\theta}} p^{\ell_i^s}(\boldsymbol{\theta}_0) \\ &= \frac{1}{N_I} \frac{1}{n_I} \sum_{i \in U_I^s} \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p^{\ell_i^s}(\boldsymbol{\theta}_0). \end{aligned}$$

Observe  $\frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p^{\ell_i^s}(\boldsymbol{\theta}_0)$  are independent identically distributed random variables from the design perspective by construction. In particular, the mean and variance of  $\frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p^{\ell_i^s}(\boldsymbol{\theta}_0)$

are given by

$$\begin{aligned}
\mathbb{E}_\pi \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) \right] &= \mathbb{E}_{\pi_1} \left( \mathbb{E}_{\pi_2} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) \right] \right) \\
&= \mathbb{E}_{\pi_1} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \right] \\
&= \sum_{i=1}^{N_I} p_i \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \\
&= \sum_{i=1}^{N_I} \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0), \\
\text{Var}_\pi \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) \right] &= \text{Var}_{\pi_1} \left( \mathbb{E}_{\pi_2} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) \right] \right) + \mathbb{E}_{\pi_1} \left( \text{Var}_{\pi_2} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) \right] \right) \\
&= \text{Var}_{\pi_1} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \right] + \sum_{i=1}^{N_I} p_i \text{Var}_{\pi_2} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) \right] \\
&= \sum_{i=1}^{N_I} p_i \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) - \sum_{i=1}^{N_I} \nabla_{\boldsymbol{\theta}} p \ell_i^c(\boldsymbol{\theta}_0) \right]^2 + \sum_{i=1}^{N_I} p_i \text{Var}_{\pi_2} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}_0) \right].
\end{aligned}$$

Hence an unbiased variance estimator for  $\text{Var}_\pi \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}) \right]$  is given by

$$\begin{aligned}
\widehat{\text{Var}}_\pi \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}) \right] &= \frac{1}{n_I - 1} \sum_{i \in \mathcal{U}_I^s} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}) - \frac{1}{n_I} \sum_{i \in \mathcal{U}_I^s} \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}) \right]^2 \\
&= \frac{1}{n_I - 1} \sum_{i \in \mathcal{U}_I^s} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} p \ell^s(\boldsymbol{\theta}) \right]^2.
\end{aligned} \tag{3.15}$$

In the second equality, I use the fact  $\pi_i = n_I p_i$ .

Therefore, the empirical variance estimator for  $\text{Var}_\pi \left[ \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\boldsymbol{\theta}_0) \right]$  is given by

$$\begin{aligned} \widehat{\text{Var}}_\pi \left( \frac{1}{N_I} \nabla_{\boldsymbol{\theta}} p \ell^s(\hat{\boldsymbol{\theta}}_n) \right) &= \frac{1}{N_I^2} \frac{1}{n_I} \widehat{\text{Var}}_\pi \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\hat{\boldsymbol{\theta}}) \right] \\ &= \frac{1}{N_I^2} \frac{1}{n_I(n_I - 1)} \sum_{i \in U_I^s} \left[ \frac{1}{p_i} \nabla_{\boldsymbol{\theta}} p \ell_i^s(\hat{\boldsymbol{\theta}}_n) \right]^2 \\ &= \frac{1}{N_I^2} \frac{n_I}{(n_I - 1)} \sum_{i \in U_I^s} \omega_i^2 \left[ \nabla_{\boldsymbol{\theta}} p \ell_i^s(\hat{\boldsymbol{\theta}}_n) \right]^2. \end{aligned}$$

In the second equality, I used 3.15 and the fact  $\nabla_{\boldsymbol{\theta}} p \ell^s(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$ . In the last equality, I used  $\omega_i = \frac{1}{n_I p_i}$ .

### 3.4.5 Jackknife variance estimation

I want to discuss Jackknife variance estimation for  $\text{Var}_{Y\pi}(\hat{\boldsymbol{\theta}}_n)$ . Recall

$$\text{Var}_{Y\pi}(\hat{\boldsymbol{\theta}}_n) = \mathbf{J}_1(\boldsymbol{\theta}) + \mathbf{J}_2(\boldsymbol{\theta}),$$

where  $\mathbf{J}_1(\boldsymbol{\theta}) = \mathbb{E}_Y \left[ \text{Var}_\pi(\hat{\boldsymbol{\theta}}_n) \right]$  and  $\mathbf{J}_2(\boldsymbol{\theta}) = \text{Var}_Y \left[ \mathbb{E}_\pi(\hat{\boldsymbol{\theta}}_n) \right]$ . By Lemma 3.10, if the first-stage sampling fraction is small, then the model variance  $\mathbf{J}_2(\boldsymbol{\theta})$  can be ignored.

It reduces to estimate  $\mathbf{J}_1(\boldsymbol{\theta})$ . One can estimate  $\text{Var}_\pi \left[ \hat{\boldsymbol{\theta}}_n \right]$  by resampling the sampling clusters using a jackknife. Consider a partition of the sampling clusters  $U_I^s$  into  $m$  subgroup. Let  $\hat{\boldsymbol{\theta}}_{(j)}$  be the  $j$ 's deleted Jackknife estimator, i.e., using the same method as  $\hat{\boldsymbol{\theta}}$  to estimate  $\boldsymbol{\theta}$  after deleting  $j$ 's group data. Then the Jackknife estimator for  $\text{Var}_\pi(\hat{\boldsymbol{\theta}})$  is given by

$$\widehat{\text{Var}}_\pi(\hat{\boldsymbol{\theta}}) = \frac{1}{m(m-1)} \sum_{j=1}^m (\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}})^2,$$

where

$$\begin{aligned} \hat{\boldsymbol{\theta}}_j &= m\hat{\boldsymbol{\theta}} - (m-1)\hat{\boldsymbol{\theta}}_{(j)}, \\ \hat{\boldsymbol{\theta}} &= \frac{1}{m} \sum_{j=1}^m \hat{\boldsymbol{\theta}}_j. \end{aligned}$$

## 4 When sampling and model clusters are not the same

In this chapter, I want to extend the asymptotic properties of the sample weighted pairwise likelihood estimator to the case when the sampling clusters are not the same as the model clusters. The main goal of this chapter is to establish consistency and asymptotic normality of the sample weighted pairwise likelihood estimator. In addition, I construct an empirical variance estimator for the sample weighted pairwise likelihood estimator and show it is consistent. Essentially, the same proof with minor modification will yield the weighted pairwise likelihood estimator is a consistent estimator. Once this has been done, it reduces to a standard argument to find the asymptotic distribution. But proving the consistency of the empirical variance estimator is surprisingly more difficult than it first seems and requires a new approach. I start by introducing the setting.

### 4.1 Setting: design

Let  $U = \{1, \dots, N\}$  be the population and consider two partitions of  $U$ . One partition is by the sampling design and the other is by the structure of the model. More specifically, let  $U_I^c = \{U_1^c, \dots, U_{N_I}^c\}$  be the sampling clusters and  $M_I^c = \{M_1^c, \dots, M_{T_c}^c\}$  be the model clusters. In practice, the sampling clusters are the objects corresponding to the first-stage sampling units (PSU) and the model clusters are the objects corresponding to the fibre (inverse image) of random effects. More precisely, the random effect is defined on the set of model clusters and two distinct model clusters are assigned different value of the random effects.

**Definition 4.1.** The sampling clusters  $U_I^c$  are finer than the model clusters  $M_I^c$  if  $U_I^c \subset M_I^c$ .

**Definition 4.2.** The sampling clusters  $U_I^c$  is coarser than the model clusters  $M_I^c$  if  $M_I^c \subset U_I^c$ .

**Remark.** It happens quite often that the sampling clusters and model clusters cannot be directly compared, i.e.,  $U_I^c \not\subset M_I^c$  and  $M_I^c \not\subset U_I^c$ . For example, Hispanic Community Health Study/Study of Latinos in section 1.1. One must search for a solution in this more general setting.

Define  $\mathcal{F}$  to be the set of all subsets of  $U_1^c$ , i.e.,  $\mathcal{F} = \{\bigcup_{i \in I} U_i^c : I \subset \{1, 2, \dots, N_1\}\}$ . Let  $\mathbb{P}_1$  be the first-stage design probability on measure space  $(U_1^c, \mathcal{F})$

$$\begin{aligned}\mathbb{P}_1 : \mathcal{F} &\longrightarrow [0, 1] \\ A &\longmapsto \mathbb{P}_1(A).\end{aligned}$$

Define  $\mathcal{F}_i$  to be the set of all subsets of  $U_i^c$  for  $i = 1, \dots, N_1$ . Let  $\mathbb{P}_i$  be the second-stage design probability on measure space  $(U_i^c, \mathcal{F}_i)$

$$\begin{aligned}\mathbb{P}_i : \mathcal{F}_i &\longrightarrow [0, 1] \\ A &\longmapsto \mathbb{P}_i(A).\end{aligned}$$

Let  $U_1^s \in \mathcal{F}$  be the first-stage sample. If  $U_i^c \in U_1^s$ , a second-stage sample  $U_i^s \in \mathcal{F}_i$  is selected by sampling design  $\mathbb{P}_i$ . Let  $S = \bigcup_i U_i^s$ .

**Remark.** In all of the following, I will denote  $i, i', i'', i'''$  for the clusters and  $k, l, k', l', k'', l'', k''', l'''$  for elements in the cluster.

In my construction below, I will need some notation for identifying whether observational units are in the same sampling cluster.

**Definition 4.3.** Let  $s$  be a set-valued function from the population to the set of sampling clusters

$$\begin{aligned}s : U &\longrightarrow U_1^c \\ k &\longmapsto s(k)\end{aligned}$$

such that for every observational unit  $k \in U$ ,  $s(k)$  is the unique sampling cluster contains  $k$ .

**Definition 4.4.** Let  $r \in \mathbb{N}$  and  $k_1, \dots, k_r \in U$ . Define the bracket

$$\langle k_1, \dots, k_r \rangle$$

to be a subset of  $U$  such that elements within the bracket have the same sampling clusters, i.e.,  $s(k_1) = s(k_2) = \dots = s(k_r)$ .

**Remark.** Observe  $\langle k_1, \dots, k_r \rangle \subset s(k_1)$ .

**Remark.** This notation will be needed in proving consistency of the empirical variance estimator (more about this in section 4.6 and 4.7).

**Definition 4.5.** Two brackets  $\langle k_1, \dots, k_r \rangle$  and  $\langle l_1, \dots, l_s \rangle$  are said to be disjoint if they have different sampling clusters, i.e.,  $s(k_1) \neq s(l_1)$ . I will denote by  $\langle k_1, \dots, k_r \rangle \times \langle l_1, \dots, l_s \rangle$ .

**Definition 4.6.** Let  $\{k_1, \dots, k_r\}$  and  $\{l_1, \dots, l_s\}$  be two subsets of  $U$ .  $\{k_1, \dots, k_r\}$  is said to have a common root with  $\{l_1, \dots, l_s\}$  if there exists  $1 \leq p \leq r$  and  $1 \leq q \leq s$  such that  $s(k_p) = s(l_q)$ .

**Definition 4.7.** Define the first-order, second-order and fourth-order sampling inclusion probability for clusters to be

$$\begin{aligned}\pi_i &= \sum_{A \in \mathcal{F}: U_i^c \in A} \mathbb{P}_I(A), \\ \pi_{ii'} &= \sum_{A \in \mathcal{F}: U_i^c, U_{i'}^c \in A} \mathbb{P}_I(A), \\ \pi_{ii' i'' i'''} &= \sum_{A \in \mathcal{F}: U_i^c, U_{i'}^c, U_{i''}^c, U_{i'''}^c \in A} \mathbb{P}_I(A).\end{aligned}$$

**Definition 4.8.** Define the first-order, second-order, third-order and fourth-order conditional inclusion probability to be

$$\begin{aligned}\pi_{k|s(k)} &= \sum_{A \in \mathcal{F}_{s(k)}: k \in A} \mathbb{P}_{s(k)}(A), \\ \pi_{kl|s(k)} &= \sum_{A \in \mathcal{F}_{s(k)}: k, l \in A} \mathbb{P}_{s(k)}(A), \quad \text{if } \langle kl \rangle, \\ \pi_{klk'|s(k)} &= \sum_{A \in \mathcal{F}_{s(k)}: k, l, k' \in A} \mathbb{P}_{s(k)}(A), \quad \text{if } \langle klk' \rangle, \\ \pi_{klk'l'|s(k)} &= \sum_{A \in \mathcal{F}_{s(k)}: k, l, k', l' \in A} \mathbb{P}_{s(k)}(A), \quad \text{if } \langle klk'l' \rangle.\end{aligned}$$

One can easily deduce the final first-order  $\pi_k$ , second-order  $\pi_{kl}$  and fourth-order inclusion probability  $\pi_{klk'l'}$  using the independent and invariant property as before, i.e,

$$\begin{aligned}
\pi_k &= \pi_{k|s(k)} \pi_{s(k)}, \\
\pi_{kl} &= \begin{cases} \pi_{kl|s(k)} \pi_{s(k)}, & \text{if } \langle kl \rangle, \\ \pi_{k|s(k)} \pi_{l|s(l)} \pi_{s(k)s(l)}, & \text{otherwise.} \end{cases} \\
\pi_{klk'l'} &= \begin{cases} \pi_{klk'l'|s(k)} \pi_{s(k)}, & \text{if } \langle klk'l' \rangle, \\ \pi_{kl|s(k)} \pi_{s(k)} \pi_{k'l'|s(k')} \pi_{s(k')}, & \text{if } \langle kl \rangle \langle k'l' \rangle, \\ \pi_{kk'|s(k)} \pi_{s(k)} \pi_{ll'|s(l)} \pi_{s(l)}, & \text{if } \langle kk' \rangle \langle ll' \rangle, \\ \pi_{kl'|s(k)} \pi_{s(k)} \pi_{lk'|s(l)} \pi_{s(l)}, & \text{if } \langle kl' \rangle \langle lk' \rangle, \\ \pi_{klk'l'|s(k)} \pi_{s(k)} \pi_{l'|s(l')} \pi_{s(l')}, & \text{if } \langle klk' \rangle \langle l' \rangle, \\ \pi_{kll'|s(k)} \pi_{s(k)} \pi_{k'|s(k')} \pi_{s(k')}, & \text{if } \langle kll' \rangle \langle k' \rangle, \\ \pi_{kk'l'|s(k)} \pi_{s(k)} \pi_{l|s(l)} \pi_{s(l)}, & \text{if } \langle kk'l' \rangle \langle l \rangle, \\ \pi_{lk'l'|s(k)} \pi_{s(l)} \pi_{k|s(k)} \pi_{s(k)}, & \text{if } \langle lk'l' \rangle \langle k \rangle, \\ \pi_{kl|s(k)} \pi_{s(k)} \pi_{k'|s(k')} \pi_{s(k')} \pi_{l'|s(l')} \pi_{s(l')}, & \text{if } \langle kl \rangle \langle k' \rangle \langle l' \rangle, \\ \pi_{kk'|s(k)} \pi_{s(k)} \pi_{l|s(l)} \pi_{s(l)} \pi_{l'|s(l')} \pi_{s(l')}, & \text{if } \langle kk' \rangle \langle l \rangle \langle l' \rangle, \\ \pi_{kl'|s(k)} \pi_{s(k)} \pi_{l|s(l)} \pi_{s(l)} \pi_{k'|s(k')} \pi_{s(k')}, & \text{if } \langle kl' \rangle \langle l \rangle \langle k' \rangle, \\ \pi_{lk'|s(l)} \pi_{s(l)} \pi_{k|s(k)} \pi_{s(k)} \pi_{l'|s(l')} \pi_{s(l')}, & \text{if } \langle lk' \rangle \langle k \rangle \langle l' \rangle, \\ \pi_{ll'|s(l)} \pi_{s(l)} \pi_{k|s(k)} \pi_{s(k)} \pi_{k'|s(k')} \pi_{s(k')}, & \text{if } \langle ll' \rangle \langle k \rangle \langle k' \rangle, \\ \pi_{k'l'|s(k')} \pi_{s(k')} \pi_{k|s(k)} \pi_{s(k)} \pi_{l|s(l)} \pi_{s(l)}, & \text{if } \langle k'l' \rangle \langle k \rangle \langle l \rangle, \\ \pi_{k|s(k)} \pi_{l|s(l)} \pi_{k'|s(k')} \pi_{l'|s(l')} \pi_{s(k)s(l)s(k')s(l')}, & \text{if } \langle k \rangle \langle l \rangle \langle k' \rangle \langle l' \rangle. \end{cases}
\end{aligned}$$

**Remark.** In all of the following, I assume there exists a  $\epsilon > 0$  such that  $\sup_{kl} \pi_{kl} \geq \epsilon$ ,  $\sup_{klk'l'} \pi_{klk'l'} \geq \epsilon$  and  $\sup_{klk'l'k''l''k'''l'''} \pi_{klk'l'k''l''k'''l'''} \geq \epsilon$ .



**Definition 4.9.** Define the first-order, second-order and fourth-order weights to be

$$\begin{aligned}\omega_k &= \frac{1}{\pi_k}, \\ \omega_{kl} &= \frac{1}{\pi_{kl}}, \\ \omega_{klk'l'} &= \frac{1}{\pi_{klk'l'}}.\end{aligned}$$

It is more convenient to work with the sample indicator function  $1_i, 1_k$  instead of design measure.

**Definition 4.10.** Define the sample indicator function to be

$$\begin{aligned}1_i &= 1_{U_i^c \in U_1^s}, \\ 1_k &= 1_{k \in S}.\end{aligned}$$

**Remark.** I will use the following notation:

$$\begin{aligned}1_{ii'} &= 1_i 1_{i'}, \\ 1_{ii'i''i'''} &= 1_i 1_{i'} 1_{i''} 1_{i'''}, \\ 1_{kl} &= 1_k 1_l, \\ 1_{klk'l'} &= 1_k 1_l 1_{k'} 1_{l'}, \\ 1_{klk'l'k''l''k'''l'''} &= 1_k 1_l 1_{k'} 1_{l'} 1_{k''} 1_{l''} 1_{k'''} 1_{l'''}. \end{aligned}$$

**Remark.** Observe

$$\begin{aligned}\mathbb{E}_\pi [1_{kl}] &= \pi_{kl}, \\ \mathbb{E}_\pi [1_{klk'l'}] &= \pi_{klk'l'}, \\ \mathbb{E}_\pi [1_{klk'l'k''l''k'''l'''}] &= \pi_{klk'l'k''l''k'''l'''}. \end{aligned}$$

**Definition 4.11.** Define

$$\begin{aligned}\Delta_{klk'l'} &= \text{Cov}_\pi(1_{kl}, 1_{k'l'}), \\ \Delta_{klk'l'k''l''k'''l'''} &= \text{Cov}_\pi(1_{klk'l'}, 1_{k''l''k'''l'''}).\end{aligned}$$

**Remark.** Observe

$$\Delta_{klk'l'} = \pi_{klk'l'} - \pi_{kl}\pi_{k'l'},$$

$$\Delta_{klk'l'k''l''k'''l'''} = \pi_{klk'l'k''l''k'''l'''} - \pi_{klk'l'}\pi_{k''l''k'''l'''}$$

**Definition 4.12.** Let  $N_I$  be the number of sampling clusters in the population  $U$  and let  $n_I$  be the number of sampling clusters in the sample  $S$ , i.e.,  $n_I = \sum_{i=1}^{N_I} 1_i$ .

**Definition 4.13.** Let  $N_i$  be the number of elements in the sampling cluster  $U_i^c$  and let  $n_i$  be the number of sampled elements in the sampling cluster  $U_i^s$ , i.e.,  $n_i = \sum_{k \in U_i^c} 1_{k|i}$ .

**Remark.** Observe

$$N = \sum_{i=1}^{N_I} N_i,$$

$$n = \sum_{i=1}^{N_I} 1_i n_i.$$

## 4.2 Setting: model

In contrast to previous literature, I do not assume the sampling clusters are the same as the model clusters, i.e.,  $U_I^c \neq M_I^c$ .

**Definition 4.14.** Define

$$M_i^s = M_i^c \cap S,$$

$$M_I^s = \{M_1^s, \dots, M_{T_c}^s\} \setminus \emptyset.$$

**Definition 4.15.** Let  $m$  be the set-value function from the population to the set of model clusters

$$m : U \longrightarrow M_I^c$$

$$k \longmapsto m(k)$$

such that for every observational unit  $k \in U$ ,  $m(k)$  is the unique model cluster contains  $k$ .

**Definition 4.16.** Let  $T_c$  be the number of model clusters in the population  $U$  and let  $t_s$  be the number of model clusters in the sample  $S$ , i.e.,  $t_s = |M_1^s|$ .

In all of the following, I assume if the sampling fraction for the PSU  $n_1/N_1$  is small, then the sampling fraction for the model cluster  $t_s/T_c$  is also small.

**Definition 4.17.** Let  $T_i$  be the number of elements in the model clusters  $M_i^c$  and let  $t_i$  be the number of elements in the sample  $M_i^s$ , i.e.,  $t_i = |M_i^s|$ .

Observe

$$N = \sum_{i=1}^{T_c} T_i,$$

$$n = \sum_{i=1}^{t_s} t_i.$$

Consider a two-level model,

$$Y_{ik}|X_{ik}, \mathbf{b}_i \sim f(y_{ik}|x_{ik}, \mathbf{b}_i, \boldsymbol{\theta}_1),$$

$$\mathbf{b}_i \sim g(\mathbf{b}_i|\boldsymbol{\theta}_2),$$

for observational unit  $k$  in the model clusters  $i$ , where  $i = 1, \dots, T_c$ . Observe  $\mathbf{b}_i$  is the random effect for the model clusters  $M_i^c$  and the random effects for different model clusters are independent under the model measure  $\Upsilon$ , i.e.,  $\mathbf{b}_i \perp \mathbf{b}_{i'}$  if  $i \neq i'$ .

**Definition 4.18.** Define the census pairwise log-likelihood for the model clusters  $i$  to be

$$p\ell_{M_i^c}(\boldsymbol{\theta}) = \sum_{\substack{k < l \\ k, l \in M_i^c}} p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta}), \quad i = 1, \dots, T_c,$$

where

$$p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta}) = \log \left[ \int f(y_k|x_k, \mathbf{b}_i, \boldsymbol{\theta}_1) f(y_l|x_l, \mathbf{b}_i, \boldsymbol{\theta}_1) g(\mathbf{b}_i|\boldsymbol{\theta}_2) d\mathbf{b}_i \right], \quad k, l \in M_i^c.$$

Then the census pairwise log-likelihood is given by

$$p\ell_{M^c}(\boldsymbol{\theta}) = \sum_{i=1}^{T_c} p\ell_{M_i^c}(\boldsymbol{\theta}).$$

**Remark.** The subindex  $M$  denotes that the pairwise likelihood is calculated for the model clusters not sampling clusters. The idea of writing in term of model clusters is I get a sum of independent random variable under model measure  $Y$ , i.e.,  $p\ell_{M_1^c}, \dots, p\ell_{M_{T_1}^c}$  are independent under model measure, so that I can use the pointwise law of large number and the central limit theorem (more about this in Theorem 4.23 and Theorem 4.25).

**Remark.** Under the correlated random effect assumption, one can forget about this definition, as  $p\ell_{M_i^c}(\theta)$  are correlated under model measure  $Y$ . One needs mixing conditions to establish the pointwise law of large number and the central limit theorem (more about this in Chapter 5).

**Definition 4.19.** Define the sample weighted pairwise log-likelihood for the model clusters  $i$  to be

$$p\ell_{M_i^s}(\theta) = \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} p\ell_{kl|i}(y_k, y_l, \theta), \quad i = 1, \dots, T_c.$$

Then the sample weighted pairwise log-likelihood is given by

$$p\ell_{M^s}(\theta) = \sum_{i=1}^{T_c} p\ell_{M_i^s}(\theta).$$

One can show the sample weighted pairwise log-likelihood recovers the census pairwise log-likelihood in this new context after taking expectation. More precisely,

**Lemma 4.20.** Let  $h_{kl|i}(y_k, y_l) = \sup_{\{\theta \in \Theta\}} \|\nabla_{\theta} p\ell_{kl|i}(y_k, y_l, \theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i, k, l\}} \mathbb{E}_Y \left[ h_{kl|i}^{1+\delta} \right] < \infty$ , then

$$\mathbb{E}_{\pi} [p\ell_{M^s}(\theta)] = p\ell_{M^c}(\theta).$$

*Proof.* The proof is almost exactly the same as Lemma 3.4. Observe

$$\begin{aligned}
\mathbb{E}_\pi [p\ell_{M^s}(\boldsymbol{\theta})] &= \mathbb{E}_\pi \left[ \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta}) \right] \\
&= \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta}) \\
&= p\ell_{M^c}(\boldsymbol{\theta}).
\end{aligned}$$

□

**Corollary 4.21.** Let  $h_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \Theta\}} \|\nabla_{\boldsymbol{\theta}} p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta})\|$  and  $g_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \Theta\}} \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y [h_{kl|i}^{1+\delta}] < \infty$  and  $\sup_{\{i,k,l\}} \mathbb{E}_Y [g_{kl|i}^{1+\delta}] < \infty$ , then the sample weighted pairwise score function is a design-unbiased estimator for the census pairwise score function under the design measure  $\pi$ , i.e.,

$$\begin{aligned}
\mathbb{E}_\pi [\nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta})] &= \nabla_{\boldsymbol{\theta}} p\ell_{M^c}(\boldsymbol{\theta}), \\
\mathbb{E}_\pi [\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta})] &= \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^c}(\boldsymbol{\theta}).
\end{aligned}$$

### 4.3 Consistency

My goal in this section is to establish the sample weighted pairwise log-likelihood estimator is a consistent estimator. Let  $\hat{\boldsymbol{\theta}}_n$  be the sample weighted pairwise likelihood estimator and  $\boldsymbol{\theta}_0$  be the true value of the model. More precisely,

**Definition 4.22.** The sample weighted pairwise log-likelihood estimator  $\hat{\boldsymbol{\theta}}_n$  of  $\boldsymbol{\theta}$  is defined as a solution of

$$\frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) = \mathbf{0}.$$

**Remark.** It can be shown that the true parameter  $\boldsymbol{\theta}_0$  of  $\boldsymbol{\theta}$  is a solution of

$$\mathbb{E}_Y \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^c}(\boldsymbol{\theta}) \right] = \mathbf{0}.$$

In all of the following, I assume the number of elements within any clusters is bounded, both the sample and population clusters need to diverge and the sampling fraction for the cluster should converge, i.e.,  $N_i \leq \lambda$  and  $T_i \leq \lambda$  all  $i$ ,  $n_I \rightarrow \infty$ ,  $N_I \rightarrow \infty$ ,  $\frac{n_I}{N_I} \rightarrow c$ ,  $t_s \rightarrow \infty$ ,  $T_c \rightarrow \infty$ ,  $\frac{t_s}{T_c} = O\left(\frac{n_I}{N_I}\right)$  and where  $\lambda > 0$  and  $c \in [0, 1)$ .

Let us state the main result.

**Theorem 4.23.** *Under the following regularity conditions,*

**A.1**  $\Theta$  is a compact subset of  $\mathbb{R}^p$  and  $\theta_0$  is an interior point of  $\Theta$ .

**A.2** Let  $h_{kl|i}(y_k, y_l) = \sup_{\{\theta \in \Theta\}} \|\nabla_{\theta} p \ell_{kl|i}(y_k, y_l, \theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y \left[ h_{kl|i}^{1+\delta} \right] < \infty$  and  $\sup_i \mathbb{E}_Y \|\mathbf{Y}_i\|^\delta < \infty$ , where  $\mathbf{Y}_i = \{y_k : k \in M_i^c\}$ .

**A.3** For any given  $c > 0$  and a given sequence  $\{\mathbf{y}_i\}$  satisfying  $\|\mathbf{y}_i\| \leq c$ , the sequence of function  $\{\nabla_{\theta} p \ell_{kl|i}(y_k, y_l, \theta)\}$  is equicontinuous on any open subsets  $A$  of  $\Theta$ .

**A.4** For any variable  $V_{kl|i}$  satisfy  $\frac{1}{T_I} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} V_{kl|i}^2 = O_p(1)$ , one has

$$\frac{1}{T_I} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} V_{kl|i} - \frac{1}{T_I} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} V_{kl|i} = O_p\left(t_s^{-\frac{1}{2}}\right)$$

with respect to design probability  $\pi$ .

**A.5** The number of element within any clusters is bounded, i.e.,  $\sup_i N_i \leq \lambda$  and  $\sup_i T_i \leq \lambda$  some  $\lambda > 0$ .

**A.6** For all  $\epsilon > 0$ , there exists a  $\delta > 0$  such that

$$\inf_{\{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon\}} \left\| \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\theta} p \ell_{M^s}(\theta) \right] \right\| > \delta.$$

then

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \tag{4.1}$$

under model-design measure  $Y\pi$ .

**Remark.** [A.2](#) can be relaxed to that the second moment is finite.

I proceed the same way as before to argue the sample weighted pairwise likelihood estimator is consistent by establishing a uniform law of large numbers (ULLN).

**Lemma 4.24.** *With the same conditions [A.1](#) – [A.5](#) as in Theorem [4.23](#), then one has*

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\| \frac{1}{T_I} \nabla p \ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla p \ell_{M^s}(\boldsymbol{\theta}) \right] \right\| \xrightarrow{p} 0 \quad (4.2)$$

*with respect to model-design probability  $Y\pi$ .*

*Proof.* Essentially, this can be done in the same way as Lemma [3.9](#). I omit the detail.  $\square$

Let us turn to the proof of Theorem [4.23](#).

*Proof.* The proof is essentially the same as Theorem [3.8](#).  $\square$

## 4.4 Asymptotic normality

My next goal is to establish the asymptotic normality of the sample weighted pairwise likelihood estimator. The asymptotic distribution can be constructed from a second-order Taylor series expansion at the true value  $\boldsymbol{\theta}_0$ . I slightly modify the argument from [Rubin-Bleuer and Kratina \(2005\)](#); [Boistard et al. \(2017\)](#).

**Theorem 4.25.** *Under the following regularity conditions,*

**A.1**  $\Theta$  is a compact subset of  $\mathbb{R}^p$  and  $\boldsymbol{\theta}_0$  is an interior point of  $\Theta$ .

**A.2** Let  $h_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \Theta\}} \|\nabla_{\boldsymbol{\theta}} p \ell_{kl|i}(y_k, y_l, \boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i,k,l\}} \mathbb{E}_Y \left[ h_{kl|i}^{2+\delta} \right] < \infty$  and  $\sup_i \mathbb{E}_Y \|\mathbf{Y}_i\|^\delta < \infty$ , where  $\mathbf{Y}_i = \{y_k : k \in M_i^c\}$ .

**A.3** For any given  $c > 0$  and a given sequence  $\{\mathbf{y}_i\}$  satisfying  $\|\mathbf{y}_i\| \leq c$ , the sequence of function  $\{\nabla_{\boldsymbol{\theta}} p \ell_{kl|i}(y_k, y_l, \boldsymbol{\theta})\}$  is equicontinuous on any open subsets  $A$  of  $\Theta$ .

**A.4** For any variable  $V_{kl|i}$  satisfy  $\frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} V_{kl|i}^2 = O_p(1)$ , one has

$$\frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} V_{kl|i} - \frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} V_{kl|i} = O_p(t_s^{-\frac{1}{2}})$$

with respect to design probability  $\pi$ .

**A.5** The number of element within any clusters is bounded, i.e.,  $\sup_i N_i \leq \lambda$  and  $\sup_i T_i \leq \lambda$  some  $\lambda > 0$ .

**A.6** For all  $\epsilon > 0$ , there exists a  $\delta > 0$  such that

$$\inf_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \epsilon\}} \left\| \mathbb{E}_{Y\pi} \left[ \frac{1}{T_1} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right] \right\| > \delta.$$

**A.7** Let  $g_{kl|i}(y_k, y_l) = \sup_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta}\}} \|\nabla_{\boldsymbol{\theta}}^2 p \ell_{kl|i}(y_k, y_l, \boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{i, k, l\}} \mathbb{E}_Y \left[ g_{kl|i}^{2+\delta} \right] < \infty$  and  $\sup_i \mathbb{E}_Y \|\mathbf{Y}_i\|^\delta < \infty$ , where  $\mathbf{Y}_i = \{y_k : k \in M_i^c\}$ .

**A.8** For any given  $c > 0$  and a given sequence  $\{\mathbf{y}_i\}$  satisfying  $\|\mathbf{y}_i\| \leq c$ , the sequence of function  $\{\nabla_{\boldsymbol{\theta}}^2 p \ell_{kl|i}(y_k, y_l, \boldsymbol{\theta})\}$  is equicontinuous on any open subset  $A$  of  $\boldsymbol{\Theta}$ .

**A.9** For any variable  $V_{kl|i}$  satisfies the following conditions,

$$(a) \quad \frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} V_{kl|i}^2 = O_p(1),$$

$$(b) \quad \lim t_s \sigma_\pi^2 > 0, \text{ where } \sigma_\pi^2 = \text{Var}_\pi \left( \frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} V_{kl|i} \right).$$

then

$$\sigma_\pi^{-1} \left( \frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} V_{kl|i} - \frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} V_{kl|i} \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I})$$

with respect to design probability  $\pi$ .



**A.10**  $\lim t_s \mathbf{J}_\pi > \mathbf{0}$  and  $\lim T_c \mathbf{J}_Y > \mathbf{0}$ , where  $\mathbf{J}_\pi(\boldsymbol{\theta}) = \text{Var}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right]$  and  $\mathbf{J}_Y(\boldsymbol{\theta}) = \text{Var}_Y \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^c}(\boldsymbol{\theta}) \right]$ .

**A.11** Assume  $\lim \frac{t_s}{T_c} = \zeta$ , where  $\zeta \in [0, 1]$ .

then

$$\mathbf{J}_\pi(\boldsymbol{\theta}_0)^{-\frac{1}{2}} \mathbf{H}(\boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N \left( \mathbf{0}, \mathbf{I} + \zeta \left[ (\lim t_s \mathbf{J}_\pi(\boldsymbol{\theta}_0))^{-1} (\lim T_c \mathbf{J}_Y(\boldsymbol{\theta}_0)) \right] \right), \quad (4.3)$$

where

$$\begin{aligned} \mathbf{H}(\boldsymbol{\theta}) &= \mathbb{E}_{Y\pi} \left[ -\frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}) \right], \\ \mathbf{J}_\pi(\boldsymbol{\theta}) &= \text{Var}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right], \\ \mathbf{J}_Y(\boldsymbol{\theta}) &= \text{Var}_Y \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^c}(\boldsymbol{\theta}) \right]. \end{aligned}$$

In particular, if the first-stage sampling fraction is small, then

$$\mathbf{J}_\pi(\boldsymbol{\theta}_0)^{-\frac{1}{2}} \mathbf{H}(\boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}) \quad (4.4)$$

under model-design measure  $Y\pi$ .

**Remark. A.2** is a standard  $2 + \delta$  moment assumption for the central limit theorem to hold. Observe  $2 + \delta$  moment bound is crucial, it is not enough to assume second moment exists.

**Remark. A.7** and **A.8** are used to prove the Uniform Law of Large Numbers (ULLN) for  $\frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta})$ .

**Remark.** In view of Theorem 7.21, I may assume  $p = 1$ .

Before I prove this theorem, I need two results to start with. Once this has been done, the rest of the proof is just routine.

**Lemma 4.26.** Under **A.1–A.8** of Theorem 4.25, then

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\| \frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}) \right] \right\| \xrightarrow{p} 0$$

under the model-design measure  $Y\pi$ .

*Proof.* The argument is essentially the same as the argument given in Lemma 3.9. I omit the details.  $\square$

**Lemma 4.27.** Under the same conditions as in Theorem 4.25, then

$$\mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N \left( \mathbf{0}, \mathbf{I} + \zeta \left[ (\lim t_s \mathbf{J}_\pi(\boldsymbol{\theta}))^{-1} (\lim T_c \mathbf{J}_Y(\boldsymbol{\theta})) \right] \right)$$

under model-design measure  $Y\pi$ . In particular, if the first-stage sampling fraction is small, then

$$\mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I})$$

with respect to model-design measure  $Y\pi$ .

*Proof.* Observe

$$\begin{aligned} & \mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right) \\ &= \mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right) + \\ & \quad \mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right) \\ &= \mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right) + \\ & \quad \left[ \frac{t_s}{T_c} \right]^{\frac{1}{2}} \left[ (t_s \mathbf{J}_\pi(\boldsymbol{\theta}))^{-\frac{1}{2}} (T_c \mathbf{J}_Y(\boldsymbol{\theta}))^{\frac{1}{2}} \right] \mathbf{J}_Y(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right). \end{aligned} \tag{4.5}$$

For the first term in 4.5, by **A.2**, **A.10**, **A.9**, one has

$$\mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}) \tag{4.6}$$

with respect to design measure  $\pi$ .

For the second term in 4.5, observe

$$\mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right] = \frac{1}{T_I} \sum_{i=1}^{T_c} \nabla_{\boldsymbol{\theta}} p \ell_{M_i^c}(\boldsymbol{\theta})$$

is a sum of independent random variables under model measure  $Y$ . By the central limit theorem, one has

$$\mathbf{J}_Y(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right] - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}) \quad (4.7)$$

with respect to model measure  $Y$ .

From A.10 and A.11, one has

$$\left[ \frac{t_s}{T_c} \right]^{\frac{1}{2}} \left[ (t_s \mathbf{J}_\pi(\boldsymbol{\theta}))^{-\frac{1}{2}} (T_c \mathbf{J}_Y(\boldsymbol{\theta}))^{\frac{1}{2}} \right] \xrightarrow{p} \zeta^{\frac{1}{2}} \left[ (\lim t_s \mathbf{J}_\pi(\boldsymbol{\theta}))^{-\frac{1}{2}} (\lim T_c \mathbf{J}_Y(\boldsymbol{\theta}))^{\frac{1}{2}} \right]. \quad (4.8)$$

Putting 4.7 and 4.8 together, by Theorem 4.4 in Billingsley (2013), one has

$$\left[ \frac{t_s}{T_c} \right]^{\frac{1}{2}} \left[ (t_s \mathbf{J}_\pi(\boldsymbol{\theta}))^{-\frac{1}{2}} (T_c \mathbf{J}_Y(\boldsymbol{\theta}))^{\frac{1}{2}} \right] \mathbf{J}_Y(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \mathbb{E}_\pi \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right] - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N\left(0, \zeta \left[ (\lim t_s \mathbf{J}_\pi(\boldsymbol{\theta}))^{-1} (\lim T_c \mathbf{J}_Y(\boldsymbol{\theta})) \right] \right) \quad (4.9)$$

with respect to model measure  $Y$ .

Putting 4.5, 4.6 and 4.9 together, by Theorem 5.1 in Rubin-Bleuer and Kratina (2005), one has

$$\mathbf{J}_\pi(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N\left(0, \mathbf{I} + \zeta \left[ (\lim t_s \mathbf{J}_\pi(\boldsymbol{\theta}))^{-1} (\lim T_c \mathbf{J}_Y(\boldsymbol{\theta})) \right] \right)$$

with respect to model-design measure  $Y\pi$ . This completes the proof.  $\square$

Let us turn to the proof of Theorem 4.25.

*Proof.* Observe

$$\begin{aligned}
0 &= \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\hat{\boldsymbol{\theta}}_n) \\
&= \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}_0) + \frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}_0)^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p\left(t_s^{-\frac{1}{2}}\right) \\
&= \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}_0) + \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}_0) \right]^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \\
&\quad \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}_0) - \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}_0) \right] \right)^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p\left(t_s^{-\frac{1}{2}}\right) \\
&= \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}_0) + \mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}_0) \right]^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p\left(t_s^{-\frac{1}{2}}\right) \\
&= \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}_0) - \mathbf{H}(\boldsymbol{\theta}_0)^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p\left(t_s^{-\frac{1}{2}}\right).
\end{aligned}$$

In the first equality, I used the definition of  $\hat{\boldsymbol{\theta}}_n$ . In the second equality, I applied a Taylor expansion to function  $\frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta})$  at  $\boldsymbol{\theta}_0$ . In the fourth, I used the fact that  $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$  from Theorem 4.23 and A.4. Then one has

$$\mathbf{H}(\boldsymbol{\theta}_0)^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}_0) + o_p\left(t_s^{-\frac{1}{2}}\right),$$

i.e.,

$$\mathbf{J}_{\pi}(\boldsymbol{\theta}_0)^{-\frac{1}{2}} \mathbf{H}(\boldsymbol{\theta}_0)^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{J}_{\pi}(\boldsymbol{\theta}_0)^{-\frac{1}{2}} \left( \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}_0) \right) + o_p(1).$$

By Lemma 4.27 and  $\mathbb{E}_{Y\pi} \left[ \frac{1}{T_I} \nabla_{\boldsymbol{\theta}} p\ell_{M^s}(\boldsymbol{\theta}_0) \right] = \mathbf{0}$ , one has

$$\mathbf{J}_{\pi}(\boldsymbol{\theta}_0)^{-\frac{1}{2}} \mathbf{H}(\boldsymbol{\theta}_0)^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \mathbf{I} + \zeta \left[ (\lim t_s \mathbf{J}_{\pi}(\boldsymbol{\theta}_0))^{-1} (\lim T_c \mathbf{J}_Y(\boldsymbol{\theta}_0)) \right] \right)$$

under model-design measure  $Y\pi$ . □

**Remark.** There are two problems on evaluating  $\mathbf{H}(\boldsymbol{\theta}_0)$ ,  $\mathbf{J}_{\pi}(\boldsymbol{\theta}_0)$  and  $\mathbf{J}_Y(\boldsymbol{\theta}_0)$ . First,  $\mathbf{H}$ ,  $\mathbf{J}_{\pi}$  and  $\mathbf{J}_Y$  are a sum over the population data, but one only observes a sample. Secondly,  $\mathbf{H}$ ,  $\mathbf{J}_{\pi}$  and  $\mathbf{J}_Y$  cannot be evaluated at  $\boldsymbol{\theta}_0$ , because one does not know the true value  $\boldsymbol{\theta}_0$ . This problem can be handled by using the plug-in estimator.

It remains to estimate  $\mathbf{H}(\boldsymbol{\theta}_0)$ ,  $\mathbf{J}_\pi(\boldsymbol{\theta}_0)$  and  $\mathbf{J}_Y(\boldsymbol{\theta}_0)$ . In all of the following, I am going to assume first-stage sampling fraction is small to simplify the exposition. Then it reduces to estimate  $\mathbf{H}(\boldsymbol{\theta}_0)$  and  $\mathbf{J}_\pi(\boldsymbol{\theta}_0)$ . It is straightforward to construct estimator  $\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n)$  such that

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{H}(\boldsymbol{\theta}_0)$$

under model-design measure  $Y\pi$ . My goal in the next section is to construct estimator  $\hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n)$  for  $\mathbf{J}_\pi(\boldsymbol{\theta}_0)$  such that

$$T_c \hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n) (T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0))^{-1} \xrightarrow{p} \mathbf{I}$$

under design measure  $\pi$ . This is surprisingly more difficult than it first seems. Once I prove the limit exists, then one has

$$(T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0))^{-\frac{1}{2}} \left( T_c \hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n) \right)^{\frac{1}{2}} \hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n)^{-\frac{1}{2}} \hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n)^\top (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}),$$

i.e.,

$$\hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n)^{-\frac{1}{2}} \hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n)^\top (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I})$$

under model-design measure  $Y\pi$ .

## 4.5 Variance estimation

It is straightforward to construct a consistent estimator for  $\mathbf{H}(\boldsymbol{\theta}_0)$ . Namely, one can estimate  $\mathbf{H}(\boldsymbol{\theta}_0) = \mathbb{E}_{Y\pi} \left[ -\frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}_0) \right]$  by

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\hat{\boldsymbol{\theta}}_n).$$

The proof of consistency is essentially an rephrase of Lemma 4.26 and the triangle inequality. The difficulty in here is to estimate  $\mathbf{J}(\boldsymbol{\theta}_0)$ .

### 4.5.1 Empirical variance estimation

One can estimate  $\mathbf{H}(\boldsymbol{\theta}_0) = \mathbb{E}_{Y\pi} \left[ -\frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\boldsymbol{\theta}_0) \right]$  by

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{T_I} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 p\ell_{M^s}(\hat{\boldsymbol{\theta}}_n).$$

More precisely,

**Lemma 4.28.** Under [A.1–A.8](#) of Theorem [4.25](#), then

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n) - \mathbf{H}(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0}.$$

*Proof.* Note  $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$  by Theorem [4.23](#). From the Continuous Mapping Theorem [7.20](#), one has

$$\mathbf{H}(\hat{\boldsymbol{\theta}}_n) - \mathbf{H}(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0} \quad (4.10)$$

under model-design measure  $\Upsilon\pi$ .

Hence

$$\begin{aligned} & \mathbb{P}_{\Upsilon\pi} \left( \|\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n) - \mathbf{H}(\boldsymbol{\theta}_0)\| \geq \epsilon \right) \\ & \leq \mathbb{P}_{\Upsilon\pi} \left( \|\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n) - \mathbf{H}(\hat{\boldsymbol{\theta}}_n)\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}_{\Upsilon\pi} \left( \|\mathbf{H}(\hat{\boldsymbol{\theta}}_n) - \mathbf{H}(\boldsymbol{\theta}_0)\| \geq \frac{\epsilon}{2} \right) \\ & \leq \mathbb{P}_{\Upsilon\pi} \left( \sup_{\boldsymbol{\theta} \in \Theta} \|\hat{\mathbf{H}}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}_{\Upsilon\pi} \left( \|\mathbf{H}(\hat{\boldsymbol{\theta}}_n) - \mathbf{H}(\boldsymbol{\theta}_0)\| \geq \frac{\epsilon}{2} \right) \\ & \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ & \leq \epsilon. \end{aligned}$$

The third inequality follows from Lemma [4.26](#) and [4.10](#). □

Therefore, it remains to estimate the design variance  $\mathbf{J}_\pi(\boldsymbol{\theta})$ . Observe

$$\begin{aligned} \mathbf{J}_\pi(\boldsymbol{\theta}) &= \text{Var}_\pi \left( \frac{1}{T_c} \nabla_{\boldsymbol{\theta}} p \ell_{M^s}(\boldsymbol{\theta}) \right) \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} \text{Cov}_\pi(1_{kl}, 1_{k'l'}) (\omega_{kl} \nabla_{\boldsymbol{\theta}} p \ell_{kl}(\boldsymbol{\theta}) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'}(\boldsymbol{\theta})) \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} \Delta_{klk'l'} (\omega_{kl} \nabla_{\boldsymbol{\theta}} p \ell_{kl}(\boldsymbol{\theta}) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'}(\boldsymbol{\theta})), \end{aligned}$$

where

$$\Delta_{klk'l'} = \pi_{klk'l'} - \pi_{kl}\pi_{k'l'}.$$

**Remark.** Observe if one samples all the population data, then  $\mathbf{J}_\pi(\boldsymbol{\theta}) = 0$ , as  $\Delta_{klk'l'} = 0$  for all  $k, l, k', l' \in U$ .

Observe one needs data in the population to calculate  $\mathbf{J}_\pi(\boldsymbol{\theta})$ . But one only observes a sample. Therefore,  $\mathbf{J}_\pi(\boldsymbol{\theta})$  must be estimated by the data from the sample. One can estimate  $\mathbf{J}_\pi(\boldsymbol{\theta})$  by

$$\hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) = \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\boldsymbol{\theta}} p \ell_{kl}(\boldsymbol{\theta}) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'}(\boldsymbol{\theta})).$$

**Remark.** Observe  $\hat{\mathbf{J}}_\pi(\boldsymbol{\theta})$  is an unbiased estimator for  $\mathbf{J}_\pi(\boldsymbol{\theta})$  under design measure  $\pi$ , i.e.,  $\mathbb{E}_\pi [\hat{\mathbf{J}}_\pi(\boldsymbol{\theta})] = \mathbf{J}_\pi(\boldsymbol{\theta})$ .

**Remark.** Observe

$$\begin{aligned} \text{Var}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta})] &= \text{Var}_\pi \left[ \frac{1}{T_c} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\boldsymbol{\theta}} p \ell_{kl}(\boldsymbol{\theta}) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'}(\boldsymbol{\theta})) \right] \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{i''=1}^{T_c} \sum_{i'''=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} \sum_{\substack{k'' < l'' \\ k'', l'' \in M_{i''}^c}} \sum_{\substack{k''' < l''' \\ k''', l''' \in M_{i'''}^c}} \text{Cov}_\pi (1_{klk'l'}, 1_{k''l''k'''l'''}) \omega_{klk'l'} \\ &\quad \Delta_{klk'l'} \omega_{k''l''k'''l'''} \Delta_{k''l''k'''l'''} (\omega_{kl} \nabla_{\boldsymbol{\theta}} p \ell_{kl}(\boldsymbol{\theta}) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'}(\boldsymbol{\theta})) \\ &\quad (\omega_{k''l''} \nabla_{\boldsymbol{\theta}} p \ell_{k''l''}(\boldsymbol{\theta}) \omega_{k'''l'''} \nabla_{\boldsymbol{\theta}} p \ell_{k'''l'''}(\boldsymbol{\theta})) \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{i''=1}^{T_c} \sum_{i'''=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} \sum_{\substack{k'' < l'' \\ k'', l'' \in M_{i''}^c}} \sum_{\substack{k''' < l''' \\ k''', l''' \in M_{i'''}^c}} \Delta_{klk'l'k''l''k'''l'''} \omega_{klk'l'} \Delta_{klk'l'} \omega_{k''l''k'''l'''} \\ &\quad \Delta_{k''l''k'''l'''} (\omega_{kl} \nabla_{\boldsymbol{\theta}} p \ell_{kl}(\boldsymbol{\theta}) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p \ell_{k'l'}(\boldsymbol{\theta})) (\omega_{k''l''} \nabla_{\boldsymbol{\theta}} p \ell_{k''l''}(\boldsymbol{\theta}) \omega_{k'''l'''} \nabla_{\boldsymbol{\theta}} p \ell_{k'''l'''}(\boldsymbol{\theta})), \end{aligned}$$

where

$$\Delta_{klk'l'k''l''k'''l'''} = \pi_{klk'l'k''l''k'''l'''} - \pi_{klk'l'} \pi_{k''l''k'''l'''}.$$

A natural question to ask in here is: when does one have a convergence in probability? More precisely, can we show

$$T_c \hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0) \xrightarrow{p} 0$$

under design measure  $\pi$ ? By Markov's Inequality Theorem 7.17, it boils down to show under what condition does one have  $\text{Var}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta})] \rightarrow 0$  for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . The question is surprisingly more subtle than it first seems. I will come back to this question on next two sections. More precisely, I will show that a set of reasonable sampling designs (Poisson, Stratified and SRSWOR) satisfy this condition (more about this in section 4.6 and 4.7). For the moment, let me suppose  $\text{Var}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta})] \rightarrow 0$  for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ .

**Lemma 4.29.** *With the same conditions as in Theorem 4.25 and  $\text{Var}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta})] \rightarrow 0$  for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , then one has*

$$T_c \hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0) \xrightarrow{p} 0$$

*under design measure  $\pi$ .*

To prove Lemma 4.29, I need a uniform convergence lemma to start with. In all of the following, without loss of generality, one may assume  $\boldsymbol{\theta}$  is 1-dimensional by Cramer-Wold Theorem.

**Lemma 4.30.** *With the same conditions as in Lemma 4.29, then one has*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) - T_c \mathbf{J}_\pi(\boldsymbol{\theta})\| \xrightarrow{p} 0.$$

*Proof.* First, I want to show  $T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}) \rightarrow 0$ . To see this, observe

$$\begin{aligned} \mathbb{P}_\pi \left( \|T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) - T_c \mathbf{J}_\pi(\boldsymbol{\theta})\| \geq \epsilon \right) &\leq \frac{\mathbb{E}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) - T_c \mathbf{J}_\pi(\boldsymbol{\theta})]^2}{\epsilon^2} \\ &= \frac{\text{Var}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta})]}{\epsilon^2} \rightarrow 0. \end{aligned}$$

In the first inequality, I use the Chebyshev's inequality Theorem 7.17.



Let  $\rho > 0$ , consider the upper and the lower function of the  $T_c \hat{\mathbf{J}}_\pi$  on  $B_\rho(\boldsymbol{\theta})$ , i.e.,

$$\begin{aligned}
U_\rho(\boldsymbol{\theta}) &= \sup_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}') \\
&= \sup_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} \frac{1}{T_c} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\boldsymbol{\theta}} p_{kl}(\boldsymbol{\theta}') \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p_{k'l'}(\boldsymbol{\theta}')), \\
L_\rho(\boldsymbol{\theta}) &= \inf_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}') \\
&= \inf_{\boldsymbol{\theta}' \in B_\rho(\boldsymbol{\theta})} \frac{1}{T_c} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\boldsymbol{\theta}} p_{kl}(\boldsymbol{\theta}') \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p_{k'l'}(\boldsymbol{\theta}')).
\end{aligned}$$

Then one can apply similar argument as in Lemma 3.9 to show

$$\sup_{\boldsymbol{\theta} \in \Theta} \|T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) - T_c \mathbf{J}_\pi(\boldsymbol{\theta})\| \xrightarrow{p} 0.$$

I omit the detail. □

I now proceed to the proof of Lemma 4.29.

*Proof.* Therefore, one has

$$\sup_{\boldsymbol{\theta} \in \Theta} \|T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) - T_c \mathbf{J}_\pi(\boldsymbol{\theta})\| \xrightarrow{p} 0 \quad (4.11)$$

under design measure  $\pi$ .

Note  $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \xrightarrow{p} 0$  under model-design measure  $\gamma\pi$  by Theorem 4.23. By the Continuous Mapping Theorem 7.20, one has

$$T_c \mathbf{J}_\pi(\hat{\boldsymbol{\theta}}_n) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0) \xrightarrow{p} 0 \quad (4.12)$$

under design measure  $\pi$ .

Hence

$$\begin{aligned}
& \mathbb{P}_\pi \left( \left\| T_c \hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0) \right\| \geq \epsilon \right) \\
& \leq \mathbb{P}_\pi \left( \left\| T_c \hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n) - T_c \mathbf{J}_\pi(\hat{\boldsymbol{\theta}}_n) \right\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}_\pi \left( \left\| T_c \mathbf{J}_\pi(\hat{\boldsymbol{\theta}}_n) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0) \right\| \geq \frac{\epsilon}{2} \right) \\
& \leq \mathbb{P}_\pi \left( \sup_{\boldsymbol{\theta} \in \Theta} \left\| T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}) \right\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}_\pi \left( \left\| T_c \mathbf{J}_\pi(\hat{\boldsymbol{\theta}}_n) - T_c \mathbf{J}_\pi(\boldsymbol{\theta}_0) \right\| \geq \frac{\epsilon}{2} \right) \\
& \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
& \leq \epsilon.
\end{aligned}$$

The third inequality follows from 4.11 and 4.12. This completes the proof.  $\square$

We are ready to state the main results.

**Theorem 4.31.** *With the same conditions as in Theorem 4.25. Assume the first-stage sampling fraction for PSU is small and  $\sup_{\boldsymbol{\theta} \in \Theta} \text{Var}_\pi \left[ T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) \right] \rightarrow 0$ , then*

$$\hat{\mathbf{J}}_\pi(\hat{\boldsymbol{\theta}}_n)^{-\frac{1}{2}} \hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n)^T (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, 1)$$

under model-design measure  $Y\pi$ .

*Proof.* This is clear from Theorem 4.25, Lemma 4.28 and Lemma 4.29.  $\square$

## 4.6 Consistency of empirical variance estimation: the sampling clusters are the model clusters

In all of this section, I assume the sampling clusters are the model clusters. My goal is to show Poisson, stratified and SRSWOR sampling all meet the following conditions

$$\text{Var}_\pi \left[ T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) \right] \rightarrow 0.$$

Although the argument given in here cannot be directly applied when the sampling clusters are not the model clusters, they provide an important theme: the tree structure of  $\Delta_{klk'l'}$ ,  $\Delta_{k'l'}$  and  $\Delta_{klk'l'k''l''k'''l'''}$  are crucial for establishing convergence. In all of

the following, I assume the number of elements within any clusters is bounded, i.e.,  $\sup_i N_i \leq \lambda$  and  $\sup_i T_i \leq \lambda$  for some  $\lambda > 0$ . Let  $h_{kl} = \sup_{\theta \in \Theta} \|\nabla_{\theta} p_{\ell_{kl}}(\theta)\|$ . Furthermore, assume there exists a  $c > 0$  such that  $\sup_{kl} h_{kl} \leq c$ ,  $\sup_{kl} \omega_{kl} \leq c$ ,  $\sup_{klk'l'} \omega_{klk'l'} \leq c$  and  $\sup_{klk'l'k''l''k'''l'''} \omega_{klk'l'k''l''k'''l'''} \leq c$ . I first establish  $\text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\theta) \right] \rightarrow 0$  for the Poisson sampling when the sampling clusters are the model clusters.

#### 4.6.1 Example: Poisson sampling design

Assume the sampling clusters are the model clusters, i.e.,  $U_1^c = M_1^c$ . Consider the following two-stage sample design:

- (i) First-stage: Poisson sampling with  $\pi_i$  for each sampling cluster  $i$ .
- (ii) Second-stage: SRSWR with sample size  $n_i = C_1$  from population size  $N_i = C_2$  for the sampling cluster  $i$ .

I start by making two observations.

**Lemma 4.32.** *Assume the sampling clusters are the model clusters, i.e.,  $U_1^c = M_1^c$ . Under the above Poisson sampling setting, then*

- (i) *Suppose  $klk'l'$  has two sampling cluster, i.e.,  $\langle kl \rangle \langle k'l' \rangle$ , then  $\Delta_{klk'l'} = 0$ .*
- (ii) *If  $\langle klk'l' \rangle \langle k''l''k'''l''' \rangle$ , then  $\Delta_{klk'l'k''l''k'''l'''} = 0$ .*

*Proof.* To prove (i), observe

$$\begin{aligned} \Delta_{klk'l'} &= \mathbb{E}_{\pi} 1_{klk'l'} - \mathbb{E}_{\pi} 1_{kl} \mathbb{E}_{\pi} 1_{k'l'} \\ &= \pi_{s(k)} \pi_{s(k')} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} - \pi_{s(k)} \frac{C_1}{C_2} \frac{C_1}{C_2} \pi_{s(k')} \frac{C_1}{C_2} \frac{C_1}{C_2} \\ &= 0. \end{aligned}$$

To prove (ii), observe

$$\begin{aligned} \Delta_{klk'l'k''l''k'''l'''} &= \mathbb{E}_{\pi} 1_{klk'l'k''l''k'''l'''} - \mathbb{E}_{\pi} 1_{klk'l'} \mathbb{E}_{\pi} 1_{k''l''k'''l'''} \\ &= \pi_{s(k)} \pi_{s(k'')} \left( \frac{C_1}{C_2} \right)^8 - \pi_{s(k)} \left( \frac{C_1}{C_2} \right)^4 \pi_{s(k'')} \left( \frac{C_1}{C_2} \right)^4 \\ &= 0. \end{aligned}$$

□

**Proposition 4.33.** Assume the sampling clusters are the model clusters, i.e.,  $U_1^c = M_1^c$ . Under the above Poisson sampling setting, then for all  $\theta \in \Theta$

$$\text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\theta) \right] \rightarrow 0.$$

*Proof.* The proof is based on bounding the number of non-zero terms in a straightforward expansion of the sum defining  $\hat{\mathbf{J}}_{\pi}(\theta)$ . Observe

$$\begin{aligned} \text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\theta) \right] &= \text{Var}_{\pi} \left[ \frac{1}{T_c} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\theta} p_{\ell_{kl}}(\theta) \omega_{k'l'} \nabla_{\theta} p_{\ell_{k'l'}}(\theta)) \right] \\ &= \frac{1}{T_c^2} \text{Var}_{\pi} \left[ \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_i^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\theta} p_{\ell_{kl}}(\theta) \omega_{k'l'} \nabla_{\theta} p_{\ell_{k'l'}}(\theta)) \right] \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} \sum_{\substack{k'' < l'' \\ k'', l'' \in M_i^c}} \sum_{\substack{k''' < l''' \\ k''', l''' \in M_{i'}^c}} \Delta_{klk'l'k''l''k'''l'''} \omega_{klk'l'} \Delta_{klk'l'} \omega_{k''l''k'''l'''} \\ &\quad \Delta_{k''l''k'''l'''} (\omega_{kl} \nabla_{\theta} p_{\ell_{kl}}(\theta) \omega_{k'l'} \nabla_{\theta} p_{\ell_{k'l'}}(\theta)) (\omega_{k''l''} \nabla_{\theta} p_{\ell_{k''l''}}(\theta) \omega_{k'''l'''} \nabla_{\theta} p_{\ell_{k'''l'''}}(\theta)) \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_i^c}} \sum_{\substack{k'' < l'' \\ k'', l'' \in M_i^c}} \sum_{\substack{k''' < l''' \\ k''', l''' \in M_i^c}} \Delta_{klk'l'k''l''k'''l'''} \omega_{klk'l'} \Delta_{klk'l'} \omega_{k''l''k'''l'''} \\ &\quad \Delta_{k''l''k'''l'''} (\omega_{kl} \nabla_{\theta} p_{\ell_{kl}}(\theta) \omega_{k'l'} \nabla_{\theta} p_{\ell_{k'l'}}(\theta)) (\omega_{k''l''} \nabla_{\theta} p_{\ell_{k''l''}}(\theta) \omega_{k'''l'''} \nabla_{\theta} p_{\ell_{k'''l'''}}(\theta)) \\ &= O\left(\frac{1}{T_c}\right). \end{aligned}$$

In the second equality, I used Lemma 4.32, i.e.,  $\Delta_{klk'l'} = 0$  if  $\langle kl \rangle \times \langle k'l' \rangle$ . In the fourth equality, I used Lemma 4.32, i.e.,  $\Delta_{klk'l'k''l''k'''l'''} = 0$  if  $\langle klk'l' \rangle \times \langle k''l''k'''l''' \rangle$ . □

**Remark.** The argument above can be extended to the case when the sampling clusters  $U_1^c$  are coarser than the model clusters  $M_1^c$ , i.e.,  $M_1^c \subset U_1^c$ . The idea is pairs in the same model clusters must be in the same sampling clusters.

**Remark.** When the sampling clusters are the model clusters, I used an independence identity, basically relying on the independent pair from different clusters. In particular, note the tree structure of the  $\Delta_{klk'l'}$ ,  $\Delta_{k''l''k'''l'''}$  and  $\Delta_{klk'l'k''l''k'''l'''}$  is vital. See the figure below. However, this argument no longer holds when the sampling clusters are not the same as the model clusters.

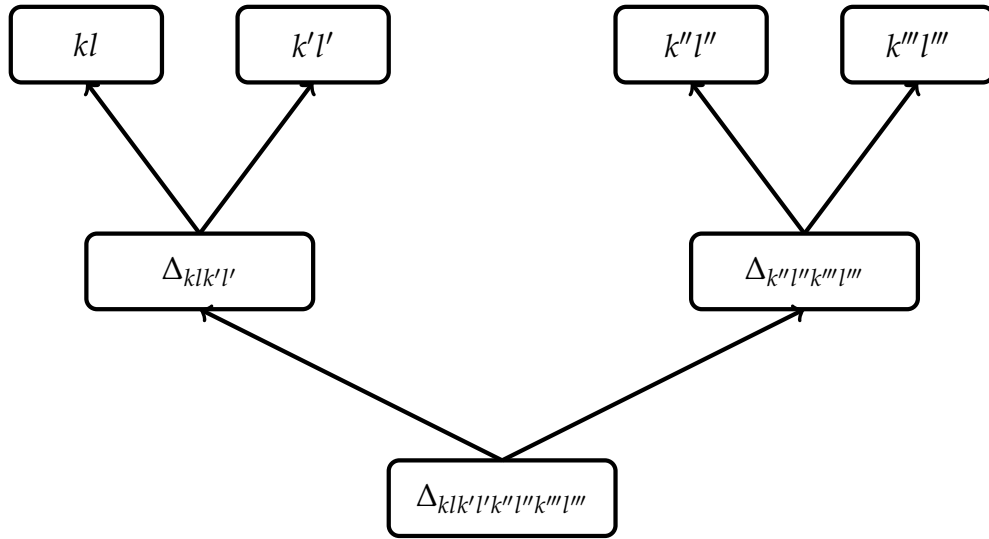


Figure 4.1: Tree structure.

#### 4.6.2 Example: SRSWOR sampling design

Assume the sampling clusters are the model clusters, i.e.,  $U_1^c = M_1^c$ . Consider the following two-stage sample design:

- (i) First-stage: SRSWOR with the sample size  $n_1$  from the population size  $N_1$  for the sampling clusters.
- (ii) Second-stage: SRSWR with the sample size  $n_i = C_1$  from the population size  $N_i = C_2$  for sampling cluster  $i$ .

I start by making two observations.

**Lemma 4.34.** Assume the sampling clusters are the model clusters, i.e.,  $U_1^c = M_1^c$ . Under the above SRSWOR sampling setting,

(i) Suppose  $klk'l'$  has two sampling clusters, i.e.,  $\langle kl \rangle \langle k'l' \rangle$ , then  $\Delta_{klk'l'} = O(T_c^{-1})$ .

(ii) If  $\langle klk'l' \rangle \langle k''l''k'''l''' \rangle$ , then  $\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1})$ .

*Proof.* To prove (i), observe

$$\begin{aligned}
\Delta_{klk'l'} &= \mathbb{E}_\pi 1_{klk'l'} - \mathbb{E}_\pi 1_{kl} \mathbb{E}_\pi 1_{k'l'} \\
&= \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I - 1}{N_I - 1} \frac{C_1}{C_2} \frac{C_1}{C_2} - \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \\
&= \frac{t_s}{T_c} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{t_s - 1}{T_c - 1} \frac{C_1}{C_2} \frac{C_1}{C_2} - \frac{t_s}{T_c} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{t_s}{T_c} \frac{C_1}{C_2} \frac{C_1}{C_2} \\
&= c \frac{t_s}{T_c} \left( \frac{t_s - 1}{T_c - 1} - \frac{t_s}{T_c} \right) \\
&= -c \frac{1}{T_c - 1} \frac{t_s}{T_c} \left( 1 - \frac{t_s}{T_c} \right) \\
&= O(T_c^{-1}).
\end{aligned}$$

To prove (ii), observe

$$\begin{aligned}
\Delta_{klk'l'k''l''k'''l'''} &= \mathbb{E}_\pi 1_{klk'l'k''l''k'''l'''} - \mathbb{E}_\pi 1_{klk'l'} \mathbb{E}_\pi 1_{k''l''k'''l'''} \\
&= \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I - 1}{N_I - 1} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} - \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \\
&= \frac{t_s}{T_c} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{t_s - 1}{T_c - 1} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} - \frac{t_s}{T_c} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{t_s}{T_c} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{C_1}{C_2} \\
&= c \frac{t_s}{T_c} \left( \frac{t_s - 1}{T_c - 1} - \frac{t_s}{T_c} \right) \\
&= -c \frac{1}{T_c - 1} \frac{t_s}{T_c} \left( 1 - \frac{t_s}{T_c} \right) \\
&= O(T_c^{-1}).
\end{aligned}$$

□

**Proposition 4.35.** Assume the sampling clusters are the model clusters, i.e.,  $U_i^c = M_i^c$ . Under the above SRSWOR sampling setting, then for all  $\theta \in \Theta$

$$\text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\theta) \right] \rightarrow 0.$$

*Proof.* The proof is based on splitting the sum into four pieces and counting the number of terms in each pieces. Recall

$$\begin{aligned} \text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\theta) \right] &= \text{Var}_{\pi} \left[ \frac{1}{T_c} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\theta} p \ell_{kl}(\theta) \omega_{k'l'} \nabla_{\theta} p \ell_{k'l'}(\theta)) \right] \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{i''=1}^{T_c} \sum_{i'''=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} \sum_{\substack{k'' < l'' \\ k'', l'' \in M_{i''}^c}} \sum_{\substack{k''' < l''' \\ k''', l''' \in M_{i'''}^c}} \Delta_{klk'l'k''l''k'''l'''} \omega_{klk'l'} \Delta_{klk'l'} \omega_{k''l''k'''l'''} \\ &\quad \Delta_{k''l''k'''l'''} (\omega_{kl} \nabla_{\theta} p \ell_{kl}(\theta) \omega_{k'l'} \nabla_{\theta} p \ell_{k'l'}(\theta)) (\omega_{k''l''} \nabla_{\theta} p \ell_{k''l''}(\theta) \omega_{k'''l'''} \nabla_{\theta} p \ell_{k'''l'''}(\theta)). \end{aligned}$$

I may split the sum over the model cluster into four pieces

$$\frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{i''=1}^{T_c} \sum_{i'''=1}^{T_c} = \frac{1}{T_c^2} \left( \sum_{\text{one distinct model cluster}} + \sum_{\text{two distinct model cluster}} + \sum_{\text{three distinct model cluster}} + \sum_{\text{four distinct model cluster}} \right) \quad (4.13)$$

and show the contribution of each term is small.

For the **first** sum in 4.13, there is only one sampling cluster. Hence

$$\frac{1}{T_c^2} \sum_{\text{one distinct model cluster}} = \frac{1}{T_c^2} O(T_c) = O(T_c^{-1}).$$

For the **second** sum in 4.13, there are two sampling clusters. There are three possibilities.

First, if both  $klk'l'$  and  $k''l''k'''l'''$  have one sampling cluster, i.e.,  $\langle klk'l' \rangle \langle k''l''k'''l''' \rangle$ , then  $\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.34. Hence

$$\frac{1}{T_c^2} \sum_{\text{two distinct model cluster}} = \frac{1}{T_c^2} O(T_c^2) O(T_c^{-1}) = O(T_c^{-1}).$$

Secondly, if one of  $klk'l'$  or  $k''l''k'''l'''$  has one cluster and the other has two clusters. WLOG, assume  $klk'l'$  has one cluster and  $k''l''k'''l'''$  has two clusters, then  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.34. Hence

$$\frac{1}{T_c^2} \sum_{\text{two distinct model cluster}} = \frac{1}{T_c^2} O(T_c^2) O(T_c^{-1}) = O(T_c^{-1}).$$

Thirdly, if both  $klk'l'$  and  $k''l''k'''l'''$  have two clusters, then  $\Delta_{klk'l'} = O(T_c^{-1})$  and  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.34. Hence

$$\frac{1}{T_c^2} \sum_{\text{two distinct model cluster}} = \frac{1}{T_c^2} O(T_c^2) O(T_c^{-1}) O(T_c^{-1}) = O(T_c^{-2}).$$

For the **third** sum in 4.13, there are three sampling clusters. Observe at least one of  $klk'l'$  or  $k''l''k'''l'''$  have two clusters. Without loss of generality, suppose  $klk'l'$  has two clusters, then  $\Delta_{klk'l'} = O(T_c^{-1})$  by Lemma 4.34. There are two possibilities here.

First, suppose  $k''l''k'''l'''$  has only one sampling cluster, i.e.,  $\langle kl \rangle \times \langle k'l' \rangle \times \langle k''l''k'''l''' \rangle$ , then

$$\begin{aligned} & \Delta_{klk'l'k''l''k'''l'''} \\ &= \mathbb{E}_\pi 1_{klk'l'k''l''k'''l'''} - \mathbb{E}_\pi 1_{klk'l'} \mathbb{E}_\pi 1_{k''l''k'''l'''} \\ &= \frac{n_I}{N_I} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I - 1}{N_I - 1} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I - 2}{N_I - 2} \left( \frac{C_1}{C_2} \right)^4 - \frac{n_I}{N_I} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I - 1}{N_I - 1} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I}{N_I} \left( \frac{C_1}{C_2} \right)^4 \\ &= \frac{t_s}{T_c} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s - 1}{T_c - 1} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s - 2}{T_c - 2} \left( \frac{C_1}{C_2} \right)^4 - \frac{t_s}{T_c} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s - 1}{T_c - 1} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s}{T_c} \left( \frac{C_1}{C_2} \right)^4 \\ &= c \frac{t_s}{T_c} \frac{t_s - 1}{T_c - 1} \left( \frac{t_s - 2}{T_c - 2} - \frac{t_s}{T_c} \right) \\ &= -c \frac{1}{T_c - 2} \frac{t_s}{T_c} \frac{t_s - 1}{T_c - 1} \left( 2 - \frac{2t_s}{T_c} \right) \\ &= O(T_c^{-1}). \end{aligned}$$

Hence

$$\frac{1}{T_c^2} \sum_{\text{three distinct model cluster}} = \frac{1}{T_c^2} O(T_c^3) O(T_c^{-1}) O(T_c^{-1}) = O(T_c^{-1}).$$



Secondly, suppose  $k''l''k'''l'''$  has only two sampling clusters, then  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.34. Hence

$$\frac{1}{T_c^2} \sum_{\text{three distinct model cluster}} = \frac{1}{T_c^2} O(T_c^3) O(T_c^{-1}) O(T_c^{-1}) = O(T_c^{-1}).$$

For the **fourth** sum in 4.13, there are four sampling clusters, i.e.,  $\langle kl \rangle \langle k'l' \rangle \langle k''l'' \rangle \langle k'''l''' \rangle$ . Then

$$\begin{aligned} & \Delta_{klk'l'k''l''k'''l'''} \\ &= \mathbb{E}_\pi 1_{klk'l'k''l''k'''l'''} - \mathbb{E}_\pi 1_{klk'l'} \mathbb{E}_\pi 1_{k''l''k'''l'''} \\ &= \frac{n_I}{N_I} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I - 1}{N_I - 1} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I - 2}{N_I - 2} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I - 3}{N_I - 3} \left( \frac{C_1}{C_2} \right)^2 - \left( \frac{n_I}{N_I} \left( \frac{C_1}{C_2} \right)^2 \frac{n_I - 1}{N_I - 1} \left( \frac{C_1}{C_2} \right)^2 \right)^2 \\ &= \frac{t_s}{T_c} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s - 1}{T_c - 1} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s - 2}{T_c - 2} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s - 3}{T_c - 3} \left( \frac{C_1}{C_2} \right)^2 - \left( \frac{t_s}{T_c} \left( \frac{C_1}{C_2} \right)^2 \frac{t_s - 1}{T_c - 1} \left( \frac{C_1}{C_2} \right)^2 \right)^2 \\ &= c \frac{t_s}{T_c} \frac{t_s - 1}{T_c - 1} \left( \frac{t_s - 2}{T_c - 2} \frac{t_s - 3}{T_c - 3} - \frac{t_s}{T_c} \frac{t_s - 1}{T_c - 1} \right) \\ &= O(T_c^{-1}). \end{aligned}$$

Observe  $\Delta_{klk'l'} = O(T_c^{-1})$  and  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.34. Therefore

$$\frac{1}{T_c^2} \sum_{\text{four distinct model cluster}} = \frac{1}{T_c^2} O(T_c^4) O(T_c^{-1}) O(T_c^{-1}) O(T_c^{-1}) = O(T_c^{-1}).$$

Putting all those estimates together, one has

$$\text{Var}_\pi \left[ T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) \right] = O(T_c^{-1}).$$

This completes the proof.  $\square$

**Remark.** When the sampling clusters are the model clusters, I used an almost independence identity, basically relying on some of term  $\Delta_{klk'l'}$ ,  $\Delta_{k''l''k'''l'''}$  or  $\Delta_{klk'l'k''l''k'''l'''}$  being  $O(T_c^{-1})$ . In particular, note the tree structure of the  $\Delta_{klk'l'}$ ,  $\Delta_{k''l''k'''l'''}$  and  $\Delta_{klk'l'k''l''k'''l'''}$  is vital. However, this argument no longer holds when the sampling clusters are not the same as the model clusters.

**Remark.** One can use a similar argument to show  $\text{Var}_\pi \left[ T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) \right] = O(T_c^{-1})$  for stratified SRSROW. I omit the detail.

## 4.7 Consistency of empirical variance estimation: the sampling clusters are not the same as the model clusters

When the sampling clusters are not the same as the model clusters, the argument is more involved to establish  $\text{Var}_\pi \left[ T_c \hat{\mathbf{J}}_1(\boldsymbol{\theta}) \right] \rightarrow 0$ . It is complicated by the structure of sampling design, i.e., pair  $kl$  in the same model cluster might not be in the same sampling cluster. I will show  $\text{Var}_\pi \left[ T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) \right] \rightarrow 0$  for Poisson, SRSWOR and stratified sampling. The proof consists of rewriting  $\text{Var}_\pi \left[ T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta}) \right]$  as the sum of terms which one can explore the decay for  $\Delta_{klk'l'}, \Delta_{k''l''k'''l'''} and  $\Delta_{klk'l'k''l''k'''l'''}$ . The key construction in here is inspired by Lumley (1998); Lumley and Mayer Hamblett (2003).$

### 4.7.1 Example: Poisson sampling design

Assume the sampling clusters are not the same as the model clusters, i.e.,  $U_1^c \neq M_1^c$ . Consider the following two-stage sample design:

- (i) First-stage: Poisson sampling with  $\pi_i$  for the sampling cluster  $i$ .
- (ii) Second-stage: SRSWR with sample size  $n_i = C_1$  from population size  $N_i = C_2$  for sampling cluster  $i$ .

I start by introducing some notation. Define

$$\mathcal{P} = \{kl \in U^2 : k < l \text{ and } m(k) = m(l)\}.$$

**Definition 4.36.** Let  $kl \in \mathcal{P}$ , define the neighbourhood of pair  $kl$  to be a set  $\mathcal{S}_{kl} \subset \mathcal{P}$  such that

- (i) If  $k'l' \in \mathcal{S}_{kl}$ , then  $kl \in \mathcal{S}_{k'l'}$ .
- (ii) If  $kl \notin \mathcal{S}_{k'l'}$  and  $k'l' \notin \mathcal{S}_{kl}$ , then  $1_{kl}$  and  $1_{k'l'}$  are independent under design measure  $\pi$ .

Define  $\bar{\mathcal{S}}_{kl}$  to be the complement of  $\mathcal{S}_{kl}$  in  $\mathcal{P}$ , i.e.,  $\bar{\mathcal{S}}_{kl} = \mathcal{P} \setminus \mathcal{S}_{kl}$ .

**Remark.** Under the above Poisson sampling setting,  $\mathcal{S}_{kl}$  is the set of pairs that share at least one sampling cluster of  $kl$ , i.e.,

$$\mathcal{S}_{kl} = \{k'l' \in \mathcal{P} : s(k') = s(k) \text{ or } s(k') = s(l) \text{ or } s(l') = s(k) \text{ or } s(l') = s(l)\}.$$

In particular, observe  $\sup_{kl \in \mathcal{P}} |\mathcal{S}_{kl}| \leq \lambda^2$  and  $\sup_{kl \in \mathcal{P}} |\bar{\mathcal{S}}_{kl}| \leq T_c \lambda^2$ .

**Lemma 4.37.** *If  $k'l' \in \bar{\mathcal{S}}_{kl}$ , then  $\Delta_{klk'l'} = 0$ .*

*Proof.* Essentially, this is just chasing down the definition. If  $k'l' \in \bar{\mathcal{S}}_{kl}$ , i.e.,  $k'l' \notin \mathcal{S}_{kl}$ , then  $kl \notin \mathcal{S}_{k'l'}$  from the definition. Then from the definition, this implies  $1_{kl}$  and  $1_{k'l'}$  are independent, i.e.,  $\Delta_{klk'l'} = 0$ .  $\square$

**Proposition 4.38.** *Under the above Poisson sampling setting, then for all  $\theta \in \Theta$*

$$\text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\theta) \right] \rightarrow 0.$$

*Proof.* Observe

$$\begin{aligned} & \text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\theta) \right] \\ &= \text{Var}_{\pi} \left[ \frac{1}{T_c} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\theta} p_{\ell_{kl}}(\theta) \omega_{k'l'} \nabla_{\theta} p_{\ell_{k'l'}}(\theta)) \right] \\ &= \frac{1}{T_c^2} \text{Var}_{\pi} \left[ \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \left( \sum_{k'l' \in \mathcal{S}_{kl}} + \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \right) 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\theta} p_{\ell_{kl}}(\theta) \omega_{k'l'} \nabla_{\theta} p_{\ell_{k'l'}}(\theta)) \right] \\ &= \frac{1}{T_c^2} \text{Var}_{\pi} \left[ \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\theta} p_{\ell_{kl}}(\theta) \omega_{k'l'} \nabla_{\theta} p_{\ell_{k'l'}}(\theta)) \right] \\ &= \frac{c}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{i''=1}^{T_c} \sum_{\substack{k'' < l'' \\ k'', l'' \in M_{i''}^c}} \sum_{k'''l''' \in \mathcal{S}_{k''l''}} \text{Cov}_{\pi}(1_{kl} 1_{k'l'}, 1_{k''l''} 1_{k'''l''''}) \omega_{klk'l'} \Delta_{klk'l'} \omega_{k''l''} \nabla_{\theta} p_{\ell_{k''l''}}(\theta) \omega_{k'''l''''} \nabla_{\theta} p_{\ell_{k'''l''''}}(\theta). \end{aligned} \quad (4.14)$$

In the third equality, I used Lemma 4.37, i.e.,  $\Delta_{klk'l'} = 0$  if  $k'l' \in \bar{\mathcal{S}}_{kl}$ .

Note if  $(1_{kl}, 1_{k'l'})$  is independent of  $(1_{k''l''}, 1_{k'''l'''})$ , then  $1_{kl}1_{k'l'}$  is independent of  $1_{k''l''}1_{k'''l'''}$  and  $\text{Cov}_\pi(1_{kl}1_{k'l'}, 1_{k''l''}1_{k'''l'''}) = 0$ . In particular, covariance is nonzero if

$$k''l'' \in \mathcal{S}_{kl}, k''l'' \in \mathcal{S}_{k'l'}, k'''l''' \in \mathcal{S}_{kl}, \text{ or } k'''l''' \in \mathcal{S}_{k'l'}. \quad (4.15)$$

So the upper bound for number of nonzero term in  $\text{Var}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta})]$  is  $T_c \lambda^8$ . To see this, observe there are  $T_c \lambda^2$  choices for  $kl$  and at most  $\lambda^2$  choices for  $k'l'$  given  $kl$ , at most  $4\lambda^2$  for 4.15 to be true, and at most  $\lambda^2$  choice for  $k'''l'''$  given  $k''l''$ . Therefore, from 4.14, one has

$$\text{Var}_\pi [T_c \hat{\mathbf{J}}_\pi(\boldsymbol{\theta})] = O(T_c^{-2}) O(T_c \lambda^2) O(\lambda^2) O(4\lambda^2) O(\lambda^2) = O(T_c^{-1}).$$

□

**Remark.** The argument above can be easily extend to the stratified Poisson sampling design. I omit the detail.

#### 4.7.2 Example: SRSWOR sampling design

Assume the sampling clusters are not the same as the model clusters, i.e.  $U_1^c \neq M_1^c$ . The argument for the SRSWOR two-stage sampling is more involved. Consider the following two-stage sample design:

- (i) First-stage: SRSWOR with the sample size  $n_1$  from the population size  $N_1$  for the sampling cluster.
- (ii) Second-stage: SRSWR with sample size  $n_i = C_1$  from population size  $N_i = C_2$  for each sampling cluster  $i$ .

I start by introducing some notations. Define

$$\mathcal{P} = \{kl \in U^2 : k < l \text{ and } m(k) = m(l)\}.$$

**Definition 4.39.** Let  $kl, k'l', k''l'' \in \mathcal{P}$ . Define

$$\begin{aligned}\mathcal{S}_{kl} &= \{k'l' \in \mathcal{P} : s(k') = s(k) \text{ or } s(k') = s(l) \text{ or } s(l') = s(k) \text{ or } s(l') = s(l)\}, \\ \mathcal{S}_{klk'l'} &= \{k''l'' \in \mathcal{P} : s(k'') = s(k) \text{ or } s(k'') = s(l) \text{ or } s(k'') = s(k') \text{ or } s(k'') = s(l') \text{ or} \\ &\quad s(l'') = s(k) \text{ or } s(l'') = s(l) \text{ or } s(l'') = s(k') \text{ or } s(l'') = s(l')\}, \\ \mathcal{S}_{klk'l'k''l''} &= \{k'''l''' \in \mathcal{P} : s(k''') = s(k) \text{ or } s(k''') = s(l) \text{ or } s(k''') = s(k') \text{ or } s(k''') = s(l') \text{ or} \\ &\quad s(k''') = s(k'') \text{ or } s(k''') = s(l'') \text{ or } s(l''') = s(k) \text{ or } s(l''') = s(l) \text{ or } s(l''') = s(k') \text{ or} \\ &\quad s(l''') = s(l') \text{ or } s(l''') = s(k'') \text{ or } s(l''') = s(l'')\}.\end{aligned}$$

Let  $\bar{\mathcal{S}}_{kl}$  be the complement of  $\mathcal{S}_{kl}$ ,  $\bar{\mathcal{S}}_{klk'l'}$  be the complement of  $\mathcal{S}_{klk'l'}$ , and  $\bar{\mathcal{S}}_{klk'l'k''l''}$  be the complement of  $\mathcal{S}_{klk'l'k''l''}$  in  $\mathcal{P}$ . More precisely,

$$\begin{aligned}\bar{\mathcal{S}}_{kl} &= \mathcal{P} \setminus \mathcal{S}_{kl}, \\ \bar{\mathcal{S}}_{klk'l'} &= \mathcal{P} \setminus \mathcal{S}_{klk'l'}, \\ \bar{\mathcal{S}}_{klk'l'k''l''} &= \mathcal{P} \setminus \mathcal{S}_{klk'l'k''l''}.\end{aligned}$$

**Remark.**  $\mathcal{S}_{kl}$  is the set of pairs that shares at least one sampling cluster of  $kl$ .  $\mathcal{S}_{klk'l'}$  is the set of pairs that shares at least one sampling cluster of  $klk'l'$ .  $\mathcal{S}_{klk'l'k''l''}$  is the set of pairs that shares at least one sampling cluster of  $klk'l'k''l''$ .

**Remark.** Let  $kl, k'l', k''l'' \in \mathcal{P}$ . Observe  $\sup_{kl} |\mathcal{S}_{kl}| \leq \lambda^2$ ,  $\sup_{klk'l'} |\mathcal{S}_{klk'l'}| \leq \lambda^2$ ,  $\sup_{klk'l'k''l''} |\mathcal{S}_{klk'l'k''l''}| \leq \lambda^2$ ,  $\sup_{kl} |\bar{\mathcal{S}}_{kl}| \leq T_c \lambda^2$ ,  $\sup_{klk'l'} |\bar{\mathcal{S}}_{klk'l'}| \leq T_c \lambda^2$  and  $\sup_{klk'l'k''l''} |\bar{\mathcal{S}}_{klk'l'k''l''}| \leq T_c \lambda^2$ .

**Lemma 4.40.** Under the above SRSWOR setting. If  $k'l' \in \bar{\mathcal{S}}_{kl}$ , then

$$\Delta_{klk'l'} = O(T_c^{-1}).$$

*Proof.* To see this, there are two cases to consider.

First, suppose  $s(k') = s(l')$ , then there are only two possibilities, either  $\langle kl \rangle \times \langle k'l' \rangle$  or  $\langle k \rangle \times \langle l \rangle \times \langle k'l' \rangle$ .

$$\begin{aligned}
\Delta_{klk'l'} &= \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I-1}{N_I-1} \frac{C_1}{C_2} \frac{C_1}{C_2} - \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \\
&= c \frac{n_I}{N_I} \left( \frac{n_I-1}{N_I-1} - \frac{n_I}{N_I} \right) \\
&= -c \frac{1}{N_I-1} \frac{n_I}{N_I} \left( 1 - \frac{n_I}{N_I} \right) \\
&= O(N_I^{-1}) \\
&= O(T_c^{-1}).
\end{aligned}$$

If  $\langle k \rangle l \rangle \langle k' l' \rangle$ , then

$$\begin{aligned}
\Delta_{klk'l'} &= \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{n_I-1}{N_I-1} \frac{C_1}{C_2} \frac{n_I-2}{N_I-2} \frac{C_1}{C_2} \frac{C_1}{C_2} - \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{n_I-1}{N_I-1} \frac{C_1}{C_2} \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \\
&= c \frac{n_I}{N_I} \frac{n_I-1}{N_I-1} \left( \frac{n_I-2}{N_I-2} - \frac{n_I}{N_I} \right) \\
&= -c \frac{1}{N_I-2} \frac{n_I}{N_I} \frac{n_I-1}{N_I-1} \left( 2 - \frac{2n_I}{N_I} \right) \\
&= O(N_I^{-1}) \\
&= O(T_c^{-1}).
\end{aligned}$$

Secondly, suppose  $s(k') \neq s(l')$ , then either  $\langle kl \rangle \langle k' \rangle \langle l' \rangle$  or  $\langle k \rangle l \rangle \langle k' \rangle \langle l' \rangle$ .

If  $\langle kl \rangle \langle k' \rangle \langle l' \rangle$ , then

$$\begin{aligned}
\Delta_{klk'l'} &= \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I-1}{N_I-1} \frac{C_1}{C_2} \frac{n_I-2}{N_I-2} \frac{C_1}{C_2} - \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{C_1}{C_2} \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{n_I-1}{N_I-1} \frac{C_1}{C_2} \\
&= c \frac{n_I}{N_I} \frac{n_I-1}{N_I-1} \left( \frac{n_I-2}{N_I-2} - \frac{n_I}{N_I} \right) \\
&= -c \frac{1}{N_I-2} \frac{n_I}{N_I} \frac{n_I-1}{N_I-1} \left( 2 - \frac{2n_I}{N_I} \right) \\
&= O(N_I^{-1}) \\
&= O(T_c^{-1}).
\end{aligned}$$

If  $\langle k \rangle \langle l \rangle \langle k' \rangle \langle l' \rangle$ , then

$$\begin{aligned}\Delta_{klk'l'} &= \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{n_I - 1}{N_I - 1} \frac{C_1}{C_2} \frac{n_I - 2}{N_I - 2} \frac{C_1}{C_2} \frac{n_I - 3}{N_I - 3} \frac{C_1}{C_2} - \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{n_I - 1}{N_I - 1} \frac{C_1}{C_2} \frac{n_I}{N_I} \frac{C_1}{C_2} \frac{n_I - 1}{N_I - 1} \frac{C_1}{C_2} \\ &= c \frac{n_I}{N_I} \frac{n_I - 1}{N_I - 1} \left( \frac{n_I - 2}{N_I - 2} \frac{n_I - 3}{N_I - 3} - \frac{n_I}{N_I} \frac{n_I - 1}{N_I - 1} \right) \\ &= O(N_I^{-1}) \\ &= O(T_c^{-1}).\end{aligned}$$

□

**Corollary 4.41.** Under the above SRSWOR setting, if  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}$ , then  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$ .

*Proof.* To see this, observe if  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}$ , then  $k'''l''' \in \bar{\mathcal{S}}_{k''l''}$ . By Lemma 4.40, one has  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$ . □

**Lemma 4.42.** Under the above SRSWOR setting, if  $k''l'' \in \bar{\mathcal{S}}_{klk'l'}$  and  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'}$ , then

$$\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1}).$$

*Proof.* Observe any element in  $k''l''k'''l'''$  has no common root with any one of  $klk'l'$ . The proof is not hard, but tedious to write down all the detail. I omit the detail. □

**Proposition 4.43.** Under the above SRSWOR setting, then for all  $\theta \in \Theta$

$$\text{Var}_{\pi} [T_c \hat{\mathbf{J}}_{\pi}(\theta)] \rightarrow 0.$$

*Proof.* Recall

$$\begin{aligned}\text{Var}_{\pi} [T_c \hat{\mathbf{J}}_{\pi}(\theta)] &= \text{Var}_{\pi} \left[ \frac{1}{T_c} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} (\omega_{kl} \nabla_{\theta} p \ell_{kl}(\theta) \omega_{k'l'} \nabla_{\theta} p \ell_{k'l'}(\theta)) \right] \\ &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{i''=1}^{T_c} \sum_{i'''=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} \sum_{\substack{k'' < l'' \\ k'', l'' \in M_{i''}^c}} \sum_{\substack{k''' < l''' \\ k''', l''' \in M_{i'''}^c}} \Delta_{klk'l'k''l''k'''l'''} \omega_{klk'l'} \Delta_{klk'l'} \omega_{k''l''k'''l'''} \\ &\quad \Delta_{k''l''k'''l'''} (\omega_{kl} \nabla_{\theta} p \ell_{kl}(\theta) \omega_{k'l'} \nabla_{\theta} p \ell_{k'l'}(\theta)) (\omega_{k''l''} \nabla_{\theta} p \ell_{k''l''}(\theta) \omega_{k'''l'''} \nabla_{\theta} p \ell_{k'''l'''}(\theta)).\end{aligned}$$

One can write the sum in the following

$$\begin{aligned}
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \left( \sum_{k'l' \in \mathcal{S}_{kl}} + \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \right) \left( \sum_{k''l'' \in \mathcal{S}_{klk'l'}} + \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \right) \left( \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} + \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} \right) \\
&= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} + \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} + \\
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} + \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} + \\
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} + \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} + \\
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} + \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}}. \tag{4.16}
\end{aligned}$$

For the **first** term in 4.16, one can show

$$\begin{aligned}
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} \\
&= O(T_c^{-2}) O(T_c) O(\lambda^2) O(\lambda^2) O(\lambda^2) O(\lambda^2) \\
&= O(T_c^{-1}).
\end{aligned}$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $\lambda^2$  choices for  $k'l'$  given  $kl$ , at most  $\lambda^2$  choices for  $k''l''$  given  $klk'l'$  and at most  $\lambda^2$  choices for  $k'''l'''$  given  $klk'l'k''l''$ .



For the **second** term in 4.16, one can show

$$\begin{aligned}
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} \\
&= O(T_c^{-2}) O(T_c) O(\lambda^2) O(\lambda^2) O(T_c \lambda^2) O(\lambda^2) O(T_c^{-1}) \\
&= O(T_c^{-1}).
\end{aligned} \tag{4.17}$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $\lambda^2$  choices for  $k'l'$  given  $kl$ , at most  $T_c \lambda^2$  choices for  $k''l''$  and at most  $\lambda^2$  choices for  $k'''l'''$  given  $klk'l'k''l''$ . For the last term  $O(T_c^{-1})$  in 4.17, then either **(A)** at least one of  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  or  $\Delta_{klk'l'k''l''} = O(T_c^{-1})$  or **(B)** one may prove directly  $\frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} = O(T_c^{-1})$ .

To see this, there are two possibilities for  $k'''l'''$ , only one of  $k'''l'''$  has a common root with at least one of  $klk'l'k''l''$  or both of  $k'''l'''$  have common root with at least one of  $klk'l'k''l''$ .

First, suppose one and only one of  $k'''l'''$  (say  $k'''$ ) has a common root with at least one of  $klk'l'k''l''$ . In particular, then the other (i.e.,  $l'''$ ) has no common root with any one of  $klk'l'k''l''$ . Then either  $k'''$  has a common root with at least one of  $klk'l'$  or  $k'''$  has a common root with at least one of  $k''l''$ . Observe those events are mutually disjoint.

If  $k'''$  has a common root with at least one of  $klk'l'$ , then  $k'''$  has no common root with  $k''l''$ . Hence  $k'''l''' \in \bar{\mathcal{S}}_{k''l''}$ . Therefore  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.40.

If  $k'''$  has a common root with at least one of  $k''l''$ , then  $k'''$  has no common root with any one of  $klk'l'$ . Hence  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'}$ . Observe  $k''l'' \in \bar{\mathcal{S}}_{klk'l'}$  by construction. Therefore  $\Delta_{klk'l'k''l''} = O(T_c^{-1})$  from Lemma 4.42.

Secondly, suppose both of  $k'''l'''$  have common root with at least one of  $klk'l'k''l''$ , then either both of  $k'''l'''$  have common root with at least one of  $klk'l'$ , or both of  $k'''l'''$  have common root with at least one of  $k''l''$ , or one of  $k'''l'''$  (say  $k'''$ ) have common root with at least one of  $klk'l'$  and the other (i.e.,  $l'''$ ) have common root with at least one of  $k''l''$ .

If both of  $k'''l'''$  have common root with at least one of  $klk'l'$ , then  $k'''l''' \in \bar{\mathcal{S}}_{k''l''}$ . Hence one has  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.40.

If both of  $k'''l'''$  have common root with at least one of  $k''l''$ , then both of  $k'''l'''$  have no common root with any one of  $klk'l'$ . Hence  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'}$ . Observe  $k''l'' \in \bar{\mathcal{S}}_{klk'l'}$  by construction. Therefore  $\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1})$  from Lemma 4.42.

If  $k'''$  has common root with at least one of  $klk'l'$  and  $l'''$  has common root with at least one of  $k''l''$ , then one can show

$$\begin{aligned} & \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \mathcal{S}_{klk'l'k''l''}} \\ &= O(T_c^{-2}) O(T_c) O(\lambda^2) O(\lambda^2) O(\lambda) O(\lambda) O(\lambda^2) \\ &= O(T_c^{-1}). \end{aligned}$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$ , at most  $\lambda^2$  choices for  $k'l'$  given  $kl$ , at most  $\lambda$  choice for  $k'''$  given  $klk'l'$ , at most  $\lambda$  choice for  $l'''$  given  $k'''$ , at most  $\lambda^2$  choice for  $k'''l'''$  given  $l'''$ .

For the **third** term in 4.16, one can show

$$\begin{aligned} & \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} \\ &= O(T_c^{-2}) O(T_c) O(\lambda^2) O(T_c \lambda^2) O(\lambda^2) O(\lambda^2) O(T_c^{-1}) \\ &= O(T_c^{-1}). \end{aligned} \tag{4.18}$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $T_c \lambda^2$  choices for  $k'l'$ , at most  $\lambda^2$  choices for  $k''l''$  given  $klk'l'$  and at most  $\lambda^2$  choices for  $k'''l'''$  given  $klk'l'k''l''$ . For the last term  $O(T_c^{-1})$  in 4.18, observe  $k'l' \in \bar{\mathcal{S}}_{kl}$ , hence  $\Delta_{klk'l'} = O(T_c^{-1})$  by Lemma 4.40.

For the **fourth** term in 4.16, one can show

$$\begin{aligned}
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{S}_{kl}} \sum_{k''l'' \in \bar{S}_{klk'l'}} \sum_{k'''l''' \in \bar{S}_{klk'l'k''l''}} \\
&= O(T_c^{-2}) O(T_c) O(\lambda^2) O(T_c \lambda^2) O(T_c \lambda^2) O(\lambda^2) O(T_c^{-1}) O(T_c^{-1}) \quad (4.19) \\
&= O(T_c^{-1}).
\end{aligned}$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $T_c \lambda^2$  choices for  $k'l'$ , at most  $T_c \lambda^2$  choices for  $k''l''$  and at most  $\lambda^2$  choices for  $k'''l'''$  given  $klk'l'k''l''$ . For the first  $O(T_c^{-1})$  in 4.19, observe  $k'l' \in \bar{S}_{kl}$ , hence  $\Delta_{klk'l'} = O(T_c^{-1})$  by Lemma 4.40. For the last term  $O(T_c^{-1})$  in 4.19. I use the following fact at least one of  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  or  $\Delta_{klk'l'k''l''} = O(T_c^{-1})$ . If this does not hold, then one can prove directly  $\frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{S}_{kl}} \sum_{k''l'' \in \bar{S}_{klk'l'}} \sum_{k'''l''' \in \bar{S}_{klk'l'k''l''}} = O(T_c^{-1})$

To see this, there are two possibilities for  $k'''l'''$ , only one of  $k'''l'''$  has a common root with at least one of  $klk'l'k''l''$  or both of  $k'''l'''$  have common root with at least one of  $klk'l'k''l''$ .

First, suppose one and only one of  $k'''l'''$  (say  $k'''$ ) has a common root with at least one of  $klk'l'k''l''$ . In particular, the other (i.e.,  $l'''$ ) has no common root with any one of  $klk'l'k''l''$ . Then either  $k'''$  has a common root with at least one of  $klk'l'$  or  $k'''$  has a common root with at least one of  $k''l''$ . Observe those events are mutually disjoint.

If  $k'''$  has a common root with at least one of  $klk'l'$ , then  $k'''$  has no common root with  $k''l''$ . Hence  $k'''l''' \in \bar{S}_{k''l''}$ . Therefore  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.40.

If  $k'''$  has a common root with at least one of  $k''l''$ , then  $k'''$  has no common root with any one of  $klk'l'$ . Hence  $k'''l''' \in \bar{S}_{klk'l'}$ . Observe  $k''l'' \in \bar{S}_{klk'l'}$  by construction. Therefore  $\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1})$  from Lemma 4.42.

Secondly, suppose both of  $k'''l'''$  have common root with at least one of  $klk'l'k''l''$ , then either both of  $k'''l'''$  have common root with at least one of  $klk'l'$ , or both of  $k'''l'''$  have common root with at least one of  $k''l''$ , or one of  $k'''l'''$  (say  $k'''$ ) have common root with at least one of  $klk'l'$  and the other root (say  $l'''$ ) have common root with at least one of  $k''l''$ .

If both of  $k'''l'''$  have common root with at least one of  $klk'l'$ , then  $k'''l''' \in \bar{\mathcal{S}}_{k''l''}$ . Hence one has  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Lemma 4.40.

If both of  $k'''l'''$  have common root with at least one of  $k''l''$ , then both  $k'''l'''$  have no common root with any one of  $klk'l'$ . Hence, any element in  $k''l''k'''l'''$  has no common root with any one of  $klk'l'$ . Therefore  $\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1})$  from Lemma 4.42.

If  $k'''$  has common root with at least one of  $klk'l'$  and  $l'''$  has common root with at least one of  $k''l''$ , then

$$\begin{aligned} & \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} \\ &= O(T_c^{-2}) O(T_c) O(\lambda^2) O(T_c \lambda^2) O(\lambda) O(\lambda) O(\lambda^2) O(T_c^{-1}) \\ &= O(T_c^{-1}). \end{aligned} \quad (4.20)$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $T_c \lambda^2$  choices for  $k'l'$ , at most  $\lambda$  choices for  $k'''$  given  $klk'l'$ , at most  $\lambda$  choices for  $l'''$  given  $k'''$ , at most  $\lambda^2$  choices for  $k''l''$  given  $l'''$ . For the last  $O(T_c^{-1})$  in 4.20, observe  $k'l' \in \bar{\mathcal{S}}_{kl}$ , hence  $\Delta_{klk'l'} = O(T_c^{-1})$  by Lemma 4.40.

For the **fifth** term in 4.16, one can show

$$\begin{aligned} & \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} \\ &= O(T_c^{-2}) O(T_c) O(\lambda^2) O(\lambda^2) O(\lambda^2) O(T_c \lambda^2) O(T_c^{-1}) \\ &= O(T_c^{-1}). \end{aligned} \quad (4.21)$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $\lambda^2$  choices for  $k'l'$  given  $kl$ , at most  $\lambda^2$  choices for  $k''l''$  given  $klk'l'$  and at most  $T_c \lambda^2$  choices for  $k'''l'''$ . For the last term  $O(T_c^{-1})$  in 4.21, observe  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}$ . Therefore  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Corollary 4.41.

For the **sixth** term in in 4.16, one can show

$$\begin{aligned}
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \mathcal{S}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} \\
&= O(T_c^{-2}) O(T_c) O(\lambda^2) O(\lambda^2) O(T_c \lambda^2) O(T_c \lambda^2) O(T_c^{-1}) O(T_c^{-1}) \quad (4.22) \\
&= O(T_c^{-1}).
\end{aligned}$$

To see this, note there are  $T_c$  choices for  $i$ , at most  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $\lambda^2$  choices for  $k'l'$  given  $kl$ , at most  $T_c \lambda^2$  choices for  $k''l''$  and at most  $T_c \lambda^2$  choices for  $k'''l'''$ . For the first  $O(T_c^{-1})$  in 4.22, observe  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}$ , hence  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Corollary 4.41. For the last term  $O(T_c^{-1})$  in 4.22, observe  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'}$  and  $k'l' \in \bar{\mathcal{S}}_{klk'l'}$  by construction. Therefore  $\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1})$  from Lemma 4.42.

For the **seventh** term in in 4.16, one has

$$\begin{aligned}
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \mathcal{S}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} \\
&= O(T_c^{-2}) O(T_c) O(\lambda^2) O(T_c \lambda^2) O(\lambda^2) O(T_c \lambda^2) O(T_c^{-1}) O(T_c^{-1}) \quad (4.23) \\
&= O(T_c^{-1}).
\end{aligned}$$

To see this, note there are  $T_c$  choices for  $i$ ,  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $T_c \lambda^2$  choices for  $k'l'$ , at most  $\lambda^2$  choices for  $k''l''$  given  $klk'l'$  and at most  $T_c \lambda^2$  choices for  $k'''l'''$ . For the last two term  $O(T_c^{-1}) O(T_c^{-1})$  in 4.23, observe  $k'l' \in \bar{\mathcal{S}}_{kl}$  and  $k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}$ , hence  $\Delta_{klk'l'} = O(T_c^{-1})$  by Lemma 4.40 and  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Corollary 4.41.

For the **eighth** sum in in 4.16, one has

$$\begin{aligned}
& \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{k'l' \in \bar{\mathcal{S}}_{kl}} \sum_{k''l'' \in \bar{\mathcal{S}}_{klk'l'}} \sum_{k'''l''' \in \bar{\mathcal{S}}_{klk'l'k''l''}} \\
&= O(T_c^{-2}) O(T_c) O(\lambda^2) O(T_c \lambda^2) O(T_c \lambda^2) O(T_c \lambda^2) O(T_c^{-1}) O(T_c^{-1}) O(T_c^{-1}) \quad (4.24) \\
&= O(T_c^{-1}).
\end{aligned}$$

To see this, note there are  $T_c$  choices for  $i$ ,  $\lambda^2$  choices for  $kl$  given  $i$ , at most  $T_c\lambda^2$  choices for  $k'l'$ , at most  $T_c\lambda^2$  choices for  $k''l''$  and at most  $T_c\lambda^2$  choices for  $k'''l'''$ . For the first two  $O(T_c^{-1})O(T_c^{-1})$  in 4.24, observe  $k'l' \in \bar{S}_{kl}$  and  $k'''l''' \in \bar{S}_{klk'l'k''l''}$ , hence  $\Delta_{klk'l'} = O(T_c^{-1})$  by Lemma 4.40 and  $\Delta_{k''l''k'''l'''} = O(T_c^{-1})$  by Corollary 4.41. For the last term  $O(T_c^{-1})$  in 4.24, observe  $k'''l''' \in \bar{S}_{klk'l'}$  and  $k'l' \in \bar{S}_{klk'l'}$  by construction. Therefore  $\Delta_{klk'l'k''l''k'''l'''} = O(T_c^{-1})$  from Lemma 4.42.

Putting all those estimates together, one has

$$\text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\boldsymbol{\theta}) \right] = O(T_c^{-1}) \rightarrow 0.$$

This completes the proof.  $\square$

**Remark.** Note the tree structure of the  $\Delta_{klk'l'}$ ,  $\Delta_{k''l''k'''l'''}$  and  $\Delta_{klk'l'k''l''k'''l'''}$  is vital to establish the convergence.

## 4.8 Model-based variance estimation

Instead of estimating  $\mathbf{J}_1(\boldsymbol{\theta}_0)$  by empirical variance

$$\hat{\mathbf{J}}_{\pi}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} \left( \omega_{kl} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl}}(\hat{\boldsymbol{\theta}}_n) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p_{\ell_{k'l'}}(\hat{\boldsymbol{\theta}}_n) \right).$$

I could estimate  $\mathbf{J}_1(\boldsymbol{\theta}_0)$  by model variance

$$\hat{\mathbf{J}}_{\pi}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} \omega_{kl} \omega_{k'l'} \mathbb{E}_Y \left[ \nabla_{\boldsymbol{\theta}} p_{\ell_{kl}}(\hat{\boldsymbol{\theta}}_n) \nabla_{\boldsymbol{\theta}} p_{\ell_{k'l'}}(\hat{\boldsymbol{\theta}}_n) \right].$$

Using the same argument as in the previous section, one can show

$$\text{Var}_{\pi} \left[ T_c \hat{\mathbf{J}}_{\pi}(\boldsymbol{\theta}) \right] = 8O(T_c^{-1}) \rightarrow 0.$$

## 4.9 Simulation: Random intercept model

A simulation study for a random intercept model was conducted to examine performance of the weighted pairwise likelihood estimation and consistency of variance estimation under various uninformative and informative sampling design. Based on 150 simulation replicates, median bias, median absolute deviation (mad) and median estimated standard deviation (esd) are computed to measure the performance of the:

- (i) naive maximum likelihood estimation (NMLE).
- (ii) pairwise likelihood estimation (PLE).
- (iii) weighted pairwise likelihood estimation (WPLE).

The NMLE can be implemented by using R lme4 package (Bates, Maechler, Bolker, Walker et al., 2014). There is no R package for PLE and WPLE. I wrote the R codes for PLE and WPLE. Sampling design can be implemented by using R sampling package (Tillé and Matei, 2009). The code for these simulations is publicly available at <https://github.com/Xudong3/>.

### 4.9.1 Model

A random intercept model is given by

$$\begin{aligned} Y_{ik}|b_i &\sim N(\beta_0 + \beta_1 X_{ik} + b_i, \sigma^2), \\ b_i &\sim N(0, \tau^2), \end{aligned}$$

for  $i = 1, \dots, T_c$  and  $k = 1, \dots, T_i$ . The parameters I want to estimate are  $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2, \tau^2)^T$ .

The sample weighted pairwise likelihood is given by

$$p\ell_{M^s}(\boldsymbol{\theta}) = \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta}),$$

where

$$p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta}) = -\frac{1}{2} \log d - \frac{1}{2} \frac{1}{d} \left[ r_l^2(\sigma^2 + \tau^2) - 2r_l r_k \tau^2 + r_k^2(\sigma^2 + \tau^2) \right],$$


$$d = (\sigma^2 + \tau^2)^2 - \tau^4,$$

$$r_k = y_k - \beta_0 - \beta_1 x_k.$$

The pairwise score function is given by

$$\nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \beta_0} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \beta_1} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \sigma^2} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \tau^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{d} [r_k(\sigma^2 + \tau^2) - r_k \tau^2 - r_l \tau^2 + r_l(\sigma^2 + \tau^2)] \\ \frac{1}{d} [r_k(\sigma^2 + \tau^2)x_k - r_k \tau^2 x_l - r_l \tau^2 x_k + r_l(\sigma^2 + \tau^2)x_l] \\ -\frac{1}{2} \frac{r_k^2 + r_l^2}{d} + \frac{[r_k^2(\sigma^2 + \tau^2) - 2r_k r_l \tau^2 + r_l^2(\sigma^2 + \tau^2)](\sigma^2 + \tau^2)}{d^2} - \frac{\sigma^2 + \tau^2}{d} \\ -\frac{1}{2} \frac{r_k^2 - 2r_k r_l \tau^2 + r_l^2}{d} - \frac{[r_k^2(\sigma^2 + \tau^2) - 2r_k r_l \tau^2 + r_l^2(\sigma^2 + \tau^2)]\sigma^2}{d^2} + \frac{\sigma^2}{d} \end{pmatrix}.$$

I first generated a target population of size  $100 \times 100$ , which consists of 100 sample clusters of size 100. Then I generated 100 model clusters. The parameter 'overlap' is the percentage of observational units in each sampling cluster such that the sample clusters match the model clusters. I assume the 'overlap' is the same across all the sampling clusters. I set the parameter 'overlap' = 4/5 in all of the setting. See figure below. In the figure below, I work on 5 clusters of size 5. The extension to 100 clusters of size 100

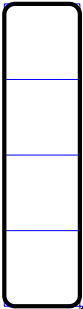
can be done similarly. Each small rectangle  represents an observational unit and the whole rectangle represents the population of size 25. It consists of 5 clusters of size 5



5, i.e.,  $N_I = T_c = 5$  and  $N_i = T_i = 5$  for all  $i$ . Each column



cluster of size 5, i.e.,  $U_I^c = \{U_1^c, \dots, U_5^c\}$ . Each



(right) represent a model cluster of size 5, i.e.,  $M_I^c = \{M_1^c, \dots, M_5^c\}$ .

$U_1^c$	$U_2^c$	$U_3^c$	$U_4^c$	$U_5^c$
$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$	$M_5^c$
$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$	$M_5^c$
$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$	$M_5^c$
$M_5^c$	$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$
$M_4^c$	$M_5^c$	$M_1^c$	$M_2^c$	$M_3^c$

$U_1^c$	$U_2^c$	$U_3^c$	$U_4^c$	$U_5^c$
$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$	$M_5^c$
$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$	$M_5^c$
$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$	$M_5^c$
$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$	$M_5^c$
$M_5^c$	$M_1^c$	$M_2^c$	$M_3^c$	$M_4^c$

Figure 4.2: Overlap between the sampling clusters and the model clusters when the overlap parameter is  $\frac{3}{5}$  (left) and  $\frac{4}{5}$  (right).

#### 4.9.2 Basic notation

For the simulations below, I want to introduce some notation. Let  $q$  be the number of simulations replicate.

**Definition 4.44.** Let  $\theta_0$  be the true value, then the median bias of  $\hat{\theta}$  is defined to be

$$\text{median bias}(\hat{\theta}) = \text{median}\{\hat{\theta}_r - \theta_0 : r = 1, \dots, q\}.$$

**Definition 4.45.** The median absolute deviation (mad) of  $\hat{\theta}$  is defined to be

$$\text{mad}(\hat{\theta}) = \frac{1}{\phi^{-1}(\frac{3}{4})} \times \text{median}\{|\hat{\theta}_r - \text{median}\{\hat{\theta}_r : r = 1, \dots, q\}| : r = 1, \dots, q\},$$

where  $\phi$  is the standard normal density function.

**Definition 4.46.** The sample pairwise likelihood estimator  $\tilde{\theta}$  of  $\theta$  is defined as a solution of

$$\frac{1}{T_I} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \nabla_{\theta} p_{\ell_{kl|i}}(y_k, y_l, \theta) = \mathbf{0}.$$

**Definition 4.47.** The median estimated standard deviation (esd) of the sample pairwise likelihood estimator  $\tilde{\theta}$  is defined to be

$$\text{esd}(\tilde{\theta}) = \text{median}\{(\text{diag}\{\tilde{\mathbf{H}}_r^{-1}(\tilde{\theta}_r) \tilde{\mathbf{J}}_r(\tilde{\theta}_r) \tilde{\mathbf{H}}_r^{-1}(\tilde{\theta}_r)\})^{\frac{1}{2}} : r = 1, \dots, q\},$$

where

$$\begin{aligned} \tilde{\mathbf{H}}_r(\tilde{\theta}_r) &= -\frac{1}{T_I} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} \nabla_{\theta}^2 p_{\ell_{kl}}(\tilde{\theta}_r), \\ \tilde{\mathbf{J}}_r(\tilde{\theta}_r) &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} \left( \nabla_{\theta} p_{\ell_{kl}}(\tilde{\theta}_r) \nabla_{\theta} p_{\ell_{k'l'}}(\tilde{\theta}_r)^T \right). \end{aligned}$$

**Definition 4.48.** The median estimated standard deviation (esd) of the sample pairwise score  $PS(\tilde{\theta})$  is defined to be

$$\text{esd}(PS(\tilde{\theta})) = \text{median}\{(\text{diag}\{\tilde{\mathbf{J}}_r(\tilde{\theta}_r)\})^{\frac{1}{2}} : r = 1, \dots, q\}.$$

**Definition 4.49.** The sample weighted pairwise likelihood estimator  $\hat{\theta}$  of  $\theta$  is defined as a solution of

$$\frac{1}{T_I} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} \nabla_{\theta} p_{\ell_{kl|i}}(y_k, y_l, \theta) = \mathbf{0}.$$

**Definition 4.50.** The median estimated standard deviation (esd) of the sample weighted pairwise likelihood estimator  $\hat{\theta}$  is defined to be

$$\text{esd}(\hat{\theta}) = \text{median}\{(\text{diag}\{\hat{\mathbf{H}}_r^{-1}(\hat{\theta}_r) \hat{\mathbf{J}}_r(\hat{\theta}_r) \hat{\mathbf{H}}_r^{-1}(\hat{\theta}_r)\})^{\frac{1}{2}} : r = 1, \dots, q\},$$

where

$$\begin{aligned}\hat{\mathbf{H}}_r(\hat{\boldsymbol{\theta}}_r) &= -\frac{1}{T_1} \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} \nabla_{\boldsymbol{\theta}}^2 p_{\ell_{kl}}(\hat{\boldsymbol{\theta}}_r), \\ \hat{\mathbf{J}}_r(\hat{\boldsymbol{\theta}}_r) &= \frac{1}{T_c^2} \sum_{i=1}^{T_c} \sum_{i'=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} \sum_{\substack{k' < l' \\ k', l' \in M_{i'}^c}} 1_{klk'l'} \omega_{klk'l'} \Delta_{klk'l'} \left( \omega_{kl} \nabla_{\boldsymbol{\theta}} p_{\ell_{kl}}(\hat{\boldsymbol{\theta}}_r) \omega_{k'l'} \nabla_{\boldsymbol{\theta}} p_{\ell_{k'l'}}(\hat{\boldsymbol{\theta}}_r)^T \right).\end{aligned}$$

**Definition 4.51.** The median estimated standard deviation (esd) of the sample weighted pairwise score  $WPS(\hat{\boldsymbol{\theta}})$  is defined to be

$$\text{esd}(WPS(\hat{\boldsymbol{\theta}})) = \text{median}\{(\text{diag}\{\hat{\mathbf{J}}_r(\hat{\boldsymbol{\theta}}_r)\})^{\frac{1}{2}} : r = 1, \dots, q\}.$$

**Remark.** For the implementation of PLE and WPL, boundary constraint (i.e.,  $\sigma^2 > 0$  and  $\tau^2 > 0$ ) and initial starting value have to be given for the optimisation algorithm. It was found that solution might not converge and is very sensitive to initial starting value using "BFGS" and "L-BFGS-B" methods from optim function in R. One can use bobyqa function from R minqa package (Bates, Mullen, Nash and Varadhan, 2014) to overcome those issues. This is due to Thomas Lumley.

**Remark.** The simulations are being done to assess convergence in distribution of  $\hat{\boldsymbol{\theta}}_n$  and in probability of  $\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_n)$ ,  $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}_n)$ . Observe convergence in probability and distribution implies convergence of median and median absolute deviation, but not the mean and standard deviation, which is why I am using median and median absolute deviation.

**Remark.** Ideally, one wants to test the performance of the weighted pairwise likelihood estimation by varying model parameters, design parameters, overlap parameter and sample size. In particular, consistency of the empirical variance estimator requires a large sample size. However,  $\mathbf{J}(\boldsymbol{\theta})$  is computational demanding. Despite considerable effort has been made to optimize the codes such as writing the sampling inclusion probability in C, computation of  $\mathbf{J}(\boldsymbol{\theta})$  is still slow and cannot be completely vectorise due to memory constraint. Computational load restricts the number of replicates and the sample size.

**Remark.** I want to mention why the full-likelihood is intractable when the sampling clusters are not equal to the model clusters.

Consider a random intercept model with informative Poisson sampling at stage 1 and simple random sampling at stage 2. Assume the sampling clusters are equal to model clusters, the sampling is independent across the clusters. Let  $g(b_i, \theta)$  be the density of  $b_i$  and  $f(y_{ik}, \theta|b_i)$  be the conditional density of  $y_{ik}$ . The population likelihood is

$$\prod_{i=1}^{N_I} \int g(b_i, \theta) \prod_{k \in U_i^c} f(y_{ik}, \theta|b_i) db_i.$$

Let  $g_R$  be the sampling likelihood at stage 1. Then the likelihood of the data is proportional to

$$\prod_{i=1}^{N_I} \int g_R(b_i, \theta) \prod_{k \in U_i^s} f(y_{ik}, \theta|b_i) db_i.$$

We have a product of one-dimensional numerical integrals for the full likelihood to maximise, which is completely feasible. The sample likelihood is also tractable.

When the sampling clusters are not the same as the model clusters, the sampling likelihood  $g_R$  for a particular sample cluster depends on the  $b_i$  and  $Y_{ij}$  for all model clusters that intersect it, so the product over  $i$  cannot simply be taken outside the integral. In the simulation setting in the thesis ( $100 \times 100$ ) with overlap of 0.6, 40  $b_i$  contribute to each sampling probability so even under Poisson sampling we have 40-dimensional numerical integrals to compute the likelihood. This is just intractable, which is why we want to use pseudo-likelihood.

#### 4.9.3 Design: stratified sampling

Consider an **uninformative** stratified SRSWOR, i.e., SRSWOR with the sample size  $n_i = 10$  for each sampling cluster  $i$ .

Under an **uninformative** sampling design, the main simulation results are the following:

- (i) NMLE, PLE and WPLE have similar levels of bias.
- (ii) The mad is closely estimated by the esd for both PLE and WPLE. There is little loss of efficiency from using PLE.

(iii) It is not clear which estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for uninformative sampling is 0.04.

Table 1: Performance of NMLE, PLE and WPLE under an **uninformative** stratified sampling design when the true value  $\beta_0 = 3, \beta_1 = 1, \sigma^2 = 1, \tau^2 = 0.8$ .

Uninformative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.10	0.03	0.10	0.04	0.05	0.10	0.04	0.04
$\beta_1$	-0.02	0.03	0.00	0.03	0.03	0.00	0.03	0.03
$\sigma^2$	0.00	0.05	0.00	0.05	0.05	0.00	0.05	0.05
$\tau^2$	0.00	0.05	-0.02	0.06	0.07	-0.02	0.06	0.07

Table 2: Pairwise score and weighted pairwise score under an **uninformative** stratified sampling design when the true value  $\beta_0 = 3, \beta_1 = 1, \sigma^2 = 1, \tau^2 = 0.8$ .

Uninformative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	3.54	1.48	1.64	376.43	157.63	171.48
$\beta_1$	0.45	3.11	2.90	36.88	338.97	305.64
$\sigma^2$	0.20	1.42	1.31	23.83	147.27	138.51
$\tau^2$	-0.10	1.05	1.07	-13.09	107.88	112.92

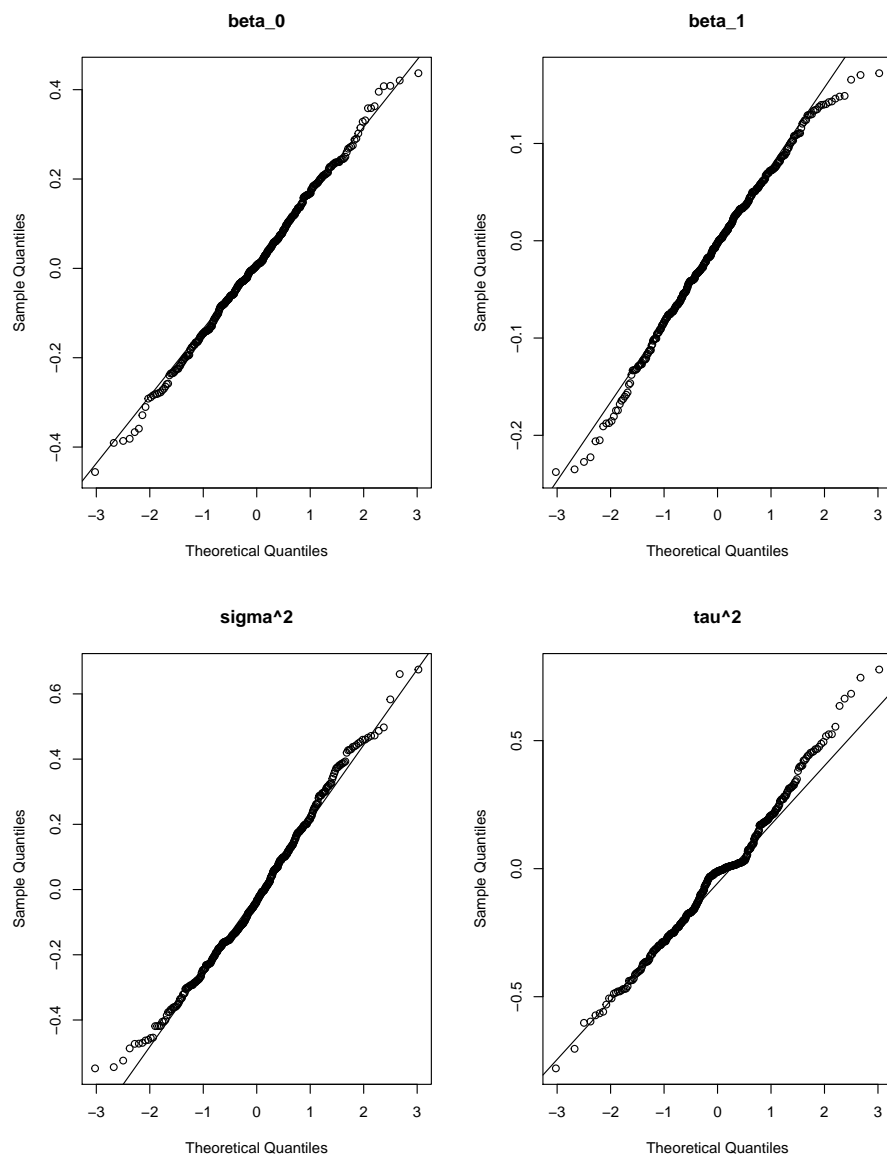


Figure 4.3: QQ plot for Weighted pairwise likelihood estimator under an **uninformative** stratified sampling design when the true value  $\beta_0 = 3$ ,  $\beta_1 = 1$ ,  $\sigma^2 = 1$ ,  $\tau^2 = 0.8$ .



Consider an **informative** stratified SRSWOR, i.e., SRSWOR with the sample size

$$n_i = \left\lceil \frac{a \exp(-br_i)}{1 + a \exp(-br_i)} N_i \right\rceil,$$

where

$$r_i = \sum_{k \in U: s(k)=i} (y_k - \beta_0 - \beta_1 x_k),$$

$$a = 0.15,$$

$$b = 0.45.$$

Under an **informative** sampling design, the main simulation results are the following:

- (i) NMLE and WPLE have similar levels of bias, but PLE has large bias.
- (ii) The mad is closely estimated by the esd for both PLE and WPLE.

The coefficients of variation of the weight  $\omega_{kl}$  for informative sampling is 0.61.

Table 3: Performance of NMLE, PLE and WPLE under an **informative** stratified sampling design when the true value  $\beta_0 = 3, \beta_1 = 1, \sigma^2 = 1, \tau^2 = 0.8$ .

Informative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.09	0.03	-0.06	0.04	0.04	0.10	0.04	0.04
$\beta_1$	-0.03	0.02	-0.21	0.02	0.02	-0.01	0.02	0.02
$\sigma^2$	0.00	0.04	0.03	0.04	0.04	-0.01	0.04	0.04
$\tau^2$	0.01	0.04	0.24	0.07	0.08	-0.02	0.06	0.06

Table 4: Pairwise score and weighted pairwise score under an **informative** stratified sampling design when the true value  $\beta_0 = 3, \beta_1 = 1, \sigma^2 = 1, \tau^2 = 0.8$ .

Informative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	-10.54	3.53	2.70	376.49	150.07	147.95
$\beta_1$	-61.41	5.10	5.16	-68.72	263.20	268.63
$\sigma^2$	7.31	2.16	2.21	-15.78	122.08	119.13
$\tau^2$	14.34	2.94	1.69	3.27	84.97	94.00

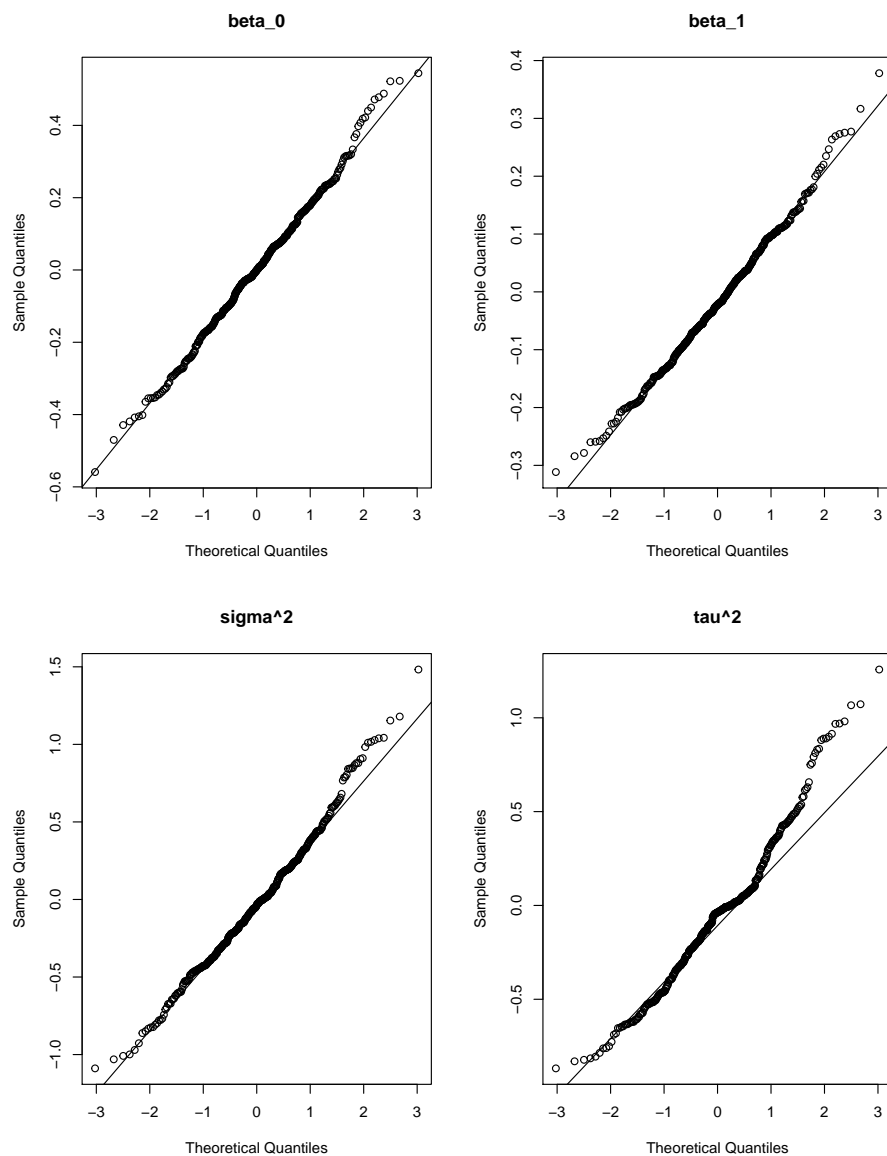


Figure 4.4: QQ plot for weighted pairwise likelihood estimator under an **uninformative** stratified sampling design when the true value  $\beta_0 = 3$ ,  $\beta_1 = 1$ ,  $\sigma^2 = 1$ ,  $\tau^2 = 0.8$ . .

#### 4.9.4 Design: two-stage SRSWOR

Consider an **uninformative** two-stage sample design,

- (i) First stage: SRSWOR with the sample size  $n_I = 10$ .
- (ii) Second stage: SRSWOR with the sample size  $n_i = 10$  for each sampling cluster  $i$ .

Under an **uninformative** sampling design, the main simulation results are the following:

- (i) PLE and WPLE have similar levels of bias, but NMLE has small bias.
- (ii) The esd underestimates the mad for both PLE and WPLE.
- (iii) It is clear NMLE estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for uninformative sampling is 0.87.

Table 5: Performance of NMLE, PLE and WPLE under an **uninformative** two-stage SRSWOR sampling design when the true value  $\beta_0 = 2, \beta_1 = 4, \sigma^2 = 0.5, \tau^2 = 1$ .

Uninformative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	-0.14	0.18	-0.12	0.29	0.29	-0.09	0.32	0.27
$\beta_1$	-0.04	0.10	-0.06	0.14	0.11	-0.07	0.15	0.11
$\sigma^2$	0.01	0.14	0.03	0.15	0.14	-0.01	0.18	0.14
$\tau^2$	-0.12	0.39	-0.32	0.39	0.28	-0.41	0.37	0.27

Table 6: Pairwise score and weighted pairwise score under an **uninformative** two-stage SRSWOR sampling design when the true value  $\beta_0 = 2, \beta_1 = 4, \sigma^2 = 0.5, \tau^2 = 1$ .

Uninformative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	-0.24	0.58	0.70	-280.23	1102.93	1277.43
$\beta_1$	-0.32	0.80	0.81	-587.28	1942.44	1543.60
$\sigma^2$	0.04	0.36	0.33	-69.17	692.46	634.71
$\tau^2$	-0.08	0.28	0.33	-155.78	561.73	619.67

Consider an **informative** two-stage sample design,

- (i) First stage: SRSWOR with the sample size  $n_I = 10$ .
- (ii) Second stage: SRSWOR with the sample size

$$n_i = \left\lceil \frac{a \exp(-br_i)}{1 + a \exp(-br_i)} N_i \right\rceil,$$

where

$$\begin{aligned} r_i &= \sum_{k \in U: s(k)=i} (y_k - \beta_0 - \beta_1 x_k), \\ a &= 0.05, \\ b &= 0.45. \end{aligned}$$

Under an **informative** sampling design, the main simulation results are the following:

- (i) PLE and WPLE have similar levels of bias, but NMLE has small bias.
- (ii) The esd underestimates the mad for both PLE and WPL.
- (iii) It is clear NMLE estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for informative sampling is 1.04.

Table 7: Performance of NMLE, PLE and WPLE under an **informative** two-stage SR-SWOR sampling design when the true value  $\beta_0 = 2, \beta_1 = 4, \sigma^2 = 0.5, \tau^2 = 1$ .

Informative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	-0.12	0.11	-0.24	0.26	0.26	-0.13	0.30	0.25
$\beta_1$	-0.03	0.05	-0.14	0.13	0.08	-0.06	0.11	0.09
$\sigma^2$	0.03	0.07	0.05	0.08	0.08	0.00	0.11	0.09
$\tau^2$	-0.09	0.24	-0.39	0.34	0.22	-0.34	0.35	0.25

Table 8: Pairwise score and weighted pairwise score under an **informative** two-stage SRSWOR sampling design when the true value  $\beta_0 = 2, \beta_1 = 4, \sigma^2 = 0.5, \tau^2 = 1$ .

Informative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	-5.60	6.85	6.98	-466.77	1111.37	1146.43
$\beta_1$	-11.8	14.35	6.95	-613.53	1285.81	1278.82
$\sigma^2$	0.41	2.60	2.37	7.08	569.78	468.09
$\tau^2$	-0.38	2.85	3.19	-163.01	527.12	525.69

#### 4.9.5 Design: two-stage Poisson

Consider an **uninformative** two-stage sample design,

- (i) First stage: Poisson sampling with sampling inclusion probability  $\pi_i$ , where  $\pi_i$  is generated from uniform distribution on  $[0, 1]$ .
- (ii) Second stage: SRSWOR with the sample size  $n_i = 10$ .

Under an **uninformative** sampling design, the main simulation results are the following:

- (i) PLE and WPLE have similar levels of bias, but NMLE has small bias.
- (ii) The esd under-estimates the mad for both PLE and WPL.
- (iii) It is clear NMLE estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for uninformative sampling is 1.83.



Table 9: Performance of NMLE, PLE and WPLE under an **uninformative** Poisson sampling design when the true value  $\beta_0 = 1, \beta_1 = 3, \sigma^2 = 1, \tau^2 = 0.5$ .

Uninformative	NMLE		PLE			WPLE		
parameters	media bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	-0.02	0.06	-0.04	0.08	0.08	-0.03	0.13	0.10
$\beta_1$	0.00	0.05	0.00	0.05	0.04	0.00	0.07	0.05
$\sigma^2$	0.00	0.07	0.00	0.07	0.07	0.00	0.09	0.08
$\tau^2$	-0.07	0.08	-0.09	0.10	0.08	-0.12	0.13	0.09

Table 10: Pairwise score and weighted pairwise score under an **uninformative** Poisson sampling design when the true value  $\beta_0 = 1, \beta_1 = 3, \sigma^2 = 1, \tau^2 = 0.5$ .

Uninformative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	-0.53	1.44	1.53	-105.40	549.73	465.32
$\beta_1$	0.04	2.31	1.98	-57.50	763.15	602.51
$\sigma^2$	-0.33	0.75	0.87	-82.28	269.34	255.38
$\tau^2$	-0.62	0.89	0.91	-162.41	295.49	268.44

Consider an **informative** two-stage sample design,

- (i) First stage: Poisson sampling with sampling inclusion probability

$$\pi_i = \frac{a \exp(-br_i)}{1 + a \exp(-br_i)},$$

where

$$r_i = \sum_{k \in U: s(k)=i} (y_k - \beta_0 - \beta_1 x_k),$$

$$a = 0.2$$

$$b = 0.45.$$

- (ii) Second stage: SRSWOR with the sample size  $n_i = 10$ .

Under an **informative** sampling design, the main simulation results are the following:

- (i) PLE and WPLE have similar levels of bias, but NMLE has small bias.
- (ii) The esd underestimates the mad for both PLE and WPL.
- (iii) It is clear NMLE estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for informative sampling is 0.95.

Table 11: Performance of NMLE, PLE and WPLE under an **informative** Poisson sampling design when the true value  $\beta_0 = 1, \beta_1 = 3, \sigma^2 = 1, \tau^2 = 0.5$ .

Informative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.02	0.13	0.01	0.19	0.16	0.01	0.20	0.16
$\beta_1$	-0.02	0.06	-0.06	0.08	0.08	-0.02	0.10	0.08
$\sigma^2$	-0.02	0.12	-0.01	0.12	0.12	-0.02	0.14	0.12
$\tau^2$	-0.06	0.19	-0.15	0.20	0.13	-0.19	0.20	0.13

Table 12: Pairwise score and weighted pairwise score under an **informative** Poisson sampling design when the true value  $\beta_0 = 1, \beta_1 = 3, \sigma^2 = 1, \tau^2 = 0.5$ .

Informative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	0.15	0.78	0.89	84.82	941.84	939.00
$\beta_1$	-0.61	1.01	1.00	-224.39	1044.01	1130.23
$\sigma^2$	-0.17	0.39	0.44	-198.02	461.58	438.59
$\tau^2$	-0.26	0.51	0.50	-240.66	500.04	492.43

## 4.10 Simulation: Random slope model

A simulation study for a random slope model was conducted to examine performance of the weighted pairwise likelihood estimation and consistency of variance estimation under various uninformative and informative sampling design. Based on 150 simulation replicates, median bias, median absolute deviation (mad) and median estimated standard deviation (esd) are computed to measure the performance of the

- (i) naive maximum likelihood estimation (NMLE).
- (ii) pairwise likelihood estimation (PLE).
- (iii) weighted pairwise likelihood estimation (WPLE).

The code for these simulations is publicly available at <https://github.com/Xudong3/>.

### 4.10.1 Model

A random slope model is given by

$$\begin{aligned} Y_{ik}|a_i, b_i &\sim N(\beta_0 + \beta_1 X_{ik} + a_i + b_i X_{ik}, \sigma^2), \\ \begin{pmatrix} a_i \\ b_i \end{pmatrix} &\sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11}^2 & \tau_{12} \\ \tau_{12} & \tau_{22}^2 \end{pmatrix} \right), \end{aligned}$$

for  $i = 1, \dots, T_c$  and  $k = 1, \dots, T_i$ . The parameters I want to estimate are  $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2, \tau_{11}^2, \tau_{12}, \tau_{22}^2)^T$ .

Observe

$$\begin{aligned}
& \text{Cov}_Y(Y_{ik}, Y_{il}) \\
&= \text{Cov}_Y(\beta_0 + \beta_1 X_{ik} + a_i + b_i X_{ik} + \epsilon_{ik}, \beta_0 + \beta_1 X_{il} + a_i + b_i X_{il} + \epsilon_{il}) \\
&= \text{Cov}_Y(a_i + b_i X_{ik} + \epsilon_{ik}, a_i + b_i X_{il} + \epsilon_{il}) \\
&= \text{Cov}_Y(a_i + b_i X_{ik}, a_i + b_i X_{il}) + \text{Cov}_Y(\epsilon_{ik}, \epsilon_{il}) \\
&= \text{Cov}_Y(a_i, a_i) + X_{il} \text{Cov}_Y(a_i, b_i) + X_{ik} \text{Cov}_Y(b_i, a_i) + X_{ik} X_{il} \text{Cov}_Y(b_i, b_i) + \text{Cov}_Y(\epsilon_{ik}, \epsilon_{il}) \\
&= \tau_{11}^2 + X_{il} \tau_{12} + X_{ik} \tau_{12} + X_{ik} X_{il} \tau_{22}^2 + \text{Cov}_Y(\epsilon_{ik}, \epsilon_{il}) \\
&= \begin{cases} \tau_{11}^2 + 2X_{ik} \tau_{12} + X_{ik}^2 \tau_{22}^2 + \sigma^2, & \text{if } k = l. \\ \tau_{11}^2 + X_{il} \tau_{12} + X_{ik} \tau_{12}^2 + X_{ik} X_{il} \tau_{22}^2, & \text{otherwise.} \end{cases}
\end{aligned}$$

The sample weighted pairwise log likelihood is given by

$$p\ell_{M^s}(\boldsymbol{\theta}) = \sum_{i=1}^{T_c} \sum_{\substack{k < l \\ k, l \in M_i^c}} 1_{kl} \omega_{kl} p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta}),$$

where

$$\begin{aligned}
p\ell_{kl|i}(y_k, y_l, \boldsymbol{\theta}) &= -\frac{1}{2} \log(d_{kl}) - \frac{1}{2} d_{kl}^{-1} \left[ r_k^2 c_{ll} - 2r_k r_l c_{kl} + r_l^2 c_{kk} \right], \\
c_{kk} &= \tau_{11}^2 + 2x_{ik} \tau_{12} + x_{ik}^2 \tau_{22}^2 + \sigma^2, \\
c_{kl} &= \tau_{11}^2 + x_{il} \tau_{12} + x_{ik} \tau_{12}^2 + x_{ik} x_{il} \tau_{22}^2, \\
c_{ll} &= \tau_{11}^2 + 2x_{il} \tau_{12} + x_{il}^2 \tau_{22}^2 + \sigma^2, \\
d_{kl} &= \left( \tau_{11}^2 + 2x_{ik} \tau_{12} + x_{ik}^2 \tau_{22}^2 + \sigma^2 \right) \left( \tau_{11}^2 + 2x_{il} \tau_{12} + x_{il}^2 \tau_{22}^2 + \sigma^2 \right) - \\
&\quad \left( \tau_{11}^2 + x_{il} \tau_{12} + x_{ik} \tau_{12}^2 + x_{ik} x_{il} \tau_{22}^2 \right)^2, \\
r_k &= y_{ik} - \beta_0 - \beta_1 x_{ik}.
\end{aligned}$$

The pairwise score function is given by

$$\begin{aligned}
& \nabla_{\boldsymbol{\theta}} p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta}) \\
&= \begin{pmatrix} \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \beta_0} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \beta_1} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \sigma^2} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \tau_{11}^2} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \tau_{12}} \\ \frac{\partial p_{\ell_{kl|i}}(y_k, y_l, \boldsymbol{\theta})}{\partial \tau_{22}^2} \end{pmatrix} \\
&= \begin{pmatrix} -\frac{1}{2} \frac{1}{d_{kl}} (-2r_k c_{ll} + 2r_l c_{kl} + 2r_k c_{kl} - 2r_l c_{kk}) \\ -\frac{1}{2} \frac{1}{d_{kl}} (-2r_k x_{ik} c_{ll} + 2x_{ik} r_l c_{kl} + 2r_k x_{il} c_{kl} - 2r_l x_{il} c_{kk}) \\ -\frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \sigma^2}}{d_{kl}} + \frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \sigma^2}}{d_{kl}^2} (r_k^2 c_{ll} - 2r_k r_l c_{kl} + r_l^2 c_{kk}) - \frac{1}{2} \frac{1}{d_{kl}} (r_k^2 + r_l^2) \\ -\frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \tau_{11}^2}}{d_{kl}} + \frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \tau_{11}^2}}{d_{kl}^2} (r_k^2 c_{ll} - 2r_k r_l c_{kl} + r_l^2 c_{kk}) - \frac{1}{2} \frac{1}{d_{kl}} (r_k^2 - 2r_k r_l 1_{g(k)=g(l)} + r_l^2) \\ -\frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \tau_{12}}}{d_{kl}} + \frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \tau_{12}}}{d_{kl}^2} (r_k^2 c_{ll} - 2r_k r_l c_{kl} + r_l^2 c_{kk}) - \frac{1}{2} \frac{1}{d_{kl}} (r_k^2 2x_l - 2r_k r_l 1_{g(k)=g(l)} (x_k + x_l) + r_l^2 2x_k) \\ -\frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \tau_{22}^2}}{d_{kl}} + \frac{1}{2} \frac{\frac{\partial d_{kl}}{\partial \tau_{22}^2}}{d_{kl}^2} (r_k^2 c_{ll} - 2r_k r_l c_{kl} + r_l^2 c_{kk}) - \frac{1}{2} \frac{1}{d_{kl}} (r_k^2 x_l^2 - 2r_k r_l 1_{g(k)=g(l)} x_k x_l + r_l^2 x_k^2) \end{pmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial d_{kl}}{\partial \sigma^2} &= c_{ll} + c_{kk}, \\
\frac{\partial d_{kl}}{\partial \tau_{11}^2} &= c_{ll} + c_{kk} - 2c_{kl}, \\
\frac{\partial d_{kl}}{\partial \tau_{12}} &= 2x_k c_{ll} + c_{kk} 2x_l - 2c_{kl} (x_k + x_l), \\
\frac{\partial d_{kl}}{\partial \tau_{22}^2} &= x_k^2 c_{ll} + c_{kk} x_l^2 - 2c_{kl} x_k x_l.
\end{aligned}$$

#### 4.10.2 Design: stratified sampling

Consider a **uninformative** stratified SRSWOR, i.e., SRSWOR with the sample size  $n_i = 10$  for each sampling cluster  $i$ .

Under an **uninformative** sampling design, the main simulation results are the following:

- (i) NMLE, WPLE and PLE have similar levels of bias.
- (ii) The mad is closely estimated by the mad for both PLE and WPL.
- (iii) It is not clear which estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for uninformative sampling is 0.04.



Table 13: Performance of NMLE, PLE and WPLE under an **uninformative** stratified sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 3$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 0.8$ .

Uninformative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.02	0.03	0.00	0.05	0.05	0.00	0.05	0.05
$\beta_1$	0.01	0.03	0.01	0.06	0.05	0.01	0.06	0.05
$\sigma^2$	-0.02	0.03	-0.04	0.04	0.04	-0.04	0.04	0.04
$\tau_{11}^2$	-0.10	0.08	-0.07	0.09	0.09	-0.07	0.09	0.09
$\tau_{12}$	-0.02	0.04	0.01	0.06	0.06	0.01	0.06	0.06
$\tau_{22}^2$	0.20	0.06	0.23	0.08	0.08	0.23	0.08	0.08

Table 14: Pairwise score and weighted pairwise score under an **uninformative** stratified sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 3$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 0.8$ .

Uninformative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	0.23	1.38	1.46	23.04	145.24	152.81
$\beta_1$	0.36	2.06	1.41	36.97	219.49	147.99
$\sigma^2$	0.24	1.10	1.18	24.26	112.62	124.62
$\tau_{11}^2$	0.28	0.79	0.84	29.39	83.48	88.24
$\tau_{12}$	-2.32	1.46	1.21	-248.06	161.464	126.85
$\tau_{22}^2$	3.80	1.24	0.81	405.58	126.55	85.23

Consider an **informative** stratified SRSWOR, i.e., SRSWOR with the sample size

$$n_i = \left\lceil \frac{a \exp (by_i)}{1 + a \exp (by_i)} \right\rceil,$$

where

$$y_i = \sum_{k \in U, s(k)=i} y_k,$$

$$a = 0.15,$$

$$b = 0.25.$$

Under an **informative** sampling design, the main simulation results are the following:

- (i) NMLE, WPLE and PLE generate bias, but PLE has large bias.
- (ii) The mad is closely estimated by the mad for both PLE and WPL.
- (iii) It is not clear which estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for informative sampling is 0.71.

Table 15: Performance of NMLE, PLE and WPLE under an **informative** stratified sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 3$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 0.8$ .

Informative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.02	0.03	-0.17	0.04	0.04	0.02	0.05	0.05
$\beta_1$	0.00	0.02	-0.58	0.05	0.04	0.01	0.06	0.05
$\sigma^2$	-0.01	0.03	-0.04	0.03	0.03	-0.04	0.04	0.04
$\tau_{11}^2$	-0.11	0.07	-0.05	0.07	0.08	-0.05	0.08	0.08
$\tau_{12}$	-0.03	0.03	-0.10	0.05	0.05	0.01	0.06	0.07
$\tau_{22}^2$	0.18	0.05	0.15	0.08	0.06	0.25	0.08	0.09

Table 16: Pairwise score and weighted pairwise score under an **informative** stratified sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 3$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 0.8$ .

Informative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\alpha$	5.41	3.28	2.65	48.12	146.08	136.84
$\beta$	-45.95	4.18	2.63	51.46	196.18	151.54
$\sigma^2$	3.75	2.08	2.26	5.83	98.05	110.82
$\tau_{11}^2$	6.80	1.46	1.59	35.03	76.78	75.19
$\tau_{12}$	-21.29	4.31	2.17	-245.37	144.46	113.99
$\tau_{22}^2$	24.35	4.13	1.49	403.15	142.63	84.19

#### 4.10.3 Design: two-stage SRSWOR

Consider an **uninformative** two-stage sample design,

- (i) First stage: SRSWOR with the sample size  $n_I = 10$ .
- (ii) Second stage: SRSWOR with the sample size  $n_i = 10$ .

Under an **uninformative** sampling design, the main simulation results are the following:

- (i) NMLE, PLE and WPLE generate a biased estimate, but WPLE has large bias.
- (ii) The esd underestimates the mad for both PLE and WPLE.
- (iii) It is not clear which estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for uninformative sampling is 1.42.

Table 17: Performance of NML, PL and WPL under an **uninformative** two-stage SR-SWOR sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 2$ ,  $\sigma^2 = 1$ ,  $\tau_{11}^2 = 0.8$ ,  $\tau_{12} = 0.6$ ,  $\tau_{22}^2 = 1.2$ .

Uninformative	NMLE		PLE			WPLE		
parameters	median bias	mad	media bias	mad	esd( $\tilde{\theta}$ )	media bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.18	0.25	0.17	0.32	0.33	0.15	0.33	0.32
$\beta_1$	0.10	0.27	0.08	0.39	0.36	0.11	0.40	0.36
$\sigma^2$	-0.02	0.18	-0.03	0.21	0.20	-0.03	0.27	0.21
$\tau_{11}^2$	0.11	0.41	-0.06	0.46	0.39	-0.12	0.40	0.37
$\tau_{12}$	-0.02	0.32	-0.12	0.39	0.33	-0.15	0.41	0.33
$\tau_{22}^2$	-0.05	0.43	-0.28	0.47	0.43	-0.36	0.51	0.39

Table 18: Pairwise score and weighted pairwise score under an **uninformative** two-stage SRSWOR sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 2$ ,  $\sigma^2 = 1$ ,  $\tau_{11}^2 = 0.8$ ,  $\tau_{12} = 0.6$ ,  $\tau_{22}^2 = 1.2$ .

Uninformative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	0.24	0.49	0.52	366.15	949.93	939.71
$\beta_1$	0.09	0.58	0.57	172.32	1143.33	986.67
$\sigma^2$	0.01	0.26	0.24	-23.47	521.14	459.61
$\tau_{11}^2$	0.05	0.22	0.21	-2.33	398.6	410.37
$\tau_{12}$	-0.03	0.35	0.37	-17.63	646.14	679.29
$\tau_{22}^2$	-0.02	0.23	0.24	-64.46	495.10	497.78

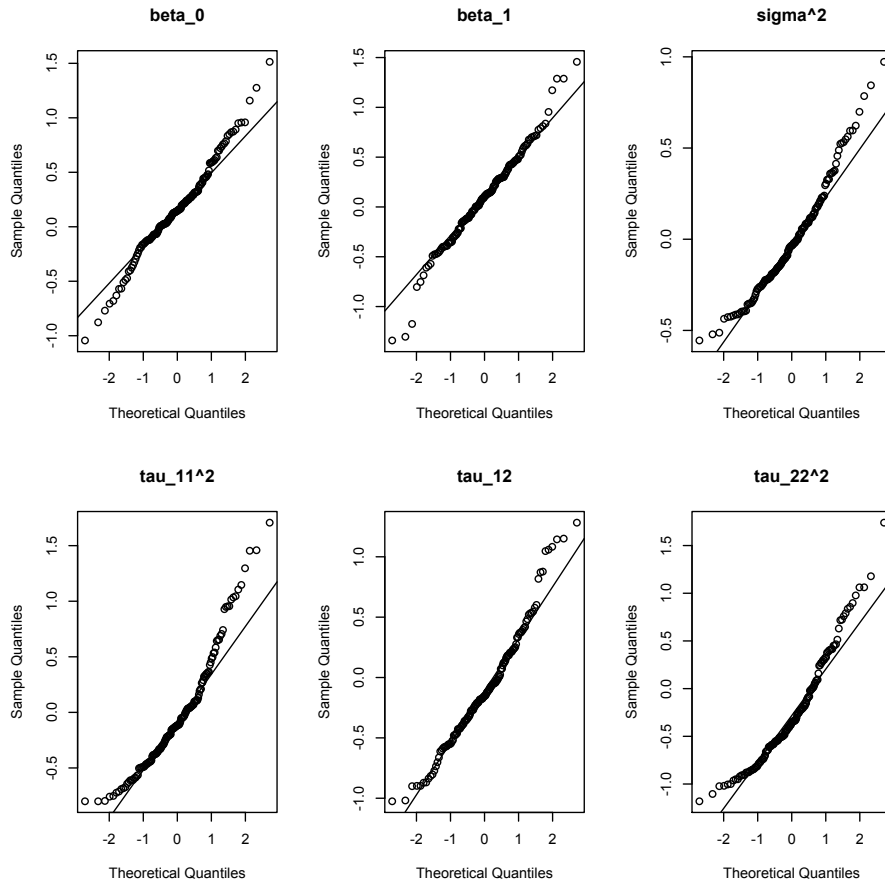


Figure 4.5: QQ plot for the weighted pairwise likelihood estimator under an **uninformative** two-stage SRSWOR sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 2$ ,  $\sigma^2 = 1$ ,  $\tau_{11}^2 = 0.8$ ,  $\tau_{12} = 0.6$ ,  $\tau_{22}^2 = 1.2$ .

Consider an **informative** two-stage sample design,

- (i) First stage: SRSWOR with the sample size  $n_I = 10$ .
- (ii) Second stage: SRSWOR with the sample size

$$n_i = \left\lceil \frac{a \exp(-br_i)}{1 + a \exp(-br_i)} N_i \right\rceil,$$

where

$$r_i = \sum_{k \in U, s(k)=i} (y_k - \beta_0 - \beta_1 x_k),$$

$$a = 0.3$$

$$b = 0.45.$$

Under an **informative** sampling design, the main simulation results are the following:

- (i) NMLE, PLE and WPLE generate a biased estimate, but PLE has very big bias.
- (ii) The esd underestimates the mad for both PLE and WPLE.
- (iii) It is not clear which estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for informative sampling is 2.79.

Table 19: Performance of NMLE, PLE and WPLE under an **informative** two-stage SR-SWOR sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 3$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 0.8$ .

Informative	NMLE		PLE			WPLE		
parameters	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.19	0.15	-0.12	0.39	0.29	0.10	0.31	0.29
$\beta_1$	0.09	0.19	-0.63	0.36	0.24	-0.01	0.44	0.31
$\sigma^2$	-0.03	0.13	0.05	0.15	0.14	0.00	0.20	0.16
$\tau_{11}^2$	-0.01	0.34	-0.17	0.34	0.26	-0.13	0.46	0.33
$\tau_{12}$	-0.01	0.26	-0.28	0.31	0.22	-0.16	0.39	0.28
$\tau_{22}^2$	-0.03	0.28	-0.72	0.35	0.28	-0.36	0.40	0.36

Table 20: Pairwise score and weighted pairwise score under an **informative** two-stage SRSWOR sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 2$ ,  $\sigma^2 = 1$ ,  $\tau_{11}^2 = 0.8$ ,  $\tau_{12} = 0.6$ ,  $\tau_{22}^2 = 1.2$ .

Informative	Pairwise score			Weighted pairwise score		
parameters	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	1.63	5.39	3.97	352.60	785.09	865.40
$\beta_1$	-6.98	7.34	4.41	-62.38	1121.25	926.64
$\sigma^2$	0.12	1.81	1.51	-35.33	357.23	377.96
$\tau_{11}^2$	0.12	1.62	1.64	0.91	407.85	351.69
$\tau_{12}$	-1.22	2.83	3.49	-40.09	496.73	594.07
$\tau_{22}^2$	-0.60	1.56	2.99	-103.49	359.13	388.20



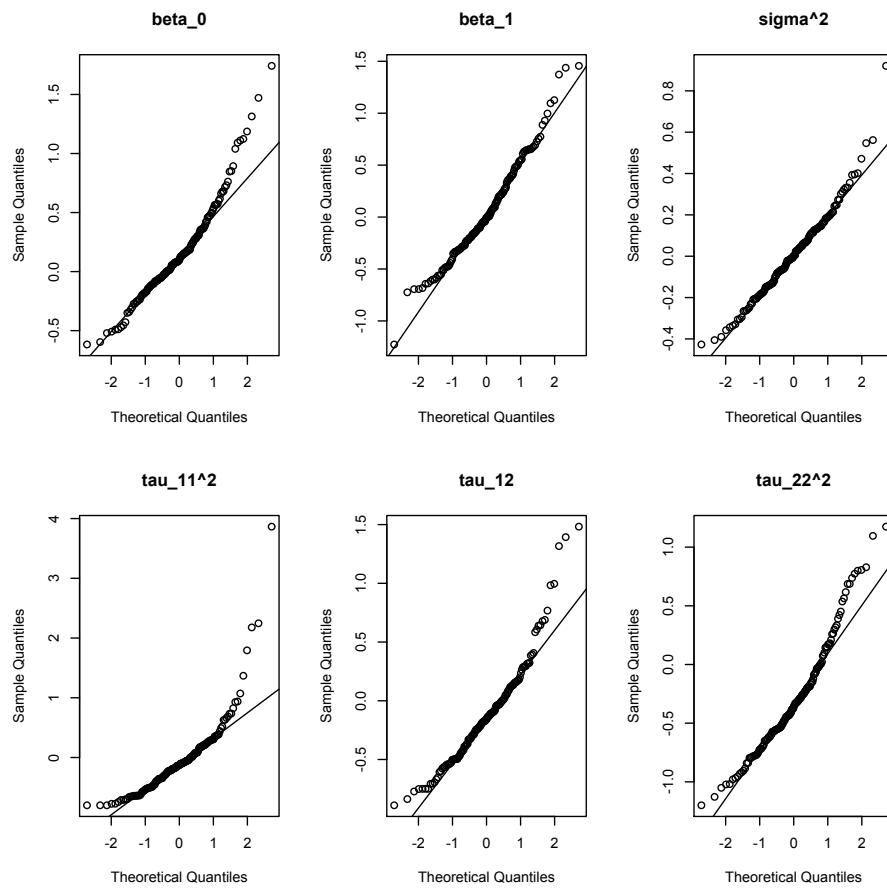


Figure 4.6: QQ plot for the weighted pairwise likelihood estimator under an **informative** two-stage SRSWOR sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 2$ ,  $\sigma^2 = 1$ ,  $\tau_{11}^2 = 0.8$ ,  $\tau_{12} = 0.6$ ,  $\tau_{22}^2 = 1.2$ .

#### 4.10.4 Design: two-stage Poisson

Consider an **uninformative** two-stage sample design with the following design,

- (i) First stage: Poisson sampling with sampling inclusion probability  $\pi_i$ , where  $\pi_i$  is generated from uniform distribution on  $[0, 1]$ .
- (ii) Second stage: SRSWOR with the sample size  $n_i = 10$ .

Under an **uninformative** sampling design, the main simulation results are the following:

- (i) NMLE, PLE and WPLE have similar levels of bias.
- (ii) The esd under-estimates the mad for both PLE and WPLE.
- (iii) It is not clear which estimation method gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for uninformative sampling is 1.74.

Table 21: Performance of NMLE, PLE and WPLE under an **uninformative** two-stage Poisson sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 1.5$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 1.2$ .

Uninformative	NMLE		PLE			WPLE		
parameters	median bias	mad	media bias	mad	esd( $\tilde{\theta}$ )	media bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	0.03	0.09	0.04	0.13	0.12	0.03	0.17	0.15
$\beta_1$	0.11	0.08	0.13	0.13	0.12	0.14	0.18	0.16
$\sigma^2$	-0.01	0.05	-0.03	0.06	0.06	-0.03	0.08	0.07
$\tau_{11}^2$	0.07	0.16	0.04	0.21	0.17	-0.03	0.27	0.20
$\tau_{12}$	0.13	0.12	0.11	0.21	0.14	0.11	0.24	0.17
$\tau_{22}^2$	0.08	0.14	0.12	0.27	0.21	0.05	0.31	0.25

Table 22: Pairwise score and weighted pairwise score under an **uninformative** two-stage Poisson sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 1.5$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 1.2$ .

Uninformative	Pairwise score			Weighted pairwise score		
parameter	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	0.07	1.12	1.12	66.3	443.13	379.73
$\beta_1$	1.11	1.19	1.13	299.46	429.85	395.88
$\sigma^2$	-0.16	0.61	0.71	-44.37	232.07	223.01
$\tau_{11}^2$	0.06	0.63	0.53	-23.76	193.70	166.01
$\tau_{12}$	0.48	0.99	0.82	127.93	309.88	249.39
$\tau_{22}^2$	0.20	0.73	0.55	13.56	182.79	180.84

Consider an **informative** two-stage sample design,

- (i) First stage: SRSWOR with the sampling inclusion probability

$$\pi_i = \frac{a \exp(-br_i)}{1 + a \exp(-br_i)},$$

where

$$r_i = \sum_{k \in U, s(k)=i} (y_k - \beta_0 - \beta_1 x_k),$$

$$a = 0.3$$

$$b = 0.45.$$

- (ii) Second stage: SRSWOR with the sample size  $n_i = 10$ .

Under an **informative** sampling design, the main simulation results are the following:

- (i) NMLE, PLE and WPLE generate a biased estimate, but PLE has a large bias.
- (ii) The esd under-estimates the mad for both PLE and WPLE.
- (iii) NMLE gives a better result.

The coefficients of variation of the weight  $\omega_{kl}$  for informative sampling is 0.75.

Table 23: Performance of NMLE, PLE and WPLE under an **informative** two-stage Poisson sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 1.5$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 1.2$ .

Informative	NMLE		PLE			WPLE		
parameter	median bias	mad	median bias	mad	esd( $\tilde{\theta}$ )	median bias	mad	esd( $\hat{\theta}$ )
$\beta_0$	-0.02	0.09	-0.22	0.31	0.14	-0.04	0.26	0.21
$\beta_1$	-0.01	0.09	-1.77	0.20	0.10	-0.12	0.29	0.21
$\sigma^2$	0.00	0.05	0.14	0.09	0.07	-0.04	0.13	0.09
$\tau_{11}^2$	0.02	0.19	-0.06	0.15	0.18	-0.21	0.34	0.22
$\tau_{12}$	0.12	0.11	-0.20	0.18	0.09	-0.11	0.25	0.17
$\tau_{22}^2$	0.10	0.16	-0.67	0.26	0.12	-0.26	0.33	0.22

Table 24: Pairwise score and weighted pairwise score under an **informative** two-stage Poisson sampling design when the true value  $\beta_0 = 1$ ,  $\beta_1 = 1.5$ ,  $\sigma^2 = 0.8$ ,  $\tau_{11}^2 = 1$ ,  $\tau_{12} = 0.5$ ,  $\tau_{22}^2 = 1.2$ .

Informative	Pairwise score			Weighted pairwise score		
parameter	median bias	mad	esd(PS( $\tilde{\theta}$ ))	median bias	mad	esd(WPS( $\hat{\theta}$ ))
$\beta_0$	17.06	3.72	3.68	-31.06	627.60	506.47
$\beta_1$	-76.66	5.92	5.75	-136.20	575.57	556.00
$\sigma^2$	5.26	2.07	1.93	-153.15	324.10	258.76
$\tau_{11}^2$	11.47	1.34	1.52	-106.65	256.37	209.32
$\tau_{12}$	-33.5	4.72	4.10	53.62	325.65	298.49
$\tau_{22}^2$	47.40	3.70	4.58	-123.43	201.49	204.02

## 5 Matérn spatial model

### 5.1 Introduction

In this chapter, I want to keep extending the asymptotic properties of the sample weighted pairwise likelihood to the case when the random effects could potentially be correlated. The main goal of this chapter is to establish consistency and asymptotic normality of the sample weighted pairwise likelihood estimator under the Matérn spatial model. The key step is to prove the pointwise law of large numbers (PLLN) and the central limit theorem (CLT) for the random field  $\nabla_{\theta} p\ell(\theta) = \{\nabla_{\theta} p\ell_{kl}(\theta) : kl \in \mathcal{A}_{\epsilon}^c\}$ . Once this has been done, the rest is just routine.

I start by introducing the setting, essentially there is nothing new in the design. But for the model, until I set up a lot of definitions, I would not be able to state precisely what I mean. In all of the following, the limit should be interpreted as expanding domain, i.e., I assume both the sample size and population size go to infinity in the limit.

### 5.2 Setting: design

Let  $U \subset \mathbb{R}^m$  be a finite population with spatial location in  $\mathbb{R}^m$  and let  $\|\cdot\|$  be the standard Euclidean norm. Let  $\omega_{kl}$  be the pairwise sampling weight for  $kl \in U^2$ .

### 5.3 Alternative approaches on model

There are two popular approaches on modelling spatially correlated random field  $b = \{b_k : k \in U\}$ :

- (i) Gaussian Markov Random Fields, i.e., directly modelling a sparse precision matrix (Besag, 1974).
- (ii) Marginal models under a mixing condition (Guyon, 1995; Lumley and Heagerty, 1999).

However, both approaches have problems. One possible solution is to model the covariance structure explicitly. This is the approach I will take.

### 5.3.1 Why not Gaussian Markov Random Fields (GMRF)?

Consider a undirected graph  $G = (U, \epsilon)$ , where  $U$  is the set of nodes (i.e., the set of population) and  $\epsilon$  is the set of undirected edges. Recall the precision matrix is defined to be the inverse of covariance matrix.

**Definition 5.1.** A random field  $X = \{X_k : k \in U\}$  with mean  $\mu$  and precision matrix  $Q$  is a Gaussian Markov Random Fields (GMRF) with respect to a undirected graph  $G = (U, \epsilon)$  if the joint distribution function is Gaussian, i.e.,

$$f(x) = (2\pi)^{-\frac{N}{2}} |Q|^{\frac{1}{2}} \exp \left[ -\frac{1}{2} (x - \mu)^T Q (x - \mu) \right]$$

and  $Q_{kl} = 0$  iff  $kl \notin \epsilon$ .

**Remark.** In my definition, there is no restriction on the cardinality of  $\epsilon$ . One could have  $|\epsilon| = O(N^2)$ . But the point is to assume  $|\epsilon|$  is small so that precision matrix is sparse, i.e., if  $|\epsilon| = O(N)$ , then  $Q$  contains only  $O(N)$  non-zero elements.

**Remark.** There is another equivalent formulation of GMRF using the full conditional distribution function  $f(x_k | x_{-k})$ . This approach was first developed by [Besag \(1974\)](#) and known as conditional autoregressive model (CAR). The Hammersley-Clifford Theorem shows under certain regularity conditions the full conditional distribution function  $f(x_k | x_{-k})$  uniquely determines the joint distribution function  $X = \{X_k : k \in U\}$  ([Rue and Held, 2005](#)).

**Lemma 5.2.** Let  $X$  be a GMRF with respect to a undirected graph  $G = (U, \epsilon)$  with mean  $\mu$  and precision matrix  $Q$ , then  $\text{Corr}_Y(x_k, x_l | x_{-kl}) = -\frac{Q_{kl}}{\sqrt{Q_{kk}Q_{ll}}}$ .

*Proof.* For more detail, see [Rue and Held \(2005\)](#). □

**Remark.** In particular, observe

$$x_k \perp x_l | x_{-kl} \iff Q_{kl} = 0.$$

Note  $Q_{kl} = 0 \implies x_k \perp x_l$ .

The problem with GMRF is modelling the precision matrix  $\mathbf{Q}$  does not help us to construct the pairwise likelihood function. To see this, consider a spatial random intercept model,

$$\begin{aligned} Y_k | \mathbf{X}_k, b_k &\sim N(\mathbf{X}_k^T \boldsymbol{\beta} + b_k, \sigma^2), \\ \mathbf{b} &\sim N(\mathbf{0}, \mathbf{Q}^{-1}), \end{aligned}$$

where  $\sigma^2 > 0$ ,  $\mathbf{b} = \{b_k : k \in U\}$  and  $\mathbf{Q}$  is the precision matrix of  $\mathbf{b}$ .

To construct the pairwise likelihood, one needs to know the covariance between all possible correlated observational units from the precision matrix, i.e., one needs to compute  $\text{Cov}_Y(b_k, b_l)$  from  $\mathbf{Q}$ . This is computationally expensive, as one needs to find the inverse of the precision matrix  $\mathbf{Q}^{-1}$ , which is a  $N \times N$  matrix. Observe it is not sufficient to invert a submatrix of  $\mathbf{Q}$ . The way to remediate this problem is by forgetting about the precision matrix and instead focus on modelling the covariance matrix directly (more about this later).

### 5.3.2 Why not marginal models under a mixing conditions?

There are many ways to define the mixing condition in the literature (Bradley et al., 2005). In all of the following, I will study exclusively the  $\alpha$ -mixing conditions.

**Definition 5.3.** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and  $\mathcal{A}, \mathcal{B}$  be two sub- $\sigma$ -algebras of  $\mathcal{F}$ , then the  $\alpha$ -mixing coefficients between  $\mathcal{A}$  and  $\mathcal{B}$  is defined by

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{A}, B \in \mathcal{B}\}.$$

I am working with the weighted pairwise likelihood which are naturally defined on  $U^2 \subset \mathbb{R}^{2m}$ . I start by defining the distance function on  $U^2$ .

**Definition 5.4.** Define the distance function  $\rho(kl, k'l')$  between pairs of observational units to be

$$\begin{aligned} \rho : U^2 \times U^2 &\longrightarrow \mathbb{R}^+ \\ kl \times k'l' &\longmapsto \rho(kl, k'l') = \min\{\|k - k'\|, \|k - l'\|, \|l - k'\|, \|l - l'\|\}. \end{aligned}$$



**Definition 5.5.** Let  $X = \{X_{kl} : kl \in U^2\}$  be a random field and  $p, q$  be two positive integers. Then the  $\alpha$ -mixing coefficients for the random field  $X$  is defined as

$$\alpha_{p,q}^X(r) = \sup\{\alpha(\mathcal{A}, \mathcal{B}) : |A| \leq p, |B| \leq q, \text{dist}(A, B) \geq r\},$$

where  $A, B \subset U^2$ ,  $\text{dist}(A, B) = \min_{kl \in A, k'l' \in B} \rho(kl, k'l')$ ,  $\mathcal{A}$  is the  $\sigma$ -algebra generated by the random variables  $\{X_{kl} : kl \in A\}$  and  $\mathcal{B}$  is the  $\sigma$ -algebra generated by the random variables  $\{X_{kl} : kl \in B\}$ , i.e.,

$$\begin{aligned}\mathcal{A} &= \sigma(X_{kl} : kl \in A), \\ \mathcal{B} &= \sigma(X_{kl} : kl \in B).\end{aligned}$$

Define

$$\begin{aligned}\alpha_{p,\infty}^X(r) &= \sup\{\alpha_{p,q}^X(r) : q \in \mathbb{N}\}, \\ \alpha_{\infty,\infty}^X(r) &= \sup\{\alpha_{p,\infty}^X(r) : p \in \mathbb{N}\}.\end{aligned}$$

**Remark.** Observe  $\alpha_{p,q}^X(r)$  is increasing in  $p, q$  and decreasing in  $r$ . In particular, observe

$$\alpha_{p,q}^X(r) < \alpha_{\infty,\infty}^X(r) \tag{5.1}$$

for all  $p, q \in \mathbb{N}$ .

**Lemma 5.6.** Let  $\psi$  be a real-valued measurable function and  $X = \{X_{kl} : kl \in U^2\}$  be a random field. Define  $\psi(X) = \{\psi(X_{kl}) : kl \in U^2\}$ . Then

$$\alpha_{p,q}^{\psi(X)}(r) \leq \alpha_{p,q}^X(r)$$

for all  $p, q \in \mathbb{N}$  and  $r > 0$ .

*Proof.* Essentially, this is just chasing down the definition. I omit the detail.  $\square$

One can establish the Pointwise Law of Large Numbers (PLLN) and the Central Limit Theorem (CLT) under appropriate mixing conditions as in Theorem 3.3.1 (Guyon, 1995). From this, one can deduce the Uniform Law of Large Numbers (ULLN) from equicontinuity conditions. Then it reduces to a standard argument to show the weighted

pairwise likelihood estimator is consistent and asymptotically normality (more about this in section 5.6 and 5.7).

But there is a problem with the marginal model. I am working under a parametric mixed model. In particular, I am interested in the variance components. The marginal model is basically a semiparametric approach. By construction, a marginal model does not specify variance components.

## 5.4 Matérn covariance function

After settling down on some preliminary definitions, my main goal is to introduce Matérn random field which is one of the most popular isotropic random field in spatial statistics. The key properties of Matérn random field is that it has an exponential decay covariance structure, which is crucial for establishing mixing conditions for the point-wise law of large number and the central limit theorem (more about this in section 5.6 and 5.7).

**Definition 5.7.** The set of random variables

$$X = \{X_k : k \in U \subset \mathbb{R}^m, m \geq 1\}$$

is called a random field.

**Remark.** In all of the following, I assume  $m = 2$ , i.e.,  $U$  is a finite subset of  $\mathbb{R}^2$ . An example is the partition of the US into the counties as shown in the Figure 1.1. Observe the percentage of diagnosed diabetes are attached to each county.

**Definition 5.8.** Let  $X = \{X_k : k \in U\}$  be a random field. Define the mean function  $\mu_k^X$  and covariance function  $C^X(k, l)$  to be

$$\begin{aligned}\mu_k^X &= \mathbb{E}_Y[X_k], \\ C^X(k, l) &= \text{Cov}_Y(X_k, X_l).\end{aligned}$$

**Definition 5.9.** The distribution of the random field  $X = \{X_k : k \in U\}$  is uniquely determined by its finite-dimensional distribution

$$F_{k_1, \dots, k_r}^X(x_1, \dots, x_r) = \mathbb{P}(X_{k_1} \leq x_1, \dots, X_{k_r} \leq x_r)$$

for all  $r \in \mathbb{N}$  and  $k_1, \dots, k_r \in U$ .

**Remark.** One of the most important random field is Gaussian random field. The distribution of Gaussian random field is completely determined by its mean function  $\mu_k^X$  and covariance function  $C^X(k, l)$ .

**Definition 5.10.** A random field  $X = \{X_k : k \in U\}$  is called weakly stationary if its mean and covariance function are translation invariant, i.e.,

$$\begin{aligned}\mu_k^X &= \mu_{k+h}^X, \\ C^X(k+h, l+h) &= C^X(k, l) = C^X(k-l),\end{aligned}$$

for all  $h \in \mathbb{R}^d$  and  $k, l \in U$ .

**Definition 5.11.** An isotropic random field  $X = \{X_k : k \in U\}$  is a weakly stationary random field with an isotropic covariance function, i.e., covariance only depends on the distance between observational units

$$C^X(k, l) = C^X(\|k - l\|)$$

for all  $k, l \in U$ .

In all of the following, I will exclusively study the Matérn covariance function, which is one of the most important isotropic covariance function ([Matérn, 1960](#)). More precisely,

**Definition 5.12.** A Matérn random field  $X = \{X_k : k \in U\}$  is an isotropic random field with a Matérn covariance function, i.e., covariance is given by

$$C_{\tau^2, v, \kappa}^X(k, l) = C_{\tau^2, v, \kappa}^X(\|k - l\|) = \frac{\tau^2}{2^{v-1}\Gamma(v)} (\kappa\|k - l\|)^v K_v(\kappa\|k - l\|),$$

where  $\Gamma$  is the gamma function,  $\tau^2 > 0$  is the marginal variance, and  $K_v$  is the modified Bessell function of the second-kind with shape parameter  $v > 0$  and scale parameter  $\kappa > 0$ . In particular, observe  $C_{\tau^2, v, \kappa}^X(0) = \tau^2$ .

**Remark.** The modified Bessel function of the second-kind is defined as a solution of

$$x^2 y'' + xy' - (x^2 + v^2)y = 0. \quad (5.2)$$

One can show the solution to 5.2 is given by

$$K_v(x) = \frac{\pi}{2} \frac{I_{-v}(x) - I_v(x)}{\sin(v\pi)},$$

where

$$I_v(x) = \left(\frac{x}{2}\right)^v \sum_{i=0}^{\infty} \frac{1}{\Gamma(i+1)\Gamma(v+i+1)} \left(\frac{x}{2}\right)^{2i}. \quad (5.3)$$

Note the power series in 5.3 has an infinite radius of convergence and convergence is rapid (Boyce et al., 1992). Implementations of the Bessel functions are widely available.

**Remark.** As  $\|k - l\| \rightarrow \infty$ , then one has

$$K_v(\kappa\|k - l\|) \approx \sqrt{\frac{\pi}{2\kappa\|k - l\|}} \exp[-\kappa\|k - l\|],$$

i.e.,

$$C_{\tau^2, v, \kappa}^X(\|k - l\|) \approx \frac{\sigma^2}{2^{v-1}\Gamma(v)} (\kappa\|k - l\|)^v \sqrt{\frac{\pi}{2\kappa\|k - l\|}} \exp[-\kappa\|k - l\|].$$

In particular,  $C_{\tau^2, v, \kappa}^X(\|k - l\|)$  is a rapidly decaying function in  $\|k - l\|$  for all  $\tau^2, v > 0$  and  $\kappa > 0$ , i.e.,  $C_{\tau^2, v, \kappa}^X(\|k - l\|) = O(\exp[-\kappa\|k - l\|])$ . This rapid decay simplifies proofs because the integral of  $C_{\tau^2, v, \kappa}^X(r)$  outside a ball of radius  $r$  is of the same or small order than  $\exp[-\kappa\|r\|]$ .

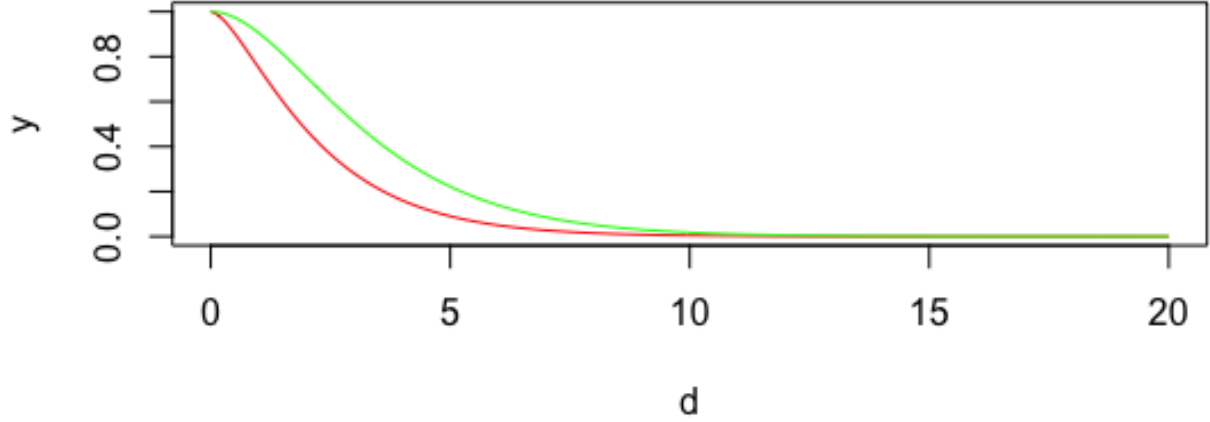


Figure 5.1: Matérn correlation function shown for  $\tau^2 = 1$ ,  $v = 1$ ,  $\kappa = 1$  (red line) and  $\tau^2 = 1$ ,  $v = 2$ ,  $\kappa = 1$  (green line) .

**Remark.** The Matérn covariance function can be thought as a discrete version of solutions to stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} x(k) = B(k),$$

where  $\alpha = v + \frac{m}{2}$ ,  $\kappa > 0$ ,  $v > 0$ ,  $\Delta = \sum_{i=1}^m \frac{\partial^2}{\partial k_i^2}$  and  $B(k)$  is a standard Gaussian random variable (Lindgren et al., 2011).

## 5.5 Setting: Matérn spatial random intercept model

The goal of this section is to introduce Matérn spatial random intercept model and derive some properties of sample weighted pairwise likelihood in this new context. Essentially there is nothing new in here.

Consider a Matérn spatial random intercept model

$$Y_k = \beta_0 + \beta_1 X_{k1} + \cdots + \beta_p X_{kp} + b_k + \epsilon_k,$$

where  $\epsilon = \{\epsilon_k : k \in U\}$  is a Gaussian random field with  $\mu_k^\epsilon = 0$  and  $C^\epsilon(k, l) = \sigma^2 1_{k=l}$  and  $b = \{b_k : k \in U\}$  is a Gaussian Matérn random field with  $\mu_k^b = 0$  and  $C_{\tau^2, v, \kappa}^b(k, l) = C_{\tau^2, v, \kappa}^b(\|k - l\|) = \frac{\tau^2}{2^{v-1}\Gamma(v)} (\kappa\|k - l\|)^v K_v(\kappa\|k - l\|)$ .

The parameters are  $\theta = (\beta_0, \beta_1, \dots, \beta_p, \sigma^2, \tau^2, v, \kappa)^T$ .

**Remark.** Under a Matérn spatial random intercept model, observe  $\text{Cov}_Y(y_k, y_l) = \text{Cov}_Y(b_k, b_l)$  for  $k \neq l$ . In particular,  $\text{Cov}_Y(y_k, y_l) = O(\exp[-\kappa\|k - l\|])$ .

**Definition 5.13.** Let  $\epsilon > 0$ . Define

$$\begin{aligned}\mathcal{A}_\epsilon^c &= \{kl \in U^2 : C_{\tau^2, v, \kappa}^b(\|k - l\|) \geq \epsilon\}, \\ \mathcal{A}_\epsilon^s &= \{kl \in S^2 : C_{\tau^2, v, \kappa}^b(\|k - l\|) \geq \epsilon\}.\end{aligned}$$

**Remark.** The exact value of  $\epsilon$  are user defined.  $\epsilon$  determines which pairs one should use for the calculation of pairwise likelihood. If  $\epsilon = 0$ , then  $\mathcal{A}_\epsilon^c$  is all possible pairs.

**Remark.** For every  $\epsilon \geq 0$ , there exists a  $\delta \geq 0$  such that  $\mathcal{A}_\epsilon^c = \{kl \in U^2 : \|k - l\| \leq \delta\}$ .

**Definition 5.14.** Define the census pairwise log-likelihood to be

$$p\ell^c(\theta) = \sum_{kl \in \mathcal{A}_\epsilon^c} p\ell_{kl}(\theta),$$

where

$$\begin{aligned}p\ell_{kl}(\theta) &= -\frac{1}{2} \log |\Sigma_{kl}| - \frac{1}{2} \mathbf{r}_{kl}^T \Sigma_{kl}^{-1} \mathbf{r}_{kl}, \\ \Sigma_{kl} &= \begin{pmatrix} \tau^2 + \sigma^2, & C_{\tau^2, v, \kappa}^b(\|k - l\|) \\ C_{\tau^2, v, \kappa}^b(\|k - l\|), & \tau^2 + \sigma^2 \end{pmatrix}, \\ \mathbf{r}_{kl} &= \begin{pmatrix} y_k - \beta_0 - \beta_1 x_{k1} - \dots - \beta_p x_{kp} \\ y_l - \beta_0 - \beta_1 x_{l1} - \dots - \beta_p x_{lp} \end{pmatrix}.\end{aligned}$$

**Definition 5.15.** Define the sample weighted pairwise log-likelihood to be

$$p\ell^s(\theta) = \sum_{kl \in \mathcal{A}_\epsilon^c} 1_{kl} \omega_{kl} p\ell_{kl}(\theta).$$

**Lemma 5.16.** *Suppose there exists a  $\delta > 0$  such that  $\sup_{\{k,l\}} \mathbb{E}_Y \left[ \sup_{\{\theta \in \Theta\}} p\ell_{kl}^{1+\delta}(\theta) \right] < \infty$ , then the sample weighted pairwise log-likelihood is a design-unbiased estimator for the census pairwise log-likelihood under design measure  $\pi$ , i.e.,*

$$\mathbb{E}_\pi [p\ell^s(\theta)] = p\ell^c(\theta).$$

*Proof.* The proof is almost exactly the same as Lemma 3.4. Observe

$$\begin{aligned} \mathbb{E}_\pi [p\ell^s(\theta)] &= \mathbb{E}_\pi \left[ \sum_{kl \in \mathcal{A}_\epsilon^c} 1_{kl} \omega_{kl} p\ell_{kl}(\theta) \right] \\ &= \sum_{kl \in \mathcal{A}_\epsilon^c} p\ell_{kl}(\theta) \\ &= p\ell^c(\theta). \end{aligned}$$

□

**Corollary 5.17.** *Let  $h_{kl}(y_k, y_l) = \sup_{\{\theta \in \Theta\}} \|\nabla_{\theta} p\ell_{kl|i}(y_k, y_l, \theta)\|$  and  $g_{kl}(y_k, y_l) = \sup_{\{\theta \in \Theta\}} \|\nabla_{\theta\theta}^2 p\ell_{kl|i}(y_k, y_l, \theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{k,l\}} \mathbb{E}_Y [h_{kl}^{1+\delta}] < \infty$  and  $\sup_{\{k,l\}} \mathbb{E}_Y [g_{kl}^{1+\delta}] < \infty$ , then the sample weighted pairwise score function is a design-unbiased estimator for the census pairwise score function under the design measure  $\pi$ , i.e.,*

$$\begin{aligned} \mathbb{E}_\pi [\nabla_{\theta} p\ell^s(\theta)] &= \nabla_{\theta} p\ell^c(\theta), \\ \mathbb{E}_\pi [\nabla_{\theta\theta}^2 p\ell^s(\theta)] &= \nabla_{\theta\theta}^2 p\ell^c(\theta). \end{aligned}$$

## 5.6 Pointwise Law of Large Numbers

The goal of this section is to show under a Matérn spatial random intercept model, the random field  $\nabla_{\theta} p\ell(\theta) = \{\nabla_{\theta} p\ell_{kl}(\theta) : kl \in \mathcal{A}_\epsilon^c\}$  satisfies mixing condition 5.4 so that one can use the Pointwise Law of Large Numbers (PLLN) (Guyon, 1995; Jenish and Prucha, 2009). There are general conditions in the literature to establish PLLN. Since I only consider Matérn covariance structure, i.e., observations are nearly independent at large distance, essentially all those conditions are trivially satisfied. Let me start by stating two general results.

**Lemma 5.18.** Assume the following regularity conditions,

**A.1** Let  $U_\infty^2 \subset \mathbb{R}^{2m}$  be a countable infinite set such that there exists a  $\delta > 0$  such that  $\rho(kl, k'l') \geq \delta$  for  $kl, k'l' \in U_\infty^2$ .

**A.2** Let  $\{U_N^2 \subset U_\infty^2 : N \in \mathbb{N}\}$  be a sequence of finite population with  $|U_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$ .

**A.3** Let  $X = \{X_{kl} : kl \in \mathcal{A}_\epsilon^c\}$  be a random field. Suppose there exists a  $\delta > 0$  such that  $\sup_{\{kl\}} \mathbb{E}_Y X_{kl}^{1+\delta} < \infty$ .

**A.4** Assume the random field  $X$  satisfies the following mixing conditions

$$\int_0^\infty r^{2m-1} \alpha_{1,1}^X(r) dr < \infty. \quad (5.4)$$

then

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} X_{kl} - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} X_{kl} \right] \xrightarrow{p} 0$$

under the model measure  $Y$ .

**Remark. A.1** is a technical assumption to ensure there is a minimum separation between pairs in the  $U_\infty^2$ . In particular, this implies

$$\sup_{kl \in \mathbb{R}^{2m}} |B(kl, r) \cap U_N| = O(r^{2m}),$$

which is a sufficient condition for convergence when dealing with irregular set  $U_N^2 \subset \mathbb{R}^{2m}$ . One can drop **A.1**, if  $U^2$  is just a lattice, i.e.,  $U^2 \subset \mathbb{Z}^{2m}$ . For more detail, see [Guyon \(1995\)](#).

*Proof.* See [Jenish and Prucha \(2009\)](#) for the proof. □

**Lemma 5.19.** Let  $X = \{X_{kl} : kl \in \mathbb{Z}^{2m}\}$  be a stationary Gaussian random field such that  $\text{Cov}_Y(X_{kl}, X_{k'l'}) = O(\rho(kl, k'l')^{-t})$  for some  $t > 2m$  and the characteristic function of  $X$  is bounded below, then  $\alpha_{\infty, \infty}^X(r) = O(r^{2m-t})$ .

*Proof.* For the proof, see page 59 from [Doukan \(1994\)](#). □



I now turn to the Pointwise Law of Large Numbers (PLLN) for the random field  $\nabla_{\theta} p\ell(\theta)$  under a Matérn spatial random intercept model. To prove this, I need to introduce some notation to start with.

**Definition 5.20.** Define the random field  $a(\theta) = \{a_{kl}(\theta) : kl \in \mathcal{A}_{\epsilon}^c\}$  and  $d(\theta) = \{d_{kl}(\theta) : kl \in \mathcal{A}_{\epsilon}^c\}$  associated with random field  $Y(\theta) = \{Y_k(\theta) : k \in U\}$  to be

$$\begin{aligned} a_{kl}(\theta) &= Y_k(\theta) + Y_l(\theta), \\ d_{kl}(\theta) &= Y_k(\theta) - Y_l(\theta). \end{aligned}$$

**Remark.** Observe

$$\begin{aligned} & \text{Cov}_Y(a_{kl}(\theta), a_{k'l'}(\theta)) \\ &= \text{Cov}_Y(Y_k(\theta) + Y_l(\theta), Y_{k'}(\theta) + Y_{l'}(\theta)) \\ &= \text{Cov}_Y(Y_k(\theta), Y_{k'}(\theta)) + \text{Cov}_Y(Y_k(\theta), Y_{l'}(\theta)) + \text{Cov}_Y(Y_l(\theta), Y_{k'}(\theta)) + \text{Cov}_Y(Y_l(\theta), Y_{l'}(\theta)) \\ &= O(\exp[-\kappa\rho(kl, k'l')]), \end{aligned} \tag{5.5}$$

where  $\rho(kl, k'l') = \min\{\|k - k'\|, \|k - l'\|, \|l - k'\|, \|l - l'\|\}$ . In particular,  $a(\theta)$  is a Gaussian random field with the exponential decay covariance structure.

Similarly, one can show

$$\text{Cov}_Y(d_{kl}(\theta), d_{k'l'}(\theta)) = O(\exp[-\kappa\rho(kl, k'l')]), \tag{5.6}$$

$$\text{Cov}_Y(a_{kl}(\theta), d_{k'l'}(\theta)) = O(\exp[-\kappa\rho(kl, k'l')]). \tag{5.7}$$

**Lemma 5.21.** Observe  $a_{kl}(\theta)$  is independent of  $d_{kl}(\theta)$  under the model measure  $Y$ , i.e.,  $a_{kl} \perp d_{kl}$ .

*Proof.* Note  $a_{kl}(\theta)$  and  $d_{kl}(\theta)$  are Gaussian random variables from Theorem 7.22. Thus it suffices to show they are uncorrelated. Observe

$$\begin{aligned} & \text{Cov}_Y(a_{kl}(\theta), d_{kl}(\theta)) \\ &= \text{Cov}_Y(Y_k(\theta) + Y_l(\theta), Y_k(\theta) - Y_l(\theta)) \\ &= \text{Cov}_Y(Y_k(\theta), Y_k(\theta)) - \text{Cov}_Y(Y_k(\theta), Y_l(\theta)) + \text{Cov}_Y(Y_l(\theta), Y_k(\theta)) - \text{Cov}_Y(Y_l(\theta), Y_l(\theta)) \\ &= 0. \end{aligned}$$

This completes the proof.  $\square$

My strategy below is to decompose  $\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta})$  into two pieces (Lemma 5.22), and for each piece I have an explicit bound on the mixing coefficients (Lemma 5.23) so that I can apply Lemma 5.18. More precisely,

**Lemma 5.22.** *There exists a real-valued measurable function  $\psi_1$  and  $\psi_2$  such that*

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) + \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})).$$

*Proof.* Observe

$$(Y_k(\boldsymbol{\theta}), Y_l(\boldsymbol{\theta})) = \left( \frac{a_{kl}(\boldsymbol{\theta}) + d_{kl}(\boldsymbol{\theta})}{2}, \frac{a_{kl}(\boldsymbol{\theta}) - d_{kl}(\boldsymbol{\theta})}{2} \right).$$

In particular, the pairwise density function  $f(y_k(\boldsymbol{\theta}), y_l(\boldsymbol{\theta}))$  can be expressed as a function of  $a_{kl}(\boldsymbol{\theta})$  and  $d_{kl}(\boldsymbol{\theta})$ , i.e., there exists a measurable function  $\zeta$  such that

$$f(y_k(\boldsymbol{\theta}), y_l(\boldsymbol{\theta})) = \zeta(a_{kl}(\boldsymbol{\theta}), d_{kl}(\boldsymbol{\theta})).$$

From Lemma 5.21,  $a_{kl}(\boldsymbol{\theta})$  is independent from  $d_{kl}(\boldsymbol{\theta})$  under model measure  $Y$ . Hence there exists a measurable function  $\zeta_1$  and  $\zeta_2$  such that

$$\zeta(a_{kl}(\boldsymbol{\theta}), d_{kl}(\boldsymbol{\theta})) = \zeta_1(a_{kl}(\boldsymbol{\theta})) \zeta_2(d_{kl}(\boldsymbol{\theta})),$$

i.e.,

$$f(y_k(\boldsymbol{\theta}), y_l(\boldsymbol{\theta})) = \zeta_1(a_{kl}(\boldsymbol{\theta})) \zeta_2(d_{kl}(\boldsymbol{\theta})).$$

Taking logs on both sides and then differentiating, one has

$$\nabla_{\boldsymbol{\theta}} p^{\ell_{kl}}(\boldsymbol{\theta}) = \psi_1(a_{kl}(\boldsymbol{\theta})) + \psi_2(d_{kl}(\boldsymbol{\theta})), \quad (5.8)$$

where  $\psi_1(a_{kl}(\boldsymbol{\theta})) = \nabla_{\boldsymbol{\theta}} \log(\zeta_1(a_{kl}(\boldsymbol{\theta})))$  and  $\psi_2(d_{kl}(\boldsymbol{\theta})) = \nabla_{\boldsymbol{\theta}} \log(\zeta_2(d_{kl}(\boldsymbol{\theta})))$ . Summing over all possible pairs and then normalising by  $\frac{1}{|\mathcal{A}_\epsilon^c|}$ , one has

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) + \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})).$$

$\square$

**Lemma 5.23.** Let  $\mathbf{Y} = (Y_1, \dots, Y_N)^\top$  follow a Matérn spatial random intercept model. The random field  $\psi_1(a(\boldsymbol{\theta})) = \{\psi_1(a_{kl}(\boldsymbol{\theta})) : kl \in \mathcal{A}_\epsilon^c\}$  and  $\psi_2(d(\boldsymbol{\theta})) = \{\psi_2(d_{kl}(\boldsymbol{\theta})) : kl \in \mathcal{A}_\epsilon^c\}$  satisfy the following mixing condition

$$\begin{aligned}\alpha_{\infty,\infty}^{\psi_1(a(\boldsymbol{\theta}))}(r) &= O(r^{2m-t}), \\ \alpha_{\infty,\infty}^{\psi_2(d(\boldsymbol{\theta}))}(r) &= O(r^{2m-t}),\end{aligned}$$

for every integer  $t > 2m$ .

*Proof.* From 5.5, one has  $\text{Cov}_Y(a_{kl}(\boldsymbol{\theta}), a_{k'l'}(\boldsymbol{\theta})) = O(\exp[-\kappa\rho(kl, k'l')])$ . In particular, observe

$$\text{Cov}_Y(a_{kl}(\boldsymbol{\theta}), a_{k'l'}(\boldsymbol{\theta})) = O(\rho(kl, k'l')^{-t}) \quad (5.9)$$

for every integer  $t > 2m$ .

Note the random field  $a(\boldsymbol{\theta}) = \{a_{kl}(\boldsymbol{\theta}) : kl \in \mathcal{A}_\epsilon^c\}$  is a Gaussian random field. From 5.9 and Lemma 5.19, one has

$$\alpha_{\infty,\infty}^{a(\boldsymbol{\theta})}(r) = O(r^{2m-t}) \quad (5.10)$$

for every integer  $t > 2m$ .

Therefore, from 5.10 and Lemma 5.6, one gets

$$\alpha_{\infty,\infty}^{\psi_1(a(\boldsymbol{\theta}))}(r) = O(r^{2m-t})$$

for every integer  $t > 2m$ .

The case of  $\psi_2(d(\boldsymbol{\theta}))$  can be handle similarly, I omit the detail. □

**Corollary 5.24.** Observe

$$\begin{aligned}\int_0^\infty r^{2m-1} \alpha_{1,1}^{\psi_1(a(\boldsymbol{\theta}))}(r) dr &< \infty, \\ \int_0^\infty r^{2m-1} \alpha_{1,1}^{\psi_2(d(\boldsymbol{\theta}))}(r) dr &< \infty.\end{aligned}$$

*Proof.* This is clear from Lemma 5.23 and 5.1. □

**Corollary 5.25.** Observe

- (i) There exists a  $\delta > 0$  such that  $\int_0^\infty r^{2m-1} \alpha_{1,1}^{\psi_1(a(\boldsymbol{\theta}))}(r)^{\frac{\delta}{2+\delta}} dr < \infty$ .
- (ii)  $\int_0^\infty r^{2m-1} \alpha_{p,q}^{\psi_1(a(\boldsymbol{\theta}))}(r) dr < \infty$  if  $p + q \leq 4$ .
- (iii)  $\alpha_{1,\infty}^{\psi_1(a(\boldsymbol{\theta}))}(r) = O(r^{-2m})$ .
- (iv) There exists a  $\delta > 0$  such that  $\int_0^\infty r^{2m-1} \alpha_{1,1}^{\psi_2(d(\boldsymbol{\theta}))}(r)^{\frac{\delta}{2+\delta}} dr < \infty$ .
- (v)  $\int_0^\infty r^{2m-1} \alpha_{p,q}^{\psi_2(d(\boldsymbol{\theta}))}(r) dr < \infty$  if  $p + q \leq 4$ .
- (vi)  $\alpha_{1,\infty}^{\psi_2(d(\boldsymbol{\theta}))}(r) = O(r^{-2m})$ .

*Proof.* This is clear from Lemma 5.23 and 5.1. □

**Remark.** Corollary 5.25 is only needed for the proof of the central limit theorem in next section.

I am ready to state the Pointwise Law of Large Numbers under a Matérn spatial random intercept model.

**Theorem 5.26.** Let  $\mathbf{Y} = (Y_1, \dots, Y_N)^\top$  follow a Matérn spatial random intercept model. In addition, assume the following regularity conditions,

- A.1 Let  $U_\infty^2 \subset \mathbb{R}^{2m}$  be a countable infinite set such that there exists a  $\delta > 0$  such that  $\rho(kl, k'l') \geq \delta$  for  $kl, k'l' \in U_\infty^2$ .
- A.2 Let  $\{U_N^2 \subset U_\infty^2 : N \in \mathbb{N}\}$  be a sequence of finite population with  $|U_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$ .
- A.3 Let  $h_{kl}(y_k, y_l) = \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla p\ell_{kl}(\boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{kl\}} \mathbb{E}_Y h_{kl}^{1+\delta} < \infty$ .

then

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p\ell^c(\boldsymbol{\theta}) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p\ell^c(\boldsymbol{\theta}) \right] \xrightarrow{p} 0$$

under the model measure  $Y$ .

*Proof.* By Lemma 5.22, it suffices to show

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) \right] \xrightarrow{p} 0, \quad (5.11)$$

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) \right] \xrightarrow{p} 0. \quad (5.12)$$

But 5.11 and 5.12 are clear from Corollary 5.24 and Lemma 5.18. This completes the proof.  $\square$

## 5.7 Central Limit Theorem

The goal of this section is to show under a Matérn spatial random intercept model, the random field  $\nabla_{\boldsymbol{\theta}} p\ell(\boldsymbol{\theta}) = \{\nabla_{\boldsymbol{\theta}} p\ell_{kl}(\boldsymbol{\theta}) : kl \in \mathcal{A}_\epsilon^c\}$  satisfies certain mixing conditions A.4 in Lemma 5.27, so that one can use the Central Limit Theorem (CLT). There are general conditions in the literature to establish CLT (Guyon, 1995). Since I only consider Matérn covariance structure, i.e., observations are nearly independent at large distance, essentially all those conditions are trivially satisfied. Let me start by stating two general results.

**Lemma 5.27.** *Assume the following regularity conditions,*

**A.1** *Let  $U_\infty^2 \subset \mathbb{R}^{2m}$  be a countable infinite set such that there exists a  $\delta > 0$  such that  $\rho(kl, k'l') \geq \delta$  for  $kl, k'l' \in U_\infty^2$ .*

**A.2** *Let  $\{U_N^2 \subset U_\infty^2 : N \in \mathbb{N}\}$  be a sequence of finite population with  $|U_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$ .*

**A.3** *Let  $X = \{X_{kl} : kl \in \mathcal{A}_\epsilon^c\}$  be a random field. Suppose there exists a  $\delta > 0$  such that  $\sup_{\{kl\}} \mathbb{E}_Y X_{kl}^{2+\delta} < \infty$ .*

**A.4** *Assume the random field  $X$  satisfies the following mixing conditions*

(i) *There exists a  $\delta > 0$  such that  $\int_0^\infty r^{2m-1} \alpha_{1,1}^X(r)^{\frac{\delta}{2+\delta}} dr < \infty$ .*

(ii)  *$\int_0^\infty r^{2m-1} \alpha_{p,q}^X(r) dr < \infty$  if  $p + q \leq 4$ .*

$$(iii) \alpha_{1,\infty}^X(r) = O(r^{-2m}).$$

$$\text{A.5 } \lim |\mathcal{A}_\epsilon^c| \sigma_Y^2 > 0, \text{ where } \sigma_Y^2 = \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} X_{kl} \right].$$

then

$$\sigma_Y^{-1} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} X_{kl} - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} X_{kl} \right] \right) \xrightarrow{\mathcal{D}} N(0, 1)$$

under the model measure  $Y$ .

*Proof.* See [Guyon \(1995\)](#); [Jenish and Prucha \(2009\)](#) for the proof.  $\square$

**Remark.** [Guyon \(1995\)](#) proves the results initially for a grid of points  $\mathbb{Z}^{2m}$ , using moments bound and Stein's Lemma ([Stein, 1981](#)). It is stated that [A.1](#) replaces the grid assumption to allow any set without accumulation point ([Guyon, 1995](#)). [Jenish and Prucha \(2009\)](#) relate the moment and mixing assumptions of [Guyon \(1995\)](#) to give [A.1](#).

**Lemma 5.28.** *Suppose*

$$\begin{aligned} X_n &\xrightarrow{\mathcal{D}} X, \\ Y_n &\xrightarrow{\mathcal{D}} Y, \end{aligned}$$

and  $X_n$  is independent of  $Y_n$ , then  $X$  is independent of  $Y$  and  $(X_n, Y_n)^T \xrightarrow{\mathcal{D}} (X, Y)^T$ .

*Proof.* Since  $X_n \xrightarrow{\mathcal{D}} X$  and  $Y_n \xrightarrow{\mathcal{D}} Y$ , then by definition, one has

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} [f(X_n)] &= \mathbb{E} [f(X)], \\ \lim_{n \rightarrow \infty} \mathbb{E} [g(Y_n)] &= \mathbb{E} [g(Y)], \end{aligned}$$

for all bounded continuous functions  $f, g$ . Hence

$$\lim_{n \rightarrow \infty} [\mathbb{E} [f(X_n)] \mathbb{E} [g(Y_n)]] = \mathbb{E} [f(X)] \mathbb{E} [g(Y)]. \quad (5.13)$$

Since  $X_n$  is independent of  $Y_n$ , by [Theorem 7.25](#), one has

$$\mathbb{E} [f(X_n)g(Y_n)] = \mathbb{E} [f(X_n)] \mathbb{E} [g(Y_n)] \quad (5.14)$$

for all bounded continuous functions  $f, g$ .

Putting 5.13 and 5.14 together, one has

$$\lim_{n \rightarrow \infty} \mathbb{E} [f(X_n)g(Y_n)] = \mathbb{E} [f(X)] \mathbb{E} [g(Y)]. \quad (5.15)$$

By Dominated Convergence Theorem 7.24, one has

$$\mathbb{E} \left[ \lim_{n \rightarrow \infty} [f(X_n)g(Y_n)] \right] = \mathbb{E} [f(X)] \mathbb{E} [g(Y)],$$

i.e.,

$$\mathbb{E} \left[ \lim_{n \rightarrow \infty} f(X_n) \lim_{n \rightarrow \infty} g(Y_n) \right] = \mathbb{E} [f(X)] \mathbb{E} [g(Y)],$$

i.e.,

$$\mathbb{E} [f(X)g(Y)] = \mathbb{E} [f(X)] \mathbb{E} [g(Y)],$$

for all bounded continuous functions  $f, g$ . Therefore, by Theorem 7.25,  $X$  is independent of  $Y$ .

To prove  $(X_n, Y_n)^T \xrightarrow{\mathcal{D}} (X, Y)^T$ , by Theorem 7.21, it suffices to show

$$c_1 X_n + c_2 Y_n \xrightarrow{\mathcal{D}} c_1 X + c_2 Y \quad (5.16)$$

for all  $c_1, c_2 \in \mathbb{R}$ .

Let  $\phi_X$  be the characteristic function of  $X$ , i.e.,

$$\begin{aligned} \phi_X : \mathbb{R} &\longrightarrow \mathbb{C} \\ t &\longmapsto \phi_X(t) = \mathbb{E} [\exp(itX)]. \end{aligned}$$

To prove 5.16, by Lévy's Continuity Theorem 7.23, it suffices to show

$$\lim_{n \rightarrow \infty} \phi_{c_1 X_n + c_2 Y_n}(t) = \phi_{c_1 X + c_2 Y}(t)$$

for all  $c_1, c_2, t \in \mathbb{R}$ .

Observe

$$\begin{aligned}
\lim_{n \rightarrow \infty} \phi_{c_1 X_n + c_2 Y_n}(t) &= \lim_{n \rightarrow \infty} \mathbb{E} [\exp(it(c_1 X_n + c_2 Y_n))] \\
&= \lim_{n \rightarrow \infty} \mathbb{E} [\exp(itc_1 X_n)] \mathbb{E} [\exp(itc_2 Y_n)] \\
&= \lim_{n \rightarrow \infty} \mathbb{E} [\exp(itc_1 X_n)] \lim_{n \rightarrow \infty} \mathbb{E} [\exp(itc_2 Y_n)] \\
&= \mathbb{E} \left[ \lim_{n \rightarrow \infty} \exp(itc_1 X_n) \right] \mathbb{E} \left[ \lim_{n \rightarrow \infty} \exp(itc_2 Y_n) \right] \\
&= \mathbb{E} [\exp(itc_1 X)] \mathbb{E} [\exp(itc_2 Y)] \\
&= \mathbb{E} [\exp(it(c_1 X + c_2 Y))] \\
&= \phi_{c_1 X + c_2 Y}(t).
\end{aligned}$$

In the second equality, I used the fact  $X_n \perp Y_n$  and Theorem 7.25. In the fourth equality, I used Dominated Convergence Theorem 7.24. In the fifth equality, I used continuous mapping theorem and the fact  $X_n \xrightarrow{\mathcal{D}} X, Y_n \xrightarrow{\mathcal{D}} Y$ . In the sixth equality, I used the fact  $X \perp Y$  and Theorem 7.25.

This completes the proof.  $\square$

I am ready to state the Central Limit Theorem under a Matérn spatial random intercept model. My strategy below is to decompose  $\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta})$  into two pieces (Lemma 5.22), and for each piece I have an explicit bound on the mixing coefficients (Lemma 5.23) so that I can apply Lemma 5.27. Then I used Lemma 5.28 to recover  $\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta})$ .

**Theorem 5.29.** *Let  $\mathbf{Y} = (Y_1, \dots, Y_N)^T$  follow a Matérn spatial random intercept model. In addition, assume the following regularity conditions,*

**A.1** *Let  $U_\infty^2 \subset \mathbb{R}^{2m}$  be a countable infinite set such that there exists a  $\delta > 0$  such that  $\rho(kl, k'l') \geq \delta$  for  $kl, k'l' \in U_\infty^2$ .*

**A.2** *Let  $\{U_N^2 \subset U_\infty^2 : N \in \mathbb{N}\}$  be a sequence of finite population with  $|U_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$ .*

**A.3** *Let  $h_{kl}(y_k, y_l) = \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla p_{kl}(\boldsymbol{\theta})\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{kl\}} \mathbb{E}_Y h_{kl}^{2+\delta} < \infty$ .*

**A.4**  *$\lim |\mathcal{A}_\epsilon^c| \mathbf{J}_Y(\boldsymbol{\theta}) > \mathbf{0}$ , where  $\mathbf{J}_Y(\boldsymbol{\theta}) = \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) \right]$ .*



then

$$\mathbf{J}_Y(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I})$$

under the model measure  $Y$ .

*Proof.* By Corollary 5.25 and Lemma 5.27, one has

$$\begin{aligned} & \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) \right]^{-\frac{1}{2}} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}), \\ & \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) \right]^{-\frac{1}{2}} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}). \end{aligned}$$

Hence,

$$\begin{aligned} & \sqrt{|\mathcal{A}_\epsilon^c|} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) \right] \right) \xrightarrow{\mathcal{D}} \\ & N \left( \mathbf{0}, \lim |\mathcal{A}_\epsilon^c| \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_1(a_{kl}(\boldsymbol{\theta})) \right] \right), \end{aligned} \quad (5.17)$$

and

$$\begin{aligned} & \sqrt{|\mathcal{A}_\epsilon^c|} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) \right] \right) \xrightarrow{\mathcal{D}} \\ & N \left( \mathbf{0}, \lim |\mathcal{A}_\epsilon^c| \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \psi_2(d_{kl}(\boldsymbol{\theta})) \right] \right). \end{aligned} \quad (5.18)$$

Putting 5.17 and 5.18 together, then by Lemma 5.28 and Lemma 5.22, one has

$$\sqrt{|\mathcal{A}_\epsilon^c|} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \lim |\mathcal{A}_\epsilon^c| \mathbf{J}_Y(\boldsymbol{\theta})),$$

i.e.,

$$\mathbf{J}_Y(\boldsymbol{\theta})^{-\frac{1}{2}} \left( \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) - \mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^c}(\boldsymbol{\theta}) \right] \right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}).$$

This completes the proof.  $\square$

## 5.8 Consistency

In all of the following,  $\hat{\theta}_n$  will be the sample weighted pairwise likelihood estimator and  $\theta_0$  will be the true value of the model. More precisely,

**Definition 5.30.** The sample weighted pairwise likelihood estimator  $\hat{\theta}_n$  of  $\theta$  is defined as a solution of

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta} p^{\ell^s}(\theta) = 0.$$

**Definition 5.31.** The true value  $\theta_0$  of  $\theta$  is defined as a solution of

$$\mathbb{E}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta} p^{\ell^c}(\theta) \right] = 0.$$

I am ready to state the weighted pairwise likelihood estimator is consistent under a Matérn spatial random intercept model. Essentially, there is nothing new here.

**Theorem 5.32.** Let  $Y = (Y_1, \dots, Y_N)^T$  follow a Matérn spatial random intercept model. In addition, assume the following regularity conditions,

**A.1** Let  $U_\infty^2 \subset \mathbb{R}^{2m}$  be a countable infinite set such that there exists a  $\delta > 0$  such that  $\rho(kl, k'l') \geq \delta$  for  $kl, k'l' \in U_\infty^2$ .

**A.2** Let  $\{U_N^2 \subset U_\infty^2 : N \in \mathbb{N}\}$  be a sequence of finite population with  $|U_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$  and  $\{S_N^2 \subset U_N^2 : N \in \mathbb{N}\}$  be a sequence of finite sample with  $|S_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$ .

**A.3**  $\Theta$  is a compact subset of  $\mathbb{R}^p$  and  $\theta_0$  is an interior point of  $\Theta$ .

**A.4** Let  $h_{kl}(y_k, y_l) = \sup_{\theta \in \Theta} \|\nabla p^{\ell_{kl}}(\theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{kl \in \mathcal{A}_\epsilon^c\}} \mathbb{E}_Y h_{kl}^{1+\delta} < \infty$ .

**A.5** For any given  $c > 0$  and a given sequence  $\{(y_k, y_l)\}$  satisfying  $(\|y_k\|^2 + \|y_l\|^2)^{\frac{1}{2}} \leq c$ , the sequence of function  $\{\nabla_{\theta} p^{\ell_{kl}}(y_k, y_l, \theta) : kl \in \mathcal{A}_\epsilon^c\}$  is equicontinuous on any open subset  $A$  of  $\Theta$ .

**A.6** For all variables  $V_{kl}$  satisfying  $\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} V_{kl}^2 = O_p(1)$ , we have

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} 1_{kl} \omega_{kl} V_{kl}(\boldsymbol{\theta}) - \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} V_{kl}(\boldsymbol{\theta}) = O_p(n_I^{-\frac{1}{2}})$$

with respect to design probability  $\pi$ .

**A.7** For all  $\epsilon > 0$ , there exists a  $\delta > 0$  such that

$$\inf_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \epsilon\}} \left\| \mathbb{E}_{Y\pi} \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}) \right] \right\| > \delta.$$

then

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$$

with respect to model-design probability  $Y\pi$ .

I proceed the same way as before to argue the sample weighted pairwise likelihood estimator is consistent by establishing a uniform law of large numbers (ULLN) on a compact set.

**Lemma 5.33.** With the same conditions **A.1** – **A.6** as in Theorem 5.32, then one has

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}) - \mathbb{E}_{Y\pi} \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\boldsymbol{\theta}} p^{\ell^s}(\boldsymbol{\theta}) \right] \right\| \xrightarrow{p} 0 \quad (5.19)$$

with respect to design-model probability  $Y\pi$ .

*Proof.* Using Theorem 5.26, this can be done exactly the same as Lemma 3.9. □

I now proceed to the proof of Theorem 5.32.

*Proof.* This can be done exactly the same as Theorem 3.8. I omit the detail. □

## 5.9 Asymptotic normality

I am ready to state the weighted pairwise likelihood estimator is asymptotical normal under a Matérn spatial random intercept model. Essentially, there is nothing new here.

**Theorem 5.34.** *Let  $\mathbf{Y} = (Y_1, \dots, Y_N)^T$  follow a Matérn spatial random intercept model. In addition, assume the following regularity conditions,*

**A.1** *Let  $U_\infty^2 \subset \mathbb{R}^{2m}$  be a countable infinite set such that there exists a  $\delta > 0$  such that  $\rho(kl, k'l') \geq \delta$  for  $kl, k'l' \in U_\infty^2$ .*

**A.2** *Let  $\{U_N^2 \subset U_\infty^2 : N \in \mathbb{N}\}$  be a sequence of finite population with  $|U_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$  and  $\{S_N^2 \subset U_N^2 : N \in \mathbb{N}\}$  be a sequence of finite sample with  $|S_N^2| \rightarrow \infty$  as  $N \rightarrow \infty$ .*

**A.3**  $\Theta$  *is a compact subset of  $\mathbb{R}^p$  and  $\theta_0$  is an interior point of  $\Theta$ .*

**A.4** *Let  $h_{kl}(y_k, y_l) = \sup_{\theta \in \Theta} \|\nabla p \ell_{kl}(\theta)\|$ . Suppose there exists a  $\delta > 0$  such that  $\sup_{\{kl\}} \mathbb{E}_Y h_{kl}^{2+\delta} < \infty$ .*

**A.5** *For any given  $c > 0$  and a given sequence  $\{(y_k, y_l)\}$  satisfying  $(\|y_k\|^2 + \|y_l\|^2)^{\frac{1}{2}} \leq c$ , the sequence of function  $\{\nabla_{\theta} p \ell_{kl}(\theta)\}$  is equicontinuous on any open subset  $A$  of  $\Theta$ .*

**A.6** *For all variables  $V_{kl}$  satisfying  $\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} V_{kl}^2 = O_p(1)$ , one has*

$$\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} 1_{kl} \omega_{kl} V_{kl} - \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} V_{kl} = O_p(n_I^{-\frac{1}{2}})$$

*with respect to design probability  $\pi$ .*

**A.7** *For all  $\epsilon > 0$ , there exists a  $\delta > 0$  such that*

$$\inf_{\{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon\}} \left\| \mathbb{E}_{Y\pi} \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta} p \ell^s(\theta) \right] \right\| > \delta.$$

**A.8** *Let  $g_{kl}(y_k, y_l) = \sup_{\theta \in \Theta} \|\nabla_{\theta}^2 p \ell_{kl}(\theta)\|$ . Suppose there exists a  $\delta_1 > 0$  such that  $\sup_{\{kl \in \mathcal{A}_\epsilon^c\}} \mathbb{E}_Y g_{kl}^{2+\delta_1} < \infty$ .*

**A.9** For any given  $c > 0$  and a given sequence  $\{(y_k, y_l)\}$  satisfying  $(\|y_k\|^2 + \|y_l\|^2)^{\frac{1}{2}} \leq c$ , the sequence of function  $\{\nabla_{\theta\theta}^2 p_{\ell_{kl}}(y_k, y_l, \theta) : kl \in \mathcal{A}_\epsilon^c\}$  is equicontinuous on any open subset  $A$  of  $\Theta$ .

**A.10** For any variable  $V_{kl}$  satisfies the following conditions,

- (a)  $\frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} 1_{kl} \omega_{kl} V_{kl}(\theta) = O(1)$ .
- (b)  $\lim |\mathcal{A}_\epsilon^c| \sigma_\pi^2 > 0$ , where  $\sigma_\pi^2 = \text{Var}_\pi \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} 1_{kl} \omega_{kl} V_{kl}(\theta) \right]$ .

then

$$\sigma_\pi^{-1} \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} 1_{kl} \omega_{kl} V_{kl}(\theta) - \frac{1}{|\mathcal{A}_\epsilon^c|} \sum_{kl \in \mathcal{A}_\epsilon^c} \omega_{kl} V_{kl}(\theta) \right] \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I})$$

with respect to design probability  $\pi$ .

**A.11**  $\lim |\mathcal{A}_\epsilon^s| \mathbf{J}_\pi > \mathbf{0}$  and  $\lim |\mathcal{A}_\epsilon^c| \mathbf{J}_Y > \mathbf{0}$ , where  $\mathbf{J}_\pi(\theta) = \text{Var}_\pi \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta\theta} p_{\ell_{M^s}}(\theta) \right]$  and  $\mathbf{J}_Y(\theta) = \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta\theta} p_{\ell_{M^c}}(\theta) \right]$ .

**A.12** Assume  $\lim \frac{|\mathcal{A}_\epsilon^s|}{|\mathcal{A}_\epsilon^c|} = \zeta$ , where  $\zeta \in [0, 1]$ .

then

$$\mathbf{J}_\pi(\theta_0)^{-\frac{1}{2}} \mathbf{H}(\theta_0) (\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N \left( \mathbf{0}, \mathbf{I} + \zeta \left[ (\lim |\mathcal{A}_\epsilon^c| \mathbf{J}_\pi(\theta))^{-1} (\lim |\mathcal{A}_\epsilon^c| \mathbf{J}_Y(\theta)) \right] \right), \quad (5.20)$$

where

$$\begin{aligned} \mathbf{H}(\theta) &= \mathbb{E}_{Y\pi} \left[ -\frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta\theta}^2 p_{\ell^s}(\theta) \right], \\ \mathbf{J}_\pi(\theta) &= \text{Var}_\pi \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta\theta} p_{\ell^s}(\theta) \right], \\ \mathbf{J}_Y(\theta) &= \text{Var}_Y \left[ \frac{1}{|\mathcal{A}_\epsilon^c|} \nabla_{\theta\theta} p_{\ell^c}(\theta) \right]. \end{aligned}$$

In particular, if the sampling fraction is small, then

$$\mathbf{J}_\pi(\theta_0)^{-\frac{1}{2}} \mathbf{H}(\theta_0) (\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}). \quad (5.21)$$

*Proof.* This essentially reduces to the same argument as Theorem 4.25. I omit the detail.  $\square$

## 6 Future work

In this chapter, I want to mention briefly what is more to be done. I will concentrate on computational aspects of variance estimation and prediction (or estimation) of random effect.

As shown in section 4.9 and 4.10, I get consistent estimation under complex sampling for the weighted pairwise likelihood estimation. However, sandwich variance estimator underestimates the variance. I believe this is due to the small simulations replicate and sample size. Of course, one really needs a simulation to show this. It is impractical to run a large simulation using current codes. Therefore computational method needs to be improved in the future. I believe this would be feasible such as writing all the codes in C to optimise the performance, but this is just a pain.

I have not talked about how to make a prediction (or estimation) of random effects in this thesis. This is not because this is unimportant. I believe this is a very hard problem, and it is not clear how to construct a solution. More explicitly, there are two problems in here. First, from 3.1, one pair of observation does not make random effects  $b_i$  identifiable. Secondly, suppose the random effects  $b_i$  is identifiable, then one gets different prediction of random effect for each pairs. It is not clear how to construct a single prediction of random effect from those.

Computation for generalized linear mixed models is also difficult because the standard approach of Laplace approximation or adaptive Gaussian quadrature requires estimation of the random effects to determine quadrature points.

## 7 Appendix I: elementary result

In this chapter, I recall some basic results, techniques or notation that will be used thorough in this thesis. I only state the results, as the proof can be found in many standard textbooks (Van der Vaart, 2000; Shao, 2003; Brockwell and Davis, 2013; Durrett, 2010; Kallenberg, 2006; Simon et al., 2015).

### 7.1 Basic notation and definition

Let  $\theta \in \Theta \subset \mathbb{R}^p$ , define

$$\begin{aligned}\nabla_{\theta} &= \left( \frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_p} \right)^T, \\ \nabla_{\theta\theta}^2 &= \left( \frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_p} \right)^T \left( \frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_p} \right) \\ &= \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2}{\partial \theta_p^2} \end{pmatrix}.\end{aligned}$$

Define

$$C^k(\Theta) = \{f : \Theta \rightarrow \mathbb{R} : f \text{ is } k \text{ times continuously differentiable}\}.$$

**Definition 7.1.** Let  $x \in \mathbb{R}^p$ , I define the Euclidean norm of  $x$  to be

$$\|x\| = \left( x_1^2 + x_2^2 + \cdots + x_p^2 \right)^{\frac{1}{2}}.$$

**Definition 7.2.** Let  $X = (x_{ij})_{i,j=1,\dots,p}$  be a  $p \times p$  matrix. Define the norm of  $X$  to be

$$\|X\| = \left( \sum_{i,j=1}^p x_{ij}^2 \right)^{\frac{1}{2}}. \quad (7.1)$$

**Definition 7.3.** Let  $A, B \subset \mathbb{R}^p$ , I define the distance between  $A$  and  $B$  to be

$$\text{dist}(A, B) = \inf\{\|a - b\| : a \in A, b \in B\}.$$

**Definition 7.4.** Let  $x \in \mathbb{R}^p$  and  $\epsilon > 0$ . Define  $\epsilon$ -ball to be

$$B_\epsilon(x) = \{y \in \mathbb{R}^p : \|y - x\| < \epsilon\}.$$

**Definition 7.5.** Let  $A$  be a bounded subset of  $\mathbb{R}^p$ , I define the diameter of  $A$  by

$$\text{diam}(A) = \sup\{\|x - y\| : x, y \in A\}.$$

**Definition 7.6.** Let  $\{X_n\}_{n \in \mathbb{N}}$  and  $\{Y_n\}_{n \in \mathbb{N}}$  be a sequence of random variables. We say  $X_n = O_p(Y_n)$  if there exists a  $c > 0$

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( \frac{|X_n|}{|Y_n|} \geq c \right) = 0.$$

**Definition 7.7.** Let  $\{X_n\}_{n \in \mathbb{N}}$  and  $\{Y_n\}_{n \in \mathbb{N}}$  a sequence of random variables. We say  $X_n = o_p(Y_n)$  if for all  $c > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \frac{|X_n|}{|Y_n|} \geq c \right) = 0.$$

**Definition 7.8.** Let  $\{X_n\}_{n \in \mathbb{N}}$  be a sequence of random variables.  $X_n \xrightarrow{\mathcal{D}} X$  if for all bounded continuous functions  $f$

$$\lim_{n \rightarrow \infty} \mathbb{E} [f(X_n)] = \mathbb{E} [f(X)].$$

**Definition 7.9.** A sequence of random variables  $\{X_n(\theta)\}_{n \in \mathbb{N}}$  is said to be equicontinuous if for all  $\epsilon > 0$ , there exists a  $\delta > 0$  such that if  $\|\theta_1 - \theta_2\| \leq \delta$ , then

$$\sup_{n \in \mathbb{N}} \|X_n(\theta_1) - X_n(\theta_2)\| \leq \epsilon.$$

**Definition 7.10.** Let  $X$  be a random variable. Then the characteristic function of  $X$  is defined to be

$$\phi_X(t) = \mathbb{E} [\exp(itX)].$$

**Definition 7.11.** Let  $(X, d)$  be a metric space and let  $A \subset X$ ,  $A$  is totally bounded if for every  $\epsilon > 0$ , it can be covered by finitely many  $\epsilon$  ball.

**Remark.** Let  $(X, d)$  be a metric space and let  $A \subset X$ .  $A$  is compact iff  $A$  is totally bounded and complete.



## 7.2 Basic results

**Lemma 7.12.** *Let  $A$  and  $B$  be nonempty bounded subsets of  $\mathbb{R}$ , then*

- (i)  $\sup A = -\inf(-A)$ , where  $-A = \{-a : a \in A\}$ .
- (ii) If  $c > 0$ ,  $\sup cA = c \sup A$ , where  $cA = \{ca : a \in A\}$ .
- (iii)  $\sup(A + B) = \sup(A) + \sup(B)$ , where  $A + B = \{a + b : a, b \in A\}$ .

**Lemma 7.13.** *Let  $X$  be a random variable, then*

- (i)  $\sup_{\theta \in \Theta} \mathbb{E}[X] \leq \mathbb{E}[\sup_{\theta \in \Theta} X]$ .
- (ii)  $\inf_{\theta \in \Theta} \mathbb{E}[X] \geq \mathbb{E}[\inf_{\theta \in \Theta} X]$ .

**Lemma 7.14** (Hölder's Inequality). *Let  $p > 1, q > 1$  such that  $\frac{1}{p} + \frac{1}{q} = 1$ , then*

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} (\mathbb{E}[|Y|^q])^{\frac{1}{q}}.$$

**Lemma 7.15** (Cauchy-Schwartz Inequality).

$$\mathbb{E}[|XY|] \leq \left(\mathbb{E}[|X|^2]\right)^{\frac{1}{2}} \left(\mathbb{E}[|Y|^2]\right)^{\frac{1}{2}}.$$

**Lemma 7.16** (Markov's Inequality). *Let  $X$  be a non-negative random variable such that  $\mathbb{E}[X] < \infty$  and let  $c > 0$ , then*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}.$$

**Lemma 7.17** (Chebyshev's Inequality). *Let  $X$  be a non-negative random variable such that  $\mathbb{E}[X] < \infty$  and let  $c > 0$ , then*

$$\mathbb{P}(|X - \mathbb{E}X| \geq c) \leq \frac{1}{c^2} \text{Var}[X].$$

**Theorem 7.18** (Taylor's Theorem). *Let  $\theta_0 \in \mathbb{R}^p$  and let  $\mathbf{r}_n$  be a sequence of vector in  $\mathbb{R}^p$  such that  $\lim_{n \rightarrow \infty} \mathbf{r}_n \rightarrow \mathbf{0}$ . Let  $\{\theta_n\}_{n \in \mathbb{N}}$  be a sequence of random vectors such that  $\theta_n - \theta_0 = O_p(\mathbf{r}_n)$ . Suppose  $f : \mathbb{R}^p \rightarrow \mathbb{R}$  is  $C^2$  differentiable, then*

$$f(\theta_n) = f(\theta_0) + \nabla_{\mathbf{X}} f(\theta_0)(\theta_n - \theta_0)^T + o_p(\mathbf{r}_n).$$

**Theorem 7.19** (Slutsky's Theorem). Let  $X_n \xrightarrow{\mathcal{D}} X$  and  $Y_n \xrightarrow{p} c$ , where  $c$  is a constant, then

$$(i) \quad X_n + Y_n \xrightarrow{\mathcal{D}} X + c.$$

$$(ii) \quad X_n Y_n \xrightarrow{\mathcal{D}} cX.$$

$$(iii) \quad Y_n^{-1} X_n \xrightarrow{\mathcal{D}} c^{-1}X.$$

**Theorem 7.20** (Continuous Mapping Theorem). Let  $X_n$  be a sequence of random variable such that  $X_n \xrightarrow{p} X$  and  $g$  is continuous function, then  $g(X_n) \xrightarrow{p} g(X)$ .

**Theorem 7.21** (Cramer-Wold Theorem). Let  $\mathbf{X}_n$  be a sequence of  $\mathbb{R}^p$ -dimensional random vector, then

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X} \iff \mathbf{a}^T \mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{a}^T \mathbf{X} \quad \text{for all } \mathbf{a} \in \mathbb{R}^p.$$

**Theorem 7.22.** A random vector  $(X_1, \dots, X_N)^T$  is a multivariate normal random vector iff

$$a_1 X_1 + \dots + a_N X_N$$

is a normal random variable for all  $a_1, \dots, a_N \in \mathbb{R}$ .

**Theorem 7.23** (Lévy's Continuity Theorem). Let  $\{X_n\}_{n \in \mathbb{N}}$  be a sequence of random variables with characteristic function  $\{\phi_{X_n}(t)\}_{n \in \mathbb{N}}$ .

$$(i) \quad \text{If } X_n \xrightarrow{\mathcal{D}} X, \text{ then } \phi_{X_n}(t) \rightarrow \phi_X(t) \text{ for all } t.$$

$$(ii) \quad \text{If } \phi_{X_n}(t) \rightarrow \phi_X(t) \text{ for all } t \text{ and } \phi_X(t) \text{ is continuous at } t = 0, \text{ then } X_n \xrightarrow{\mathcal{D}} X.$$

**Theorem 7.24** (Dominated Convergence Theorem). Let  $\{X_n\}_{n \in \mathbb{N}}$  be a sequence of random variables such that  $X_n \xrightarrow{p} X$ . Suppose there exists a random variable  $Y$  such that  $|X_n| \leq Y$  and  $\mathbb{E}[|Y|] < \infty$ , then  $\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \mathbb{E}[X]$ .

**Theorem 7.25.**  $X$  and  $Y$  are independent iff for all bounded continuous functions  $f, g$ ,

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \mathbb{E}[g(Y)]$$

## 8 References

- Ardilly, P. and Tillé, Y. (2006), *Sampling Methods: Exercises and Solutions*, Springer Science & Business Media. 15
- Bates, D., Maechler, M., Bolker, B., Walker, S. et al. (2014), 'lme4: Linear mixed-effects models using eigen and s4', *R Package Version 1*(7), 1–23. 125
- Bates, D., Mullen, K., Nash, J. and Varadhan, R. (2014), 'minqa: Derivative-free optimization algorithms by quadratic approximation', *R Package Version 1*(4). 130
- Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 192–236. 2, 60, 164, 165
- Billingsley, P. (2013), *Convergence of Probability Measures*, John Wiley & Sons. 97
- Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review/Revue Internationale de Statistique* pp. 279–292. 76
- Boistard, H., Lopuhaä, H. P., Ruiz-Gazen, A. et al. (2017), 'Functional central limit theorems for single-stage sampling designs', *The Annals of Statistics* 45(4), 1728–1758. 93
- Boyce, W. E., DiPrima, R. C. and Meade, D. B. (1992), *Elementary Differential Equations and Boundary Value Problems*, Vol. 9, Wiley New York. 170
- Bradley, R. C. et al. (2005), 'Basic properties of strong mixing conditions: a survey and some open questions', *Probability Surveys* 2, 107–144. 166
- Breslow, N. E. and Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* 88(421), 9–25. 2, 40, 41, 62
- Brockwell, P. J. and Davis, R. A. (2013), *Time Series: Theory and Methods*, Springer Science & Business Media. 189
- Carrillo-Garcia, I. A. (2008), Analysis of longitudinal surveys with missing responses, PhD thesis, University of Waterloo. 68

- Cassel, C.-M., Särndal, C. E. and Wretman, J. H. (1977), *Foundations of Inference in Survey Sampling*, Wiley. 15
- Centers for Disease Control and Prevention (CDC) (2007), Estimates of the percentage of adults with diagnosed diabetes division of diabetes translation, Technical report, U.S. Department of Health and Human Service. 10, 11
- Cochran, W. G. (2007), *Sampling Techniques*, John Wiley & Sons. 15
- Conomos, M. P., Laurie, C. A., Stilp, A. M., Gogarten, S. M., McHugh, C. P., Nelson, S. C., Sofer, T., Fernández-Rhodes, L., Justice, A. E., Graff, M. et al. (2016), 'Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos', *The American Journal of Human Genetics* 98(1), 165–184. 9
- Cox, D. R. and Reid, N. (2004), 'A note on pseudolikelihood constructed from marginal densities', *Biometrika* 91(3), 729–737. 62
- Demidenko, E. (2013), *Mixed Models: Theory and Applications with R*, John Wiley & Sons. 15, 44, 48, 50
- Diggle, P. (2002), *Analysis of Longitudinal Data*, Oxford University Press. 40
- Doukan, P. (1994), *Mixing. Lectures Notes in Statistics*, 85, Springer-Verlag. 174
- Durrett, R. (2010), *Probability: Theory and Examples*, Cambridge University Press. 189
- Fuller, W. A. (2011), *Sampling Statistics*, Vol. 560, John Wiley & Sons. 15
- Gelman, A. (2007), 'Struggles with survey weighting and regression modeling', *Statistical Science* pp. 153–164. 54
- Grace, G. Y. (2016), *Statistical Analysis with Measurement Error or Misclassification*, Springer. 15, 54
- Grömping, U. (1996), 'A note on fitting a marginal model to mixed effects log-linear regression data via GEE', *Biometrics* pp. 280–285. 50

- Guyon, X. (1995), *Random Fields on a Network: Modeling, Statistics and Applications*, Springer Science & Business Media. 164, 167, 173, 174, 179, 180
- Hall, A. R. (2005), *Generalized Method of Moments*, Oxford University Press. 15
- Heagerty, P. J. and Lele, S. R. (1998), 'A composite likelihood approach to binary spatial data', *Journal of the American Statistical Association* 93(443), 1099–1111. 58, 60
- Isaki, C. T. and Fuller, W. A. (1982), 'Survey design under the regression superpopulation model', *Journal of the American Statistical Association* 77(377), 89–96. 52
- Jenish, N. and Prucha, I. R. (2009), 'Central limit theorems and uniform laws of large numbers for arrays of random fields', *Journal of Econometrics* 150(1), 86–98. 173, 174, 180
- Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Science & Business Media. 15
- Kallenberg, O. (2006), *Foundations of Modern Probability*, Springer Science & Business Media. 189
- Kim, J. K. and Park, M. (2010), 'Calibration estimation in survey sampling', *International Statistical Review* 78(1), 21–39. 54
- Kim, J. K. and Shao, J. (2013), *Statistical Methods for Handling Incomplete Data*, CRC Press. 15, 54
- Laird, N. M. and Ware, J. H. (1982), 'Random-effects models for longitudinal data', *Biometrics* pp. 963–974. 43
- LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J. et al. (2010), 'Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos', *Annals of Epidemiology* 20(8), 642–649. 9
- Liang, K.-Y. and Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* 73(1), 13–22. 49

- Lindgren, F., Rue, H. and Lindström, J. (2011), ‘An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498. 171
- Lindsay, B. G. (1988), ‘Composite likelihood methods’, *Contemporary Mathematics* **80**(1), 221–239. 3, 60
- Lu, T.-T. and Shiou, S.-H. (2002), ‘Inverses of  $2 \times 2$  block matrices’, *Computers & Mathematics with Applications* **43**(1-2), 119–129. 59
- Lumley, T. (1998), Marginal regression modelling of weakly dependent data, PhD thesis, University of Washington (US). 112
- Lumley, T. and Heagerty, P. (1999), ‘Weighted empirical adaptive variance estimators for correlated data regression’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(2), 459–477. 164
- Lumley, T. and Mayer Hamblett, N. (2003), ‘Asymptotics for marginal generalized linear models with sparse correlations’, <http://biostats.bepress.com/cgi/viewcontent.cgi?article=1030&context=uwbiostat>. 112
- Lumley, T. and Scott, A. (2017), ‘Fitting regression models to survey data’, *Statistical Science* **32**(2), 265–278. 54
- Matérn, B. (1960), Spatial variation, PhD thesis, University of Stockholm (Sweden). 169
- McCullagh, P. and Nelder, J. (1989), *Generalised Linear Modelling*, Chapman and Hall: New York. 40
- National Heart, Lung, and Blood Institute (2010), ‘Hispanic Community Health Study/Study of Latinos’. 9
- Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991), ‘A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data’, *International Statistical Review/Revue Internationale de Statistique* pp. 25–35. 39, 50

- Pfeffermann, D. (1996), 'The use of sampling weights for survey data analysis', *Statistical Methods in Medical Research* 5(3), 239–261. 3, 8, 63
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998), 'Weighting for unequal selection probabilities in multilevel models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(1), 23–40. 8, 12, 52
- Pfeffermann, D. and Sverchkov, M. (1999), 'Parametric and semi-parametric estimation of regression models fitted to survey data', *Sankhyā: The Indian Journal of Statistics, Series B* pp. 166–186. 8
- Pfeffermann, D. and Sverchkov, M. (2009), 'Inference under informative sampling', *Handbook of Statistics* 29, 455–487. 8, 53
- Pfeffermann, D. and Sverchkov, M. Y. (2003), 'Fitting generalized linear models under informative sampling', *Analysis of Survey Data* pp. 175–195. 8
- Pinheiro, J. C. and Chao, E. C. (2006), 'Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models', *Journal of Computational and Graphical Statistics* 15(1), 58–81. 2, 41, 62
- Rabe-Hesketh, S. and Skrondal, A. (2006), 'Multilevel modelling of complex survey data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(4), 805–827. 8
- Rao, J., Verret, F. and Hidirolou, M. A. (2013), 'A weighted composite likelihood approach to inference for two-level models from survey data', *Survey Methodology* 39(2), 263–282. 3, 8, 13, 60, 61
- Rubin-Bleuer, S. and Kratina, I. S. (2005), 'On the two-phase framework for joint model and design-based inference', *The Annals of Statistics* pp. 2789–2810. 52, 93, 97
- Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, CRC Press. 165
- Särndal, C.-E., Swensson, B. and Wretman, J. (2003), *Model Assisted Survey Sampling*, Springer Science & Business Media. 15

- Särndal, C.-E., Thomsen, I., Hoem, J. M., Lindley, D., Barndorff-Nielsen, O. and Dale-  
nius, T. (1978), ‘Design-based and model-based inference in survey sampling [with  
discussion and reply]’, *Scandinavian Journal of Statistics* pp. 27–52. 15
- Scott, A. J. and Wild, C. J. (1997), ‘Fitting regression models to case-control data by  
maximum likelihood’, *Biometrika* **84**(1), 57–71. 54
- Scott, A. and Wild, C. (2001), ‘Maximum likelihood for generalised case-control studies’,  
*Journal of Statistical Planning and Inference* **96**(1), 3–27. 54
- Shao, J. (2003), *Mathematical Statistics*, Springer-Verlag. 68, 189
- Simon, B. et al. (2015), *A Comprehensive Course in Analysis*, American Mathematical Soci-  
ety Providence, Rhode Island. 189
- Song, P. X.-K. (2007), *Correlated Data Analysis: Modeling, Analytics and Applications*,  
Springer Science & Business Media. 15
- Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L.,  
Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M. et al. (2010), ‘De-  
sign and implementation of the Hispanic Community Health Study/Study of Latinos’,  
*Annals of Epidemiology* **20**(8), 629–641. 9
- Stein, C. M. (1981), ‘Estimation of the mean of a multivariate normal distribution’, *The  
Annals of Statistics* pp. 1135–1151. 180
- Tillé, Y. and Matei, A. (2009), ‘Sampling: survey sampling’, *R Package Version 2*. 125
- Van der Vaart, A. W. (2000), *Asymptotic Statistics*, Vol. 3, Cambridge University Press. 52,  
56, 189
- Varin, C., Reid, N. and Firth, D. (2011), ‘An overview of composite likelihood methods’,  
*Statistica Sinica* pp. 5–42. 3
- Wedderburn, R. W. (1974), ‘Quasi-likelihood functions, generalized linear models, and  
the gaussnewton method’, *Biometrika* **61**(3), 439–447. 44



- Xu, X. (2012), Aspects of composite likelihood estimation and prediction, PhD thesis, University of Toronto (Canada). 62
- Yi, G. Y., Rao, J. and Li, H. (2016), 'A weighted composite likelihood approach for analysis of survey data under two-level models', *Statistica Sinica* **26**, 569–587. 3, 8, 13, 60, 61, 66, 68, 75, 80
- Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988), 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics* pp. 1049–1060. 50