Distribution Agreement

Signature:

________________________________ _______________

Lijia Wang Date

Composite Conditional Likelihood

By

Lijia Wang
Doctor of Philosophy

Biostatistics

John J. Hanfelt, Ph.D.
Advisor

Andrew Hill, Ph.D.
Committee Member

Ying Guo, Ph.D.
Committee Member

Howard Chang, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Composite Conditional Likelihood

By

Lijia Wang M.S., Emory University, 2015

Advisor: John J. Hanfelt, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy
in Biostatistics
2016

Abstract

Composite Conditional Likelihood

By Lijia Wang

Sparse clustered data often arise in genetic and epidemiologic studies, such as studies of complex diseases that tend to aggregate within households. Fine stratification is needed to adjust for the population heterogeneity that is commonly encountered in such studies. In this dissertation, we consider a finely stratified study, and we aim to find a robust and flexible method to measure the pairwise associations between individual outcomes within clusters.

We propose a composite conditional likelihood approach that allows the investigator to specify only marginal densities and intracluster pairwise densities, and is insensitive to the stratum-specific nuisance parameters. We investigate the asymptotic properties of our composite conditional likelihood method under both the standard situation and the sparse data situation.

We also develop and apply general odds ratio models, which accommodate either discrete or continuous observations, for use in composite conditional likelihood. We propose a specific odds ratio model for use in household aggregation studies, which not only rewards pairwise departure from the references points but also penalizes lack of agreement. We demonstrate via simulation studies that the proposed method provides a valid and flexible way to obtain robust inference in studies of pairwise association with sparse clustered data. We apply the method to a study of drinking water quality within households, finely stratified by small geographic area..

We finally conduct an exploratory study regarding the selection of weights for each cluster in the proposed composite conditional likelihood to improve efficiency of estimation. We investigate the optimal choice of cluster-specific weights under both the standard situation and the sparse data situation. We demonstrate via simulation studies that the efficiency of estimation is improved. We reanalyze the water quality data after incorporating the proposed weights and compare the results analyzed under equal and unequal weights.

Composite Conditional Likelihood

By

Lijia Wang
M.S., Emory University, 2015

Advisor: John J. Hanfelt, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy
in Biostatistics
2016

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Sparse clustered data often arise in genetic and epidemiologic studies, such as studies of complex diseases that tend to aggregate within households. Fine stratification is needed to adjust for the population heterogeneity which is commonly encountered in such studies. The scientific interest might be in the effects of covariates on the marginal outcome and the intracluster associations. This inferential problem is challenging for several reasons. Fully parametric models would require specification of third- and higher-order associations of the observations within clusters. However, it is difficult to verify the assumptions made on the higher order associations, and the resulting inferences on interest parameters would be sensitive to such assumptions (Liang et al., 1992). The generalized estimating equation (GEE) appproach (Liang and Zeger 1986; Liang et al. 1992) and pairwise likelihood (PL) approach (Lindsay 1988; Le Cessie and Van Houwelingen 1994) do not require full specification of the likelihood, but yield inconsistent inferences of parameters with sparse data due to sensitivity to the estimation of stratum-specific parameters: this becomes the “infinitely many nuisance parameters” problem in the sparse data context (Neyman and Scott, 1948). Random-effects models avoid the above nuisance parameter problem, but often do not provide flexible models of pairwise association; moreover, random-effects models require assumptions about the joint distribution of stratum-level random effects, conditional upon other study covariates, and hence, inferences on parameters of interest might be sensitive to such assumptions.

In my dissertation, we consider a finely stratified study consisting of a total of p strata with n_i independent clusters in the ith stratum, and m_{ij} correlated observations in its jth cluster (See Table 1.1 for the data structure). We assume that individual outcomes y_{ijk} marginally are distributed according to a generalized linear model with canonical link and stratum-specific intercept λ_i:

$$f_{ijk}(y_{ijk};\lambda_i,\alpha,\phi) = \exp\{\frac{y_{ijk}(\lambda_i + x_{ijk}^T\alpha) - b(\lambda_i + x_{ijk}^T\alpha)}{a(\phi)} + c(y_{ijk},\phi)\}, k = 1,\ldots,m_{ij}, \tag{1.1}$$

where α is a parameter vector of main effects, x_{ijk} are known constants, ϕ is a dispersion parameter, and the real-valued functions $a(\cdot), b(\cdot), c(\cdot)$ are known. Assume also that a parameter vector β relates some pairwise covariates z_{ijkl} to the marginal intracluster pairwise probability $f_{ijkl}(y_{ijk}, y_{ijl};\lambda_i,\alpha,\phi,\beta)$, $k \neq l$, where $\beta = \beta^0$, say, indicates pairwise independence. We aim to find a robust and flexible method to measure the pairwise associations between individual outcomes within clusters. Our primary interest is in the sparse data situation when the number of strata is much larger than the number of observations per stratum ($p \to \infty$, $N_i = \sum\limits_{j=1}^{n_i} m_{ij} < K < \infty$), although we also investigate the standard situation when the number of observations per stratum is much larger than the number of strata and cluster sizes are uniformly bounded ($p < \infty$, $n_i \to \infty$, $m_{ij} < L < \infty$, and $\frac{n_i}{N} \to r_i$, where $N = \sum\limits_{i=1}^{p} n_i$, and $0 < r_i < R < \infty$).

Our work is motivated by a study of gastrointestinal health effects due to the consumption of drinking water (Payment et al., 1997). This prospective study was conducted over 16 months. A total of 1339 households with young children, stratified by 138 small geographic areas, were randomly drawn from a population in Quebec served by a single water treatment plant and its distribution system (Figure 1.1). Households were randomly assigned into one of the four drinking water groups: unmodified tap water; tap water with a purge valve; bottled plant water; or purified bottled water. Counts of highly credible gastrointestinal illness (HCGI) episodes and covariates including sex and age of all participating household members were recorded. Previous study suggested a relative higher risk of gastrointestinal illnesses in tap water group (Payment et al., 1997). Researchers also desired to investigate potential unmeasured household-specific environmental effects on gastrointestinal illness episodes. Therefore, it is of scientific interest to assess: 1) the effects of covariates on the counts of HCGI; 2) whether HCGI episodes tended to aggregate within households, which suggests an unmeasured household-specific environmental effect.

In my dissertation, we first develop a general composite conditional likelihood for sparse clustered data and investigate its properties (Aim 1), and then we also develop and apply general odds ratio models for use in this composite conditional likelihood (Aim 2), and finally we discuss the selection of weights for each cluster to improve efficiency of estimation (Aim 3). In the next section of this chapter, we review the existing work on the composite conditional likelihood for sparse clustered binary data, the general odds ratio function and the choices of weights for composite likelihood. In the end of this chapter, we present the outline of the dissertation.

Table 1.1: Data Structure

stratum	cluster	observation
1	1	$Y_{111}, Y_{112}, \ldots, Y_{11m_{11}}$
	2	$Y_{121}, Y_{122}, \ldots, Y_{12m_{12}}$
	$\vdots$	$\vdots$
	n_1	$Y_{1n_11}, Y_{1n_12}, \ldots, Y_{1n_1m_{1n_1}}$
2	1	$Y_{211}, Y_{212}, \ldots, Y_{21m_{21}}$
	2	$Y_{221}, Y_{222}, \ldots, Y_{22m_{22}}$
	$\vdots$	$\vdots$
	n_2	$Y_{2n_21}, Y_{2n_22}, \ldots, Y_{2n_2m_{2n_2}}$
$\vdots$	$\vdots$	$\vdots$
p	1	$Y_{p11}, Y_{p12}, \ldots, Y_{p1m_{p1}}$
	2	$Y_{p21}, Y_{p22}, \ldots, Y_{p2m_{p2}}$
	$\vdots$	$\vdots$
	n_p	$Y_{pn_p1}, Y_{pn_p2}, \ldots, Y_{pn_pm_{pn_p}}$

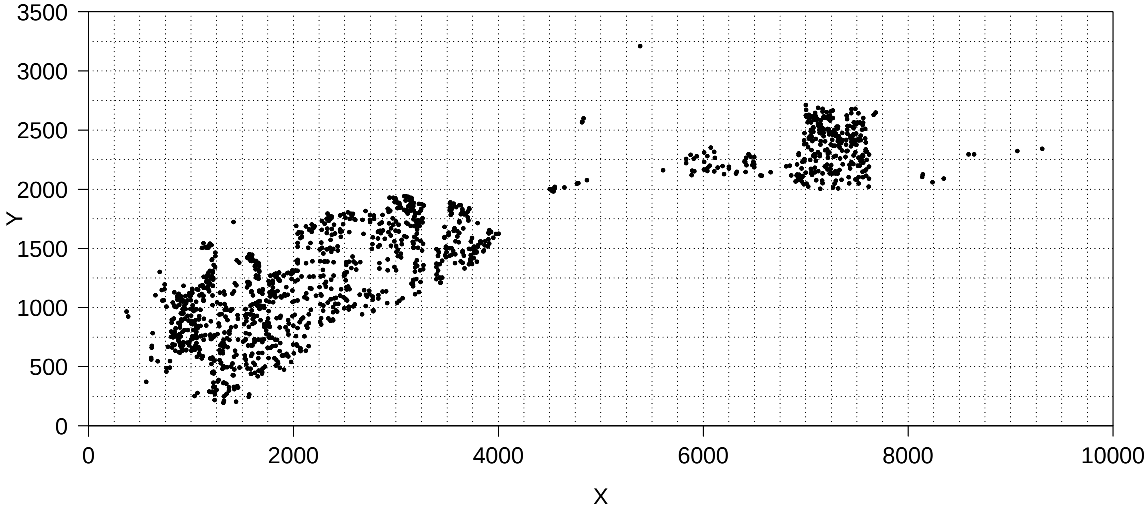

Figure 1.1: A map of households (units are in meters)

1.2 Background

1.2.1 Composite likelihood methods

In a general setting, let $L(\theta;\mathbf{Y})$ denote a parametric log-likelihood, where θ is an unknown parameter, and $\mathbf{Y}$ is a random vector. As discussed in Lindsay (1988), we write log-likelihoods $L_i(\beta)$ for a set of conditional or marginal events, for $i = 1, \ldots, n$, and construct the composite log-likelihood as:

$$CL(\theta) = \sum_{i=1}^{n} w_i L_i(\theta) \tag{1.2}$$

where w_i are weights chosen by the investigator, either on efficiency grounds or specified to be proportional to the inverse probability of ascertaining the cluster in studies where clusters have unequal probabilities of ascertainment. One could solves the equation $\frac{dCL(\beta)}{d\beta} = 0$, also called the composite score function, to obtain the estimator of θ.

Composite likelihood methods have several appealing properties: (1) composite likelihood can be a substitute for the likelihood when the maximum likelihood estimator is difficult to obtain; (2) like the score function, the composite score function is an unbiased estimating function; and (3) the resulting estimators can be consistent even when full maximum likelihood estimators are not, also called consistency robustness (Lindsay, 1988).

1.2.2 Composite conditional likelihood for sparse clustered binary data

As discussed in Section 1.1, challenges of conducting fully likelihood-based inference on sparse correlated data include modeling higher order associations within clusters, which often are not of primary scientific interest, and handling the stratum-specific nuisance parameters. By contrast, one approach that is insensitive to stratum-specific parameters and requires only assumptions about marginal densities and pairwise densities within clusters is a composite conditional likelihood approach (Hanfelt, 2004). In earlier work, Hanfelt (2004) focused on the special case where the individual outcomes are binary, i.e., the individual outcome $Y_{ijk} \sim \text{Bernoulli}(\mu_{ijk}), \mu_{ijk} = \text{expit}(\lambda_i + x_{ijk}^T\alpha)$, and the pairwise probability mass functions are given by $f_{ijkl}(y_{ijk}, y_{ijl}; \alpha, \beta)$, $k < l$. Here, α is a vector of main effects and β is a vector of intracluster pairwise associations.

Here, we briefly review this prior work. First, a composite conditional likelihood based on the pairwise conditional scheme is constructed for inference on α. Specifically, an independent pair drawn from different clusters within a stratum, Y_{ijk}, Y_{iuv}, $j \neq u$, are considered. For this pair, the sum $Y_{ijk} + Y_{iuv}$ is a complete sufficient statistic for the nuisance parameter λ_i when α is known. For binary data, $Y_{ijk} + Y_{iuv}$ can only be 0, 1, or 2, and $f(y_{ijk}|Y_{ijk} + Y_{iuv} = 0)$ and $f(y_{ijk}|Y_{ijk} + Y_{iuv} = 2)$ are degenerate. Therefore, a composite conditional likelihood for α is given by:

$$l^{(1)}(\alpha) = \sum_{i}^{(1)} l_i^{(1)}(\alpha) = \sum_{i=1}^{K} \sum_{j<u}^{n_i} w_{ij} w_{iu} \{ \sum_{k=1}^{m_{ij}} \sum_{v=1}^{m_{iu}} \Delta_{ijk,iuv} \log\, f(y_{ijk}|Y_{ijk} + Y_{iuv} = 1; \alpha)\}, \tag{1.3}$$

where $\Delta_{ijk,iuv} = 1$ if $Y_{ijk} + Y_{iuv} = 1$, or 0 otherwise. The weights w_{ij} are known for each cluster.

Next, a composite conditional likelihood for the pairwise association β is constructed based on a quadruplet consisting of two pairs of observations drawn from two different clusters within one stratum, $(Y_{ijk}, Y_{ijl}, Y_{iuv}, Y_{iuw})$. For binary data, the sum of the quadruplets, which yield non-degenerate conditional probability, can only be 1, 2, or 3. Moreover, the conditional probability is not very informative about β in the case of 'nearly concordant' quadruplets where the sum is 1 or 3. For this reason, we focus on quadruplets whose sum is 2. More specifically, the conditional probabilities conditioning on quadruplets of the form $Y_{ijk} = 1, Y_{iuv} = 1, Y_{ijl} + Y_{iuw} = 1$ or $Y_{ijk} = 0, Y_{iuv} = 0, Y_{ijl} + Y_{iuw} = 1$ are not very informative about the pairwise association, so only quadruplets of the form $Y_{ijk} = 1, Y_{iuv} = 0, Y_{ijl} + Y_{iuw} = 1$ are considered. Therefore, a composite likelihood for β is given by:

$$\begin{aligned} &l^{(2)}(\beta;\eta,\lambda) = \sum_{i=1}^{p} l_i^{(2)}(\beta;\eta,\lambda_i) \\ &= \sum_{i=1}^{p}\sum_{j<u}^{n_i} w_{ij}w_{iu}\{\sum_{k\neq l}^{m_{ij}}\sum_{v\neq w}^{m_{iu}} \Delta_{ijkl,iuvw} \log f(y_{ijl}|Y_{ijk}=1, Y_{iuv}=0, Y_{ijl}+Y_{iuw}=1;\beta,\alpha,\lambda_i)\}, \end{aligned} \tag{1.4}$$

where $\Delta_{ijkl,iuvw} = 1$ if $Y_{ijk} = 1, Y_{iuv} = 0, Y_{ijl} + Y_{iuw} = 1$, or 0 otherwise. In my dissertation, we consider a “discordant quadruplet” consisting of two pairs of observations drawn from two different clusters within one stratum, keeping two of the observations and the sum of the other two fixed, and generalize the composite conditional likelihood approach to accommodate responses of any type, such as counts or continuous data.

1.2.3 General odds ratio function

The general odds ratio function (Osius 2004; Chen 2003, 2004; Van Der Linde 2003) is defined as:

$$\psi_{ijkl}(y_{ijk}, y_{ijl}; \beta) = \frac{f_{ijkl}(y_{ijk}, y_{ijl}) f_{ijkl}(y^0_{ijk}, y^0_{ijl})}{f_{ijkl}(y_{ijk}, y^0_{ijl}) f_{ijkl}(y^0_{ijk}, y_{ijl})}, k \neq l, \qquad (1.5)$$

where f_{ijkl} refers to the pairwise density (or mass, in the discrete case) of (Y_{ijk}, Y_{ijl}), $k \neq l$, and the y^0_{ijk} are known reference values, typically $y^0_{ijk} = 0$. The general odds ratio function has several appealing features as a measure of pairwise association: it 1) accommodates responses of any type, such as counts or continuous data; 2) reduces to the usual odds ratio in the special case of binary data; 3) is invariant under prospective or retrospective proband-based sampling design; 4) is unconstrained by the marginal univariate distributions of the responses; and 5) completely characterizes the pairwise association, i.e. the information in the pairwise distribution that is not contained in the marginal univariate distributions (Osius, 2004).

When the cluster size is greater than 2, the general odds ratio function typically is restricted, but not severely so, by the joint distribution of the clustered observations (e.g. Liang et al., 1992). For multivariate normal or poisson data, the general odds ratio function is unrestricted.

One simple general odds ratio model is the bilinear log odds ratio model (Chen 2003, 2004), given by:

$$\log \psi(y_1, y_2; \beta) = \beta(y_1 - y^0_1) \times (y_2 - y^0_2), \qquad (1.6)$$

where y^0_1 and y^0_2 are reference points, and $\otimes$ would replace $\times$ for vector responses.

This model can be extended to accommodate pairwise covariates z_{12} (Chen, 2007) as:

$$\log\,\psi(y_1, y_2; z_{12}, \beta) = d(z_{12}, \beta)(y_1 - y_1^0) \times (y_2 - y_2^0),$$

where $d(\cdot)$ is a known function. Note that (1.6) makes a strong assumption that log odds ratios are linear in the responses. This assumption can be relaxed somewhat by assuming a transformed linear odds ratio model (Chen, 2007), given by:

$$\log\,\psi(y_1, y_2; \beta) = \beta\{G(y_1) - G(y_1^0)\}\{H(y_2) - H(y_2^0)\} \tag{1.7}$$

where $G(\cdot)$ and $H(\cdot)$ are known nonlinear functions, preferably monotonic, such as when $G(\cdot)$ and $H(\cdot)$ are cumulative distribution functions.

The general odds ratio model has certain advantages over other models of pairwise association. Alternatively, one could adopt a copula model (Van Ophem, 1999), which shares several of the appealing features of general odds ratio function in measuring pairwise association, but it is not invariant under different sampling designs and does not have an odds ratio interpretation. The correlation coefficient generally is inadequate as a measure of association, since it is not invariant when the sampling procedure depends on the marginal characteristics (Chen, 2007), and the range of the correlation parameter can be severely constrained by the marginal means (Lakshminarayana et al., 1999).

In my dissertation, we show that, under the composite conditional likelihood approach, a natural way to model pairwise association within clusters is via the general odds ratio function (Osius 2004; Chen 2003, 2004; Van Der Linde 2003). We also propose a specific odds ratio model for use in household aggregation studies, where all responses are measured on the same scale and an unmeasured exposure is hypothesized to affect responses within households in the same way, that we argue is more suitable than the bilinear log odds ratio model (1.6) and that not only rewards pairwise departure from the references points y_1^0 and y_2^0, but also penalizes lack of agreement.

1.3 Outline

This dissertation is organized as follows. In Chapter 2, we develop a general composite conditional likelihood and investigate its properties. We first present a composite conditional likelihood approach that allows the investigator to specify only the individual marginal densities and intracluster pairwise densities to conduct inference on sparse clustered data. Specifically, we adopt the approach in Hanfelt (2004) to construct a composite conditional likelihood that is free of stratum-specific nuisance parameters when estimating the effects of covariates on the marginal univariate responses and the dispersion parameter, $\eta = (\alpha, \phi)$, and we propose a general composite conditional likelihood, based on "discordant quadruplets", that is strongly insensitive to the stratum-specific nuisance parameters, to conduct inference on the intracluster pairwise associations β. We investigate the asymptotic properties of our composite conditional likelihood method under both the standard situation and the sparse data situation. Under the standard situation, we prove in detail the consistency and asymptotic normality for estimators of both η and β; we also investigate, in some specific examples, the asymptotic efficiency of the estimator of β and the sensitivity to the stratum-specific nuisance parameters. Under the sparse data situation, we present the consistency and asymptotic normality results for the estimate of η, and investigate the situation when consistency and asymptotic normality hold for the estimate of β for specific distributions.

In Chapter 3, we develop and apply general odds ratio models for use in composite conditional likelihood. We propose a specific general odds ratio model for use in household aggregation studies, that not only rewards pairwise departure from the references points but also penalizes lack of agreement. We demonstrate via simulation studies that the proposed model for use in composite conditional likelihood provides a valid and flexible way to obtain robust inference in studies of pairwise association with sparse clustered data by comparing the performance our approach versus GEE approach. We apply the method to investigate the marginal effects of covariates on HCGI and the aggregation of HCGI within households in the drinking water study.

In Chapter 4, we discuss the selection of weights for each cluster in the proposed composite conditional likelihood to improve efficiency of estimation. We investigate the optimal choice of cluster-specific weights via multivariate normal distributed data under both the standard situation and the sparse data situation. Specifically, we take advantage of Hajek projection to facilitate the comparison of estimators obtained by composite conditional likelihood and full likelihood, in situations where the latter is optimal, to identify a weight for each cluster that is at least approximately efficient. We demonstrate via simulation studies that the efficiency of estimation is improved. We reanalyze the water quality data after incorporating the proposed weights and compare the results analyzed under equal and unequal weights.

In Chapter 5, we summarize our current work and discuss the future direction of research.

Chapter 2

Develop A General Composite Conditional Likelihood and Investigate Its Properties

2.1 A General Composite Conditional Likelihood

Under the assumption that individual outcomes y_{ijk} marginally are distributed according to a generalized linear model (1.1), we follow the approach of Hanfelt (2004), and use the composite conditional log-likelihood below to render inferences about the main effects and dispersion parameter, i.e., $\eta = (\alpha^T, \phi)^T$, insensitive to the effects of nuisance parameters $\lambda = \{\lambda_i, i = 1, \ldots, p\}$:

$$l^{(1)}(\eta) = \sum_{i=1}^{p} l_i^{(1)}(\eta) = \sum_{i=1}^{p} \sum_{j<u}^{n_i} l_{iju}^{(1)}(\eta) = \sum_{i=1}^{p} \sum_{j<u}^{n_i} w_{ij} w_{iu} \{ \sum_{k=1}^{m_{ij}} \sum_{v=1}^{m_{iu}} \log\, f(y_{ijk}|Y_{ijk}+Y_{iuv};\eta)\}, \tag{2.1}$$

where $\{w_{ij}\}$ are known cluster-specific weights, which should be specified to be proportional to the inverse probability of ascertaining the cluster in studies where clusters have unequal probabilities of ascertainment (Hanfelt, 2004); and $f(y_{ijk}|Y_{ijk} + Y_{iuv} = a_{ijk,iuv};\eta)$ is

$$\frac{\exp\{a^{-1}(\phi) y_{ijk} (x_{ijk} - x_{iuv})^T \alpha + c(y_{ijk}, \phi) + c(a_{ijk,iuv} - y_{ijk}, \phi)\}}{\int \exp\{a^{-1}(\phi) s_{ijk} (x_{ijk} - x_{iuv})^T \alpha + c(s_{ijk}, \phi) + c(a_{ijk,iuv} - s_{ijk}, \phi)\}\, ds_{ijk}}. \tag{2.2}$$

In the case where the responses are discrete, we use summation rather than integration in the denominator of (2.2).

A greater challenge, and the focus of this chapter, is developing a composite conditional likelihood approach to conduct inference on the pairwise association parameter β. We note that under model (1.1) the collection of cluster-specific sums of responses, $\{Y_{i++} = \sum_{j=1}^{n_i} \sum_{k=1}^{m_{ij}} Y_{ijk}, i = 1, \ldots, p\}$, are locally ancillary statistics at $\beta = \beta^0$; and hence, one could consider conducting inference about pairwise association parameter β based on the following conditional log-likelihood:

$$l^{+}(\beta;\eta,\lambda) = \sum_{i=1}^{p} \log\, f\left((y_{i1}, \ldots, y_{in_i})|Y_{i++} = y_{i++};\beta,\eta,\lambda_i\right), \tag{2.3}$$

which is free of nuisance parameters $\lambda = (\lambda_1, \ldots, \lambda_p)$ locally at $\beta = \beta^0$. However, this approach requires a complete specification of the joint distribution of correlated outcomes in each cluster $y_{ij} = \{y_{ijk}, k = 1, \ldots, m_{ij}\}$.

As an alternative to model (2.3), we consider a "discordant quadruplet" consisting of two pairs of observations drawn from two different clusters within one stratum, keeping two of the observations and the sum of the other two fixed (Figure 2.1), denoted by

$$(Y_{ijk}, Y_{ijl}, Y_{iuv}, Y_{iuw})|(Y_{ijk} = y_{ijk}, Y_{iuv} = y_{iuv}, Y_{ijl}{+}Y_{iuw} = a_{ijl,iuw}), \text{ for } j \neq u, k \neq l, v \neq w.$$

We propose a composite log-likelihood for β consisting of a sum of component log-probabilities, based on a scheme of conditioning on the above discordant quadruplets:

$$\begin{aligned} l^{(2)}(\beta; \eta, \lambda) &= \sum_{i=1}^{p} l_i^{(2)}(\beta; \eta, \lambda_i) = \sum_{i=1}^{p} \sum_{j<u}^{n_i} l_{iju}^{(2)}(\beta; \eta, \lambda_i) \qquad (2.4) \\ = &\sum_{i=1}^{p} \sum_{j<u}^{n_i} w_{ij} w_{iu} \{ \sum_{k \neq l}^{m_{ij}} \sum_{v \neq w}^{m_{iu}} \log\, f(y_{ijl} | Y_{ijk} = y_{ijk}, Y_{iuv} = y_{iuv}, Y_{ijl} + Y_{iuw} = a_{ijl,iuw}; \beta, \eta, \lambda_i) \}. \end{aligned}$$

Importantly, $l^{(2)}(\beta; \eta, \lambda)$ is free of nuisance parameters λ when evaluated locally at $\beta = \beta^0$; this indicates that model (4.2) is strongly insensitive to stratum-specific effects λ, except when the intracluster pairwise assocations are quite strong, i.e., β is far from β^0. We will see in Section 3.2 some special cases when $l^{(2)}(\beta; \eta, \lambda)$ is free of λ for all β; moreover, in the next section, we show in an example that, even when $l^{(2)}(\beta; \eta, \lambda)$ depends on λ, it contains little information about λ.

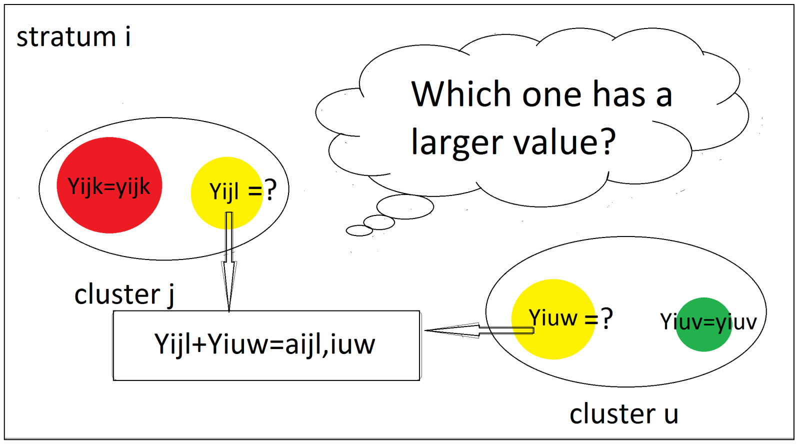

Figure 2.1: Discordant quadruplet: the red and green responses, and the sum of the two yellow responses, are considered fixed; the larger the circle is, the larger the value of observation

2.2 Asymptotic Results

We consider conducting joint inference on parameters of interest η and β in the presence of nuisance parameters λ by the following pair of composite conditional estimating functions:

$$G^{(1)}(\eta) = \frac{\partial l^{(1)}(\eta)}{\partial \eta} = \sum_{i=1}^{p} G_i^{(1)}(\eta) = \sum_{i=1}^{p} \sum_{j<u}^{n_i} G_{iju}^{(1)}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \eta) \tag{2.5}$$

$$G^{(2)}(\beta; \eta, \lambda) = \frac{\partial l^{(2)}(\beta; \eta, \lambda)}{\partial \beta} = \sum_{i=1}^{p} G_i^{(2)}(\beta; \eta, \lambda) = \sum_{i=1}^{p} \sum_{j<u}^{n_i} G_{iju}^{(2)}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \beta; \eta, \lambda_i). \tag{2.6}$$

We solve the equations $G^{(1)}(\eta) = 0$ and $G^{(2)}(\beta; \eta, \hat{\lambda}) = 0$ to obtain $\hat{\eta}$ and $\hat{\beta}$, where $\hat{\lambda}$ is obtained by maximizing the plug-in first-order unconditional composite likelihood, i.e.,

$$\hat{\lambda}_i = \arg\max_{\lambda_i} \sum_{j=1}^{n_i} \sum_{k=1}^{m_{ij}} \log f_{ijk}(y_{ijk}; \lambda_i, \hat{\eta}), (i = 1, \ldots, p). \tag{2.7}$$

We assume that $\{Y_{ij}, X_{ij}, m_{ij}, w_{ij}, j = 1, \ldots, n_i\}$ are independent and identically distributed for each stratum (Datta and Satten, 2008). We investigate the properties of $\hat{\eta}$ and $\hat{\beta}$ under two asymptotic schemes. In Section 2.2.1, we consider the standard asymptotic setting with a fixed number of strata, while the number of clusters are large, i.e., $p < \infty$, $n_i \to \infty$, and $\frac{n_i}{N} \to r_i$, where $N = \sum_{i=1}^{p} n_i$, and $0 < r_i < \infty$. In Section 2.2.2, we consider a sparse data asymptotic setting where the numbers of clusters are uniformaly bounded, while the number of strata is large, i.e., $p \to \infty$, $n_i < K < \infty$. Throughout, we assume uniformly bounded cluster sizes, i.e., $m_{ij} < L < \infty$.

2.2.1 Standard asymptotic results

The asymptotic results of $\hat{\eta}$ and $\hat{\beta}$ under the standard asymptotic setting ($p < \infty$, $n_i \to \infty$, and $\frac{n_i}{N} \to r_i$, where $N = \sum_{i=1}^{p} n_i$, and $0 < r_i < \infty$) are briefly sketched in the following proposition:

Proposition 1: Let the true parameter values of η, β, λ be $\eta_0, \beta_0, \lambda_0 = (\lambda_{i0}, i = 1, \ldots, p)$. Suppose conditions (a)-(g) in A1 in the Appendix of this chapter hold; then, $\hat{\eta}$ is strongly consistent and asymptotic normal, that is,

$$\hat{\eta} \xrightarrow{w.p.1} \eta_0, \sqrt{N}(\hat{\eta} - \eta_0) \xrightarrow{d} N(0, [A^{(1)}(\eta_0)]^{-1} B^{(1)}(\eta_0)\{[A^{(1)}(\eta_0)]^{-1}\}^T),$$

where $A^{(1)}(\eta_0) = \sum_{i=1}^{p} r_i^2 E[-G_{i12}^{(1)'}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \eta_0)]$, and $B^{(1)}(\eta_0) = \sum_{i=1}^{p} 4r_i^3 B_i^{(1)}(\eta_0)$, $B_i^{(1)}(\eta_0) = \text{var}\{E[G_{i12}^{(1)}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \eta_0)|\mathbf{Y}_{i1} = \mathbf{y}_{i1}]\}$.

Furthermore, suppose conditions (h)-(o) in A1 in the Appendix of this chapter hold; then, $\hat{\beta}$ is strongly consistent and asymptotic normal, that is,

$$\hat{\beta} \xrightarrow{w.p.1} \beta_0, \sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, [A^{(2)}(\beta_0; \eta_0, \lambda_0)]^{-1} B^{(2)}(\beta_0; \eta_0, \lambda_0)\{[A^{(2)}(\beta_0; \eta_0, \lambda_0)]^{-1}\}^T).$$

where $A^{(2)}(\beta_0; \eta_0, \lambda_0) = \sum_{i=1}^{p} r_i^2 E[-G_{i12}^{(2)'}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \beta_0; \eta_0, \lambda_0)]$, and $B^{(2)}(\beta_0; \eta_0, \lambda_0) = \sum_{i=1}^{p} 4r_i^3 B_i^{(2)}(\beta_0; \eta_0, \lambda_0)$, $B_i^{(2)}(\beta_0; \eta_0, \lambda_0) = \text{var}\{E[G_{i12}^{(*)}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \beta_0; \eta_0, \lambda_0)|\mathbf{Y}_{i1} = \mathbf{y}_{i1}]\}$.

The proof of Proposition 1 is provided in A2 in the Appendix of this chapter.

To briefly explore the asymptotic efficiency of our proposed approach, we consider a single stratum of bivariate normal data with marginal means $\lambda_1 + x_{1jk}^T \alpha$, marginal variance ϕ, and cluster sizes of 2. We assume homogeneous intracluster pairwise correlation, ρ, say, and arbitrarily apply the Fisher z-transformation,

$$\frac{1}{2}\log\ \{(1+\rho)/(1-\rho)\} = \beta, \tag{2.8}$$

so that $-\infty < \beta < \infty$. Under this scenario, or any other smooth transformation of ρ, the asymptotic relative efficiency of $\hat{\beta}$ versus the maximum likelihood estimator of β is given by ARE$(\hat{\beta}) = 1$; that is, $\hat{\beta}$ is fully efficient under the standard asymptotic setting. Moreover, under the homogeneous correlation model (2.8), $l^{(2)}(\beta; \eta, \lambda)$ is free of nuisance parameters λ.

Although $l^{(2)}(\beta; \eta, \lambda)$ depends on λ in general, we expect little sensitivity of $l^{(2)}(\beta; \eta, \lambda)$ to λ. To further examine the sensitivity of $l^{(2)}(\beta; \eta, \lambda)$ to λ, we consider a single stratum with cluster sizes of 2, consisting of bivariate normal data with marginal means $\lambda_1 + x_{1jk}^T \alpha$, marginal variances ϕ, and heterogeneous intracluster pairwise correlations $\rho_j \sim$ Unif$(0, 1)$, $j = 1, \ldots, n_1$. We compute the asymptotic relative efficiency of $\tilde{\lambda}$, where $\tilde{\lambda}$ is an estimator of λ obtained by maximizing $l^{(2)}(\beta; \eta, \lambda)$, versus the maximum likelihood estimator of λ. We find that ARE$(\tilde{\lambda})$ is close to 0 (≈ 0.003), confirming that $l^{(2)}(\beta; \eta, \lambda)$ contains relatively little information about nuisance parameters λ.

2.2.2 Results under sparse data situation

For the sparse data situation ($p \to \infty$, $n_i < K < \infty$), Lindeberg central limit theorem (Lindeberg, 1922) can be used to establish the consistency and asymptotic normality of $\hat{\eta}$; that is,

$$\hat{\eta} \xrightarrow{w.p.1} \eta_0, \sqrt{p}(\hat{\eta} - \eta_0) \xrightarrow{d} N(0, V_1) \tag{2.9}$$

where $V_1 = \lim_{p \to \infty} p[E(\frac{\partial G^{(1)}(\eta)}{\partial \eta}]^{-1} \text{var}(G^{(1)}(\eta))\{[E(\frac{\partial G^{(1)}(\eta)}{\partial \eta}]^{-1}\}^T$. Furthermore, under the assumption that $\lambda_1, \lambda_2, \ldots, \lambda_p$ are independent and identically distributed, we can show that $\hat{\beta}$ is strongly consistent and asymptotic normal (Lindeberg, 1922); that is,

$$\hat{\beta} \xrightarrow{w.p.1} \eta_0, \sqrt{p}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_2) \tag{2.10}$$

where $V_2 = \lim_{p\to\infty} p[E(\frac{\partial G^{(2)}(\beta)}{\partial\beta}]^{-1}\mathrm{var}(G^{(2)}(\beta))\{[E(\frac{\partial G^{(2)}(\beta)}{\partial\beta}]^{-1}\}^T$. When $\lambda_1, \ldots, \lambda_p$ are not independent and identically distributed, general results are lacking, although it is encouraging to note, as seen in the following examples, that under certain homogeneity conditions, $l^{(2)}(\beta; \eta, \lambda)$ is free of λ for all β, and thus the consistency and asymptotic normality of $\hat{\beta}$ is ensured.

Example 1. Let (Y_{ij1}, Y_{ij2}) follow a bivariate normal distribution with marginal means $\lambda_i + x_{ijk}^T\alpha$, $(k = 1, 2)$, and common marginal variance ϕ. Here, $l^{(2)}(\beta; \eta, \lambda)$ is free of λ when the intracluster pairwise correlations are homogeneous within strata, for any model $\rho_{ij12} = \mathrm{corr}(Y_{ij1}, Y_{ij2}) = \rho(\beta, z_i)$, where z_i denotes a pairwise covariate that is the same for all pairs of observations within clusters in stratum i.

Example 2. For correlated binary data, $l^{(2)}(\beta; \eta, \lambda)$ is free of λ under homogeneous marginal means and intracluster pairwise odds ratios within strata (Hanfelt, 2004).

Example 3. Let (Y_{ij1}, Y_{ij2}) follow a bivariate Poisson distribution (Madsen and Dalthorp, 2007) with marginal means $\exp(\lambda_i + x_{ijk}^T\alpha)$ $(k = 1, 2)$. Here, $l^{(2)}(\beta; \eta, \lambda)$ is free of λ under homogeneous means and intracluster pairwise correlations within strata, for any model $\rho_{ij12} = \mathrm{corr}(Y_{ij1}, Y_{ij2}) = \rho(\beta, z_i)$.

The above examples suggest that, in the challenging situation of sparse data with $\lambda_1, \ldots, \lambda_p$ not independent and identically distributed, $l^{(2)}(\beta; \eta, \lambda)$ generally is insensitive to nuisance parameters λ, and provides valid inference on β, provided that either β is not so far from β^0 or the covariates, x_{ijk} and z_{ijkl}, are not too heterogeneous within strata.

2.3 Appendix

A1. Regularity conditions for Proposition 1:

(a) there is an integrable and symmetric kernel g_1 such that; for all $\eta \in \Theta_1$, where Θ_1 is a compact topological space, $|G^{(1)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \eta)| < g_1(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$ and $(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}) \in \mathbf{R}^2$;

(b) there is a sequence $S^2_{M_1}$ of measurable sets such that $P(\mathbf{R}^2 - \bigcup\limits_{M_1=1}^{\infty} S^2_{M_1}) = 0$; for each M_1, $E[G^{(1)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \eta)|\mathbf{Y}_{ij} = \mathbf{y}_{ij}]$ is equicontinuous in η for $\mathbf{Y}_{ij} \in S^1_{M_1}$, and $E[G^{(1)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \eta)]$ is equicontinuous in η for $(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}) \in S^2_{M_1}$, where $S^2_{M_1} = S^1_{M_1} \times S^1_{M_1}$;

(c) $\sum\limits_{i=1}^{p} E_{\eta_0}[G^{(1)}_i(\eta)]$ exists for all $\eta \in \Theta_1$, $\sum\limits_{i=1}^{p} E_{\eta_0}[G^{(1)}_i(\eta_0)] = 0$, and $\sum\limits_{i=1}^{p} E_{\eta_0}[G^{(1)}_i(\eta)] \neq 0$ for all $\eta \neq \eta_0$ in Θ_1;

(d) $G^{(1)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \eta)$ and its first two partial derivatives with respect to η exist for all η in a neighborhood of η_0;

(e) for each η in a neighborhood of η_0, there exists a function $h_1(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$ with finite expectation, such that $|\frac{\partial^2 G^{(1)}_{ijk}(\mathbf{Y}_{ij}, \mathbf{Y}_{ik}, \eta)}{\partial \eta^2}| \leq h_1(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$ for all $(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$;

(f) $E[-G^{(1)'}_{i12}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \eta_0)]$ exists and is nonsingular;

(g) $B^{(1)}_i(\eta_0) = \text{var}\{E[G^{(1)}_{i12}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \eta_0)|\mathbf{Y}_{i1} = \mathbf{y}_{i1}]\}$ exists and is finite;

(h) there is an integrable and symmetric kernel g_2 such that; for all $\beta \in \Theta_2$, where Θ_2 is a compact topological space, $|G^{(2)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \beta; \eta_0, \lambda_{i0})| < g_2(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$ and $(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}) \in \mathbf{R}^2$;

(i) there is a sequence $S^2_{M_2}$ of measurable sets such that $P(\mathbf{R}^2 - \bigcup\limits_{M_2=1}^{\infty} S^2_{M_2}) = 0$; for each M_2, $E[G^{(2)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \beta; \eta_0, \lambda_{i0})|\mathbf{Y}_{ij} = \mathbf{y}_{ij}]$ is equicontinuous in β for $\mathbf{Y}_{ij} \in$

$S^1_{M_2}$, and $E[G^{(2)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \beta; \eta_0, \lambda_{i0})]$ is equicontinuous in β for $(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}) \in S^2_{M_2}$, where $S^2_{M_2} = S^1_{M_2} \times S^1_{M_2}$;

(j) $\sum_{i=1}^{p} E_{(\beta_0,\eta_0,\lambda_{i0})}[G^{(2)}_i(\beta; \eta, \lambda_i)]$ exists for all $\eta \in \Theta_1$ and $\beta \in \Theta_2$, $\sum_{i=1}^{p} E_{(\beta_0,\eta_0,\lambda_{i0})}[G^{(2)}_i(\beta_0; \eta_0, \lambda_{i0})] = 0$, and $\sum_{i=1}^{p} E_{(\beta_0,\eta_0,\lambda_{i0})}[G^{(2)}_i(\beta; \eta, \lambda_i)] \neq 0$ for all $\eta \neq \eta_0$ in Θ_1, $\beta \neq \beta_0$ in $\in \Theta_2$;

(k) $G^{(*)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \beta; \eta, \lambda_i)$ (defined in A2) and its first two partial derivatives with respect to β exist for all β in a neighborhood of β_0;

(l) for each β in a neighborhood of β_0, there exists a function $h_2(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$ with finite expectation, such that $|\frac{\partial^2 G^{(*)}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \beta; \eta, \lambda_i)}{\partial \beta^2}| \leq h_2(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$ for all $(\mathbf{Y}_{ij}, \mathbf{Y}_{iu})$;

(m) $E[-G^{(*)'}_{iju}(\mathbf{Y}_{ij}, \mathbf{Y}_{iu}, \beta_0; \eta_0, \lambda_{i0})]$ exists and is nonsingular;

(n) $B^{(2)}_i(\beta_0; \eta_0, \lambda_0) = \text{var}\{E[G^{(*)}_{i12}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \beta_0; \eta_0, \lambda_0)|\mathbf{Y}_{i1} = \mathbf{y}_{i1}]\}$ exists and is finite;

(o) $\hat{\lambda}$ is strongly consistent and $\sqrt{N}-$consistent, i.e., $\hat{\lambda}_i \xrightarrow{w.p.1} \lambda_i$ and $\hat{\lambda}_i - \lambda_i = \mathcal{O}_p(N^{-1/2})$.

Condition (a)-(b) and (h)-(i) are regularity conditions in Theorem 1 given by Yeo and Johnson (2001) for almost sure uniform convergence of $G^{(1)}_i(\eta)$ and $G^{(2)}_i(\beta; \eta_0, \lambda_{i0})$, $i = 1, \ldots, p$.

A2. Proof of Proposition 1:

Let $\hat{\eta}$ be the root of $\sum_{i=1}^{p} G^{(1)}_i(\eta) = 0$. If $\hat{\eta}$ is not consistent, there exists a subsequence of $\hat{\eta}$, $\{\hat{\eta}_k\}$ such that $\hat{\eta}_k \xrightarrow{w.p.1} \eta_1 \neq \eta_0$. Then

$$\begin{aligned}|\sum_{i=1}^{p} G_i^{(1)}(\hat{\eta}_k) - \sum_{i=1}^{p} E_{\eta_0}[G_i^{(1)}(\eta_1)]| &\leq |\sum_{i=1}^{p} G_i^{(1)}(\hat{\eta}_k) - \sum_{i=1}^{p} E_{\eta_0}[G_i^{(1)}(\hat{\eta}_k)]| \\ &+ |\sum_{i=1}^{p} E_{\eta_0}[G_i^{(1)}(\hat{\eta}_k)] - \sum_{i=1}^{p} E_{\eta_0}[G_i^{(1)}(\eta_1)]|\end{aligned}$$

The first term on the right hand side converges to 0 by uniform convergence of $\sum_{i=1}^{p} G_i^{(1)}(\eta)$. The second term converges to 0 by continuity of $\sum_{i=1}^{p} E_{\eta_0} G_i^{(1)}$. Thus $\sum_{i=1}^{p} G_i^{(1)}(\hat{\eta}_k) \xrightarrow{w.p.1} \sum_{i=1}^{p} E_{\eta_0}[G_i^{(1)}(\eta_1)] \neq 0$, and this contradicts $\sum_{i=1}^{p} G_i^{(1)}(\hat{\eta}_k) = 0$. Therefore, $\hat{\eta} \xrightarrow{w.p.1} \eta_0$.

$$\begin{aligned}0 = \sum_{i=1}^{p} G_i^{(1)}(\hat{\eta}) &= \sum_{i=1}^{p} [G_i^{(1)}(\eta_0) + G_i^{(1)'}(\eta_0)(\hat{\eta} - \eta_0) + G_i^{(1)''}(\eta^*)(\hat{\eta} - \eta_0)^2/2] \\ &= \sum_{i=1}^{p} [G_i^{(1)}(\eta_0) + (\hat{\eta} - \eta_0)\{G_i^{(1)'}(\eta_0) + G_i^{(1)''}(\eta^*)(\hat{\eta} - \eta_0)/2\}] \\ \Longrightarrow \sqrt{N}(\hat{\eta} - \eta_0) &= \{\sum_{i=1}^{p} -G_i^{(1)'}(\eta_0) - \sum_{i=1}^{p} G_i^{(1)''}(\eta^*)(\hat{\eta} - \eta_0)/2\}^{-1} \sqrt{N} \sum_{i=1}^{p} G_i^{(1)}(\eta_0)\end{aligned}$$

where $G_i^{(1)'}(\eta_0) = \frac{\partial G_i^{(1)}(\eta)}{\partial \eta}\Big|_{\eta=\eta_0}$, $G_i^{(1)''}(\eta^*) = \frac{\partial^2 G_i^{(1)}(\eta)}{\partial \eta^2}\Big|_{\eta=\eta^*}$.

By U-statistics theory (Serfling, 2009),

$$\begin{aligned}&\sqrt{n_i}\frac{2}{n_i(n_i-1)} G_i^{(1)}(\eta_0) \xrightarrow{d} N(0, 4B_i^{(1)}(\eta_0)) \\ \Longrightarrow &\sqrt{N}\frac{2}{N^2} G_i^{(1)}(\eta_0) \xrightarrow{d} N(0, 4r_i^3 B_i^{(1)}(\eta_0)) \\ \Longrightarrow &\sqrt{N}\sum_{i=1}^{p}\frac{2}{N^2} G_i^{(1)}(\eta_0) \xrightarrow{d} N(0, \sum_{i=1}^{p} 4r_i^3 B_i^{(1)}(\eta_0))\end{aligned}$$

By U-statistics theory (Serfling, 2009),

$$\frac{2}{n_i(n_i-1)}G_i^{(1)'}(\eta_0) \xrightarrow{P} E[-\frac{\partial G_{i12}^{(1)}(\mathbf{Y}_{i1},\mathbf{Y}_{i2},\eta)}{\partial\eta}\bigg|_{\eta=\eta_0}]$$

$$\Longrightarrow \sum_{i=1}^{p}\frac{2}{N^2}G_i^{(1)'}(\eta_0) \xrightarrow{P} \sum_{i=1}^{p} r_i^2 E[-\frac{\partial G_{i12}^{(1)}(\mathbf{Y}_{i1},\mathbf{Y}_{i2},\eta)}{\partial\eta}\bigg|_{\eta=\eta_0}] = A^{(1)}(\eta_0)$$

Since $\hat{\eta} \xrightarrow{P} \eta_0$ and by assumption $\frac{2}{N^2}\sum_{i=1}^{p} G_i^{(1)''}(\eta^*) = O_p(1)$, we have $\frac{2}{N^2}\sum_{i=1}^{p} G_i^{(1)''}(\eta^*)(\hat{\eta}-\eta_0)/2 \xrightarrow{P} 0$. By Slutsky's theorem, $\sqrt{N}(\hat{\eta}-\eta_0) \xrightarrow{d} N(0,[A^{(1)}(\eta_0)]^{-1}B^{(1)}(\eta_0)\{[A^{(1)}(\eta_0)]^{-1}\}^T)$.

Let $\hat{\beta}$ be the root of $\sum_{i=1}^{p} G_i^{(2)}(\beta;\hat{\eta},\hat{\lambda}_i) = 0$. If $\hat{\beta}$ is not consistent, there exists a subsequence of $\hat{\beta}$, $\{\hat{\beta}_k\}$ such that $\hat{\beta}_k \xrightarrow{w.p.1} \beta_1 \neq \beta_0$. Then

$$\begin{aligned}
&|\sum_{i=1}^{p} G_i^{(2)}(\hat{\beta}_k;\hat{\eta},\hat{\lambda}_i) - \sum_{i=1}^{p} E_{\beta_0,\eta_0,\lambda_{i0}}[G_i^{(2)}(\beta_1;\eta_0,\lambda_{i0})]| \\
\leq& |\sum_{i=1}^{p} G_i^{(2)}(\hat{\beta}_k;\hat{\eta},\hat{\lambda}_i) - \sum_{i=1}^{p} G_i^{(2)}(\hat{\beta}_k;\eta_0,\lambda_{i0})| + |\sum_{i=1}^{p} G_i^{(2)}(\hat{\beta}_k;\eta_0,\lambda_{i0}) - \sum_{i=1}^{p} E_{\beta_0,\eta_0,\lambda_{i0}}[G_i^{(2)}(\hat{\beta}_k;\eta_0,\lambda_{i0})]| \\
+& |\sum_{i=1}^{p} E_{\beta_0,\eta_0,\lambda_{i0}}[G_i^{(2)}(\hat{\beta}_k;\eta_0,\lambda_{i0})] - \sum_{i=1}^{p} E_{\beta_0,\eta_0,\lambda_{i0}}[G_i^{(2)}(\beta_1;\eta_0,\lambda_{i0})]|
\end{aligned}$$

The first term converges to 0 by continuity of $\sum_{i=1}^{p} G_i^{(2)}$ and the strongly consistent of $\hat{\eta}$ and $\hat{\lambda}$. The second term on the right hand side converges to 0 by uniform convergence of $\sum_{i=1}^{p} G_i^{(2)}(\beta;\eta_0,\lambda_{i0})$. The third term converges to 0 by continuity of $\sum_{i=1}^{p} E_{\beta_0,\eta_0,\lambda_{i0}} G_i^{(2)}$. Thus $\sum_{i=1}^{p} G_i^{(2)}(\hat{\beta}_k;\hat{\eta},\hat{\lambda}_i) \xrightarrow{w.p.1} \sum_{i=1}^{p} E_{\beta_0,\eta_0,\lambda_{i0}}[G_i^{(2)}(\beta_1;\eta_0,\lambda_{i0})] \neq 0$, and this contradicts $\sum_{i=1}^{p} G_i^{(2)}(\hat{\beta}_k;\hat{\eta},\hat{\lambda}_i) = 0$. Therefore, $\hat{\beta} \xrightarrow{w.p.1} \beta_0$.

Let $G_i^{(3)}(\lambda_i)$, $i=1,\ldots,p$, are estimating functions which yield $\hat{\lambda}$ in condition (d) of Proposition 1. Define $\theta_i = (\eta,\lambda_i)$, for $i=1,\ldots,p$. To prove the asymptotic normality of $\hat{\beta}$, we first define a projected estimating function $G_i^*(\beta, \theta_i)$ which is uncorrelated with (η, λ_i) as:

$$\begin{aligned} &G_i^*(\beta, \theta_i) = G_i^{(2)}(\beta, \theta_i) \\ &- \begin{pmatrix} E[G_i^{(1)}(\eta) G_i^{(2)}(\beta, \theta_i)] \\ E[G_i^{(3)}(\lambda_i) G_i^{(2)}(\beta, \theta_i)] \end{pmatrix}^T \begin{pmatrix} E\{[G_i^{(1)}(\eta)]^2\} & E[G_i^{(1)}(\eta) G_i^{(3)}(\lambda_i)] \\ E[G_i^{(3)}(\lambda_i) G_i^{(1)}(\eta)] & E\{[G_i^{(3)}(\lambda_i)]^2\} \end{pmatrix}^{-1} \begin{pmatrix} G_i^{(1)}(\eta) \\ G_i^{(3)}(\lambda_i) \end{pmatrix}, \end{aligned}$$

where $G_i^*(\beta, \theta_i)$ is a U-statistics, and $G_i^*(\hat{\beta}, \hat{\theta}_i) = 0$.

$$0 = \sum_{i=1}^{p} G_i^*(\hat{\beta}, \hat{\theta}_i) = \sum_{i=1}^{p} [G_i^*(\hat{\beta}, \theta_{i0}) + G_{i\theta_i}^{*'}(\hat{\beta}, \theta_{i0})(\hat{\theta}_i - \theta_{i0}) + G_{i\theta_i}^{*''}(\hat{\beta}, \theta_i^*)(\hat{\theta}_i - \theta_{i0})^2/2] \tag{2.11}$$

where θ_i^* is a point between θ_{i0} and $\hat{\theta}_i$, $G_{i\theta_i}^{*'}(\hat{\beta}, \theta_{i0}) = \frac{\partial G_i^*(\hat{\beta}, \theta_i)}{\partial \theta_i}\Big|_{\theta_i = \theta_{i0}}$, $G_{i\theta_i}^{*''}(\hat{\beta}, \theta_i^*) = \frac{\partial^2 G_i^*(\hat{\beta}, \theta_i)}{\partial \theta_i^2}\Big|_{\theta_i = \theta_i^*}$. By U-statistics theory (Serfling, 2009), we have $(\hat{\theta}_i - \theta_{i0}) = O_p(n_i^{-1/2})$, and $(\hat{\theta}_i - \theta_{i0})^2 = O_p(n_i^{-1})$.

$$G_{i\theta_i}^{*'}(\hat{\beta}, \theta_{i0}) = \frac{\partial G_i^*(\beta_0, \theta_i)}{\partial \theta_i}\Big|_{\theta_i = \theta_{i0}} + \frac{\partial^2 G_i^*(\beta, \theta_i)}{\partial \theta_i \partial \beta}\Big|_{\theta_i = \theta_{i0}, \beta = \beta^*}(\hat{\beta} - \beta_0), \tag{2.12}$$

where β^* is a point between β_0 and $\hat{\beta}$. Since $E[\frac{\partial G_i^*(\beta_0, \theta_i)}{\partial \theta_i}\Big|_{\theta_i = \theta_{i0}}] = 0$ and $\hat{\beta} \xrightarrow{P} \beta_0$, we have $G_{i\theta_i}^{*'}(\hat{\beta}, \theta_{i0}) = o_p(n_i^2)$. By assumption, $G_i^{*''}(\hat{\beta}, \theta_i^*) = O_p(n_i^2)$.

Next, we express (2.11) as:

$$\begin{aligned} 0 &= \sum_{i=1}^{p} [G_i^*(\hat{\beta}, \theta_{i0}) + o_p(n_i^{3/2})] \\ &= \sum_{i=1}^{p} [G_i^*(\beta_0, \theta_{i0}) + G_{i\beta}^{*'}(\beta_0, \theta_{i0})(\hat{\beta} - \beta_0) + G_{i\beta}^{*''}(\beta^*, \theta_{i0})(\hat{\beta} - \beta_0)^2/2 + o_p(n_i^{3/2})] \\ &= \sum_{i=1}^{p} \{G_i^*(\beta_0, \theta_{i0}) + (\hat{\beta} - \beta_0)[G_{i\beta}^{*'}(\beta_0, \theta_{i0}) + G_{i\beta}^{*''}(\beta^*, \theta_{i0})(\hat{\beta} - \beta_0)/2] + o_p(n_i^{3/2})\} \end{aligned}$$

$$\implies \sqrt{N}(\hat{\beta}-\beta_0) = \{\sum_{i=1}^{p} -G^{*'}_{i\beta}(\beta_0,\theta_{i0}) - \sum_{i=1}^{p} G^{*''}_{i\beta}(\beta^*,\theta_{i0})(\hat{\beta}-\beta_0)/2\}^{-1}\{\sqrt{N}\sum_{i=1}^{p} G^*_i(\beta_0,\theta_{i0}) + o(N^2)\}, \tag{2.13}$$

where $G^{*'}_{i\beta}(\beta_0,\theta_{i0}) = \frac{\partial G^*_i(\beta,\theta_{i0})}{\partial\beta}\Big|_{\beta=\beta_0}$, $G^{*''}_{i\beta}(\beta_0,\theta_{i0}) = \frac{\partial^2 G^*_i(\beta,\theta_{i0})}{\partial\beta^2}\Big|_{\beta=\beta^*}$ By U-statistics theory (Serfling, 2009),

$$\sqrt{n_i}\frac{2}{n_i(n_i-1)}G^*_i(\beta_0,\theta_{i0}) \xrightarrow{d} N(0, 4B^{(2)}_i(\beta_0;\eta_0,\lambda_0)), \tag{2.14}$$

where G^*_{ijl} is the kernal of the U-statistics G^*_i.

$$\begin{aligned}
&\implies \sqrt{N}\frac{2}{N^2}G^*_i(\beta_0,\theta_{i0}) \xrightarrow{d} N(0, 4r_iB^{(2)}_i(\beta_0;\eta_0,\lambda_0))\\
&\implies \sqrt{N}\sum_{i=1}^{p}\frac{2}{N^2}G^*_i(\beta_0,\theta_{i0}) \xrightarrow{d} N(0, \sum_{i=1}^{p} 4r_i^3B^{(2)}_i(\beta_0;\eta_0,\lambda_0))\\
&\implies \sqrt{N}\sum_{i=1}^{p}\frac{2}{N^2}G^*_i(\beta_0,\theta_{i0}) + o_p(1) \xrightarrow{d} N(0, \sum_{i=1}^{p} 4r_i^3B^{(2)}_i(\beta_0;\eta_0,\lambda_0))
\end{aligned}$$

By U-statistics theory (Serfling, 2009),

$$\begin{aligned}
&\frac{2}{n_i(n_i-1)} - G^{*'}_{i\beta}(\beta_0,\theta_{i0}) \xrightarrow{P} E[-\frac{\partial G^*_{i12}(\mathbf{Y}_{i1},\mathbf{Y}_{i2},\beta,\theta_{i0})}{\partial\beta}\Big|_{\beta=\beta_0}]\\
\implies &\sum_{i=1}^{p}\frac{2}{N^2} - G^{*'}_{i\beta}(\beta_0,\theta_{i0}) \xrightarrow{P} \sum_{i=1}^{p} r_i^2E[-\frac{\partial G^*_{i12}(\mathbf{Y}_{i1},\mathbf{Y}_{i2},\beta,\theta_{i0})}{\partial\beta}\Big|_{\beta=\beta_0}] = A^{(2)}(\beta_0;\eta_0,\lambda_0)
\end{aligned}$$

Since $\hat{\beta} \xrightarrow{P} \beta$ and by assumption $\frac{2}{N^2}\sum_{i=1}^{p} G^{*''}_{i\beta}(\beta^*,\theta_{i0}) = O_p(1)$, we have $\frac{2}{N^2}\sum_{i=1}^{p} G^{*''}_{i\beta}(\beta^*,\theta_{i0})(\hat{\beta}-\beta_0)/2 \xrightarrow{P} 0$. By Slutsky's theorem,

$$\sqrt{N}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, [A^{(2)}(\beta_0;\eta_0,\lambda_0)]^{-1}B^{(2)}(\beta_0;\eta_0,\lambda_0)\{[A^{(2)}(\beta_0;\eta_0,\lambda_0)]^{-1}\}^T).$$

Chapter 3

Develop and Apply General Odds Ratio Models for Use in Composite Conditional Likelihood

3.1 General Odds Ratio Models

3.1.1 Model intracluster pairwise association via odds ratio function under the composite conditional likelihood approach

While the proposed composite conditional likelihood (4.2) could be written in terms of the pairwise density (or mass) functions $f_{ijkl}(\cdots,\cdots;\lambda_i,\eta,\beta)$, such an approach would not be desirable, since in general we lack probability models for pairwise data that are both flexible and computationally convenient, especially for discrete data. As an alternative to specifying the pairwise probabilities $f_{ijkl}(\cdots,\cdots;\lambda_i,\eta,\beta)$ directly, in this chapter we pursue a more flexible and convenient approach, in which we indirectly specify the pairwise probabilities, via the marginal univariate probability (1.1) and a general odds ratio function.

Under the "discordant quadruplet" conditional scheme adopted in (4.2), we can write the relevant conditional probabilities in terms of the general odds ratio function $\psi_{ijkl}(\cdot,\cdot;\beta)$ as

$$\begin{aligned} &f(y_{ijl}|Y_{ijk}=y_{ijk},Y_{iuv}=y_{iuv},Y_{ijl}+Y_{iuw}=a;\beta,\eta,\lambda_i) = \\ &\qquad \frac{\psi_{ijkl}(y_{ijk},y_{ijl};\beta)\psi_{iuvw}(y_{iuv},a-y_{ijl};\beta)}{\int \psi_{ijkl}(y_{ijk},s_{ijl};\beta)\psi_{iuvw}(y_{iuv},a-s_{ijl};\beta)C_{ijkl,iuvw}(s_{ijl},y_{ijl},a)\,ds_{ijl}}, \end{aligned} \tag{3.1}$$

where, in the denominator, summation would replace integration for discrete responses. Note that the only term that depends on nuisance parameter λ_i in (3.1) is a type of between-cluster contrast given by

$$C_{ijkl,iuvw}(s_{ijl},y_{ijl},a) = \frac{f_{ijkl}(y^0_{ijk},s_{ijl})f_{iuvw}(y^0_{iuv},a-s_{ijl})}{f_{ijkl}(y^0_{ijk},y_{ijl})f_{iuvw}(y^0_{iuv},a-y_{ijl})}, j<u, k\neq l, v\neq w. \tag{3.2}$$

Under pairwise independence within clusters and adoption of marginal univariate model (1.1), it follows that $C_{ijkl,iuvw}(s_{ijl}, y_{ijl}, a)$ would reduce to the following term, which is free of λ_i:

$$\begin{aligned} C_{ijkl,iuvw}(s_{ijl}, y_{ijl}, a) = \exp\{&(s_{ijl} - y_{ijl})(x_{ijl} - x_{iuw})^T \alpha / a(\phi) + c(s_{ijl}, \phi) \\ &+ c(a - s_{ijl}, \phi) - c(y_{ijl}, \phi) - c(a - y_{ijl}, \phi)\}. \end{aligned}$$

Owing to the fact that the odds ratio function and marginal univariate distributions uniquely determine the pairwise distribution (Osius, 2004), $C_{ijkl,iuvw}(s_{ijl}, y_{ijl}, a)$ is uniquely determined by the odds ratio function and marginal univariate distributions. Furthermore, $C_{ijkl,iuvw}(s_{ijl}, y_{ijl}, a)$ can be computed based on an iterative procedure called "marginal fittings" which was proved by Osius (2004) to converge:

1) start with any initial guess of the pairwise density function $f_{Y_1Y_2}$ with ψ_{Y_1,Y_2} as its odds ratio which satisfies the existence conditions in Osius (2004);

2) adjust $f_{Y_1Y_2}$'s marginal distribution of Y_1 to the pre-specified marginal f_{Y_1}, then adjust $f_{Y_1Y_2}$'s marginal distribution of Y_2 to the pre-specified marginal f_{Y_2};

3) repeat 2) till convergence;

4) calculate $C_{ijkl,iuvw}(s_{ijl}, y_{ijl}, a)$ by (3.2).

After each step of "marginal fittings", the corresponding odds ratio remains equivalent to ψ_{Y_1,Y_2}. Refer to A1 in the Appendix in this chapter for details in existence conditions in Osius (2004), the choice of the initial density $f^{(0)}_{Y_1Y_2}$ and adjusting $f^{(0)}_{Y_1Y_2}$'s marginal distribution.

3.1.2 Develop a odds ratio model for use in composite conditional likelihood

We reviewed some general odds ratio models in Chapter 1. In the drinking water study, where all responses y_{ijk} are measured on the same scale and we desire β to be regarded as a measure of household aggregation of illness episodes, it seems desirable that the measurement of pairwise association not only rewards pairwise departure from the references points y_1^0 and y_2^0, but also penalizes lack of agreement; unfortunately, neither Model (1.6) nor Model (1.7) accomodates the latter property. We propose an alternative general odds ratio model that avoids this limitation as:

$$\log \psi(y_1, y_2; \beta) = \frac{\beta(y_1 - y_1^0) \times (y_2 - y_2^0)}{1 + 2(y_1 - y_2)^2}, \tag{3.3}$$

which can be extended to accommodate pairwise covariates z_{12} as:

$$\log \psi(y_1, y_2; z_{12}, \beta) = \frac{d(z_{12}, \beta)(y_1 - y_1^0) \times (y_2 - y_2^0)}{1 + 2(y_1 - y_2)^2}. \tag{3.4}$$

Model (3.3) is symmetric in terms of y_1 and y_2, linear in β, rewards pairwise departure from the references points y_1^0 and y_1^0 in the same direction, penalizes lack of agreement by adding the term $2(y_1 - y_2)^2$ in the denominator, and reduces to the usual odds ratio function in the case of binary data. As shown in Table 3.1, both model (1.6) (bilinear model) and model (3.3) (proposed model) reward pairwise departure from the reference points, e.g., $\log \psi(10, 10; \beta) > \log \psi(1, 1; \beta)$; however, the bilinear model (1.6) does not penalize lack of agreement, so that it assigns the equality $\log \psi(1, 100; \beta) = \log \psi(5, 20; \beta) = \log \psi(10, 10; \beta)$, which would exaggerate the extent of familial aggregation of disease, and thus seems less satisfactory than the proposed odds ratio model (3.3) in the water quality study.

We show the rational for using $2(y_1 - y_2)^2$ in the denominator rather than $(y_1 - y_2)^2$ (model 1) or $2|y_1 - y_2|$ (model 2) via a simple comparison. As shown in Table 3.2, neither model 1 nor model 2 penalize lack of agreement in certain circumstances, e.g., under model 1, $\log \psi(1, 1; \beta) = \log \psi(1, 2; \beta)$, under model 2, $\log \psi(3, 3; \beta) < \log \psi(10, 20; \beta)$, $\log \psi(2, 3; \beta) < \log \psi(5, 20; \beta)$, which are less satisfactory than the proposed model (3.3) in the water quality study.

To briefly assess the number of iterations of "marginal fittings" needed for convergence when using model (3.3), we let (Y_{ij1}, Y_{ij2}) follow a bivariate Poisson distribution with marginal means $\exp(\lambda_i + x_{ijk}^T \alpha)$ $(k = 1, 2)$, $\alpha = 0.5$, $\lambda_i \sim \text{Unif}(-1, 0)$, $w_{ij} = 1$, and we used model (3.3), with four different values of β, to model the intracluster pairwise associations. We obtained $\hat{\beta}$ by solving the equation $G^{(2)}(\beta; \eta, \hat{\lambda}) = 0$, and we used the "marginal fittings" method to evaluate $C_{ijkl,iuvw}$. We started with the initial pairwise density function $f(y_1, y_2) = \exp(\beta y_1 y_2 / (1 + 2(y_1 - y_2)^2))$, which satisfied the existence conditions in Osius (2004). The iterative estimates of β, shown in Figure 3.1, indicated that five steps of "marginal fittings" were sufficient for convergence for all values of β.

Table 3.1: Log general odds ratios under the bilinear model and the proposed model with reference values $y_1^0 = y_2^0 = 0$

y1	y2	bilinear model	proposed model
1	1	β	β
1	2	2β	0.667β
1	3	3β	0.333β
10	10	100β	100β
5	20	100β	0.222β
1	100	100β	0.056β

Table 3.2: Log general odds ratios under the proposed model and alternative models with reference values $y_1^0 = y_2^0 = 0$

y1	y2	proposed model	model 1
1	1	β	β
1	2	0.667β	β
y1	y2	proposed model	model 2
2	3	2β	2β
3	3	9β	9β
10	20	0.995β	9.524β
5	20	0.222β	3.226β

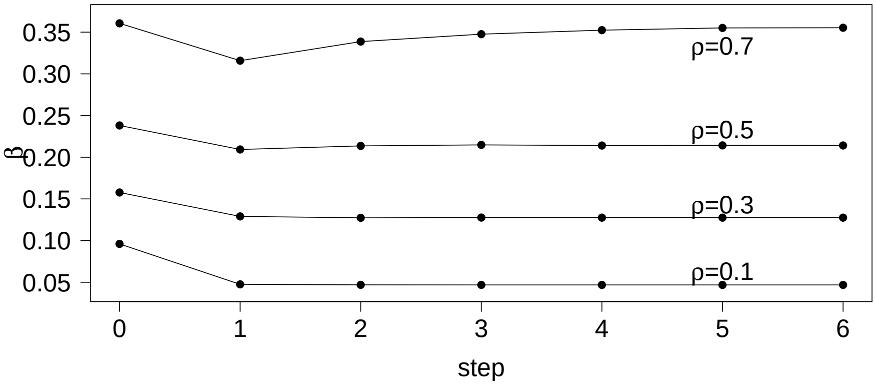

Figure 3.1: Iterative estimates using the “marginal fittings” method

3.2 Simulation Studies

We conducted a simulation study to assess the performance of $l^{(2)}(\beta;\eta,\lambda)$ incorporating the general odds ratio function. We let (Y_{ijk}, Y_{ijl}) follow a bivariate normal distribution, where marginally $Y_{ijk} \sim N(\lambda_i + x_{ijk}^T\alpha, \phi)$, and chose the bilinear log odds ratio model $\psi(y_{ijk}, y_{ijl}; z_{ijkl}, \beta) = \beta z_{ijkl} y_{ijk} y_{ijl}$, with pairwise covariate $z_{ijkl} = |x_{ijk} - x_{ijl}|$, where $x_{ijk} \sim \text{Unif}(-1,1)$, and specified $\alpha = 1$, $\phi = 1$, $\beta = 1$, $w_{ij} = 1$, $\lambda_i \sim \text{Unif}(-1,1)$. We used the Nelder-Mead method to obtain $\hat{\eta}$ that maximized $l^{(1)}(\eta)$, then estimated nuisance parameters λ using equation (2.7), and finally used a combination of golden section search and successive parabolic interpolation to obtain $\hat{\beta}$ that maximizes $l^{(2)}(\beta;\hat{\eta},\hat{\lambda})$. We compared our estimates to the naive GEE approach, which was not intended for use with sparse data. For GEE, we used algorithms in Brent (1973) first to alternately estimate α, ϕ and λ until convergence, then to estimate β based on the quadratic generalized estimating function.

We considered both the sparse data situation where we specified $p = 200$, $n_i = 6$ and $m_{ij} = 2$, and the less-sparse situation where we specified $p = 50$, $n_i = 30$ and $m_{ij} = 2$. For each scenario, we considered both extensive intracluster pairwise dependence ($\beta = 3$) and less-extensive intracluster pairwise dependence ($\beta = 1$). We estimated nuisance parameters in two ways: 1) we obtained $\hat{\lambda}_i$ by (2.7); and 2) we used a Gaussian-based shrinkage estimator $\tilde{\lambda}_i = (\hat{\lambda}_i + \sum_{i'} \hat{\lambda}_{i'}/p)/2$. Results for inferences on η and β are shown in Table 3.3. As expected, the composite conditional likelihood (CCL) approach was barely influenced by the estimation method of nuisance parameters, whereas the GEE approach was more sensitive to fitting the nuisance parameters. CCL achieved lower bias of $\hat{\beta}$ under all the scenarios, especially the sparse data scenario, than the GEE approach. GEE tended to underestimate ϕ under both sparse and less-sparse situations, while the CCL approach yielded valid inference on ϕ. Moreover, GEE estimators of β had large standard errors, and even larger sandwich-based standard error estimates, defects not seen with the CCL approach, which is designed to reduce the aberrant effects of fitting nuisance parameters. In this simulation study, both GEE and CCL yielded valid inference on α.

Next we conducted a simulation study to evaluate the performance of our two sandwich estimators of standard error, one of which was targeted for non-sparse data (Section 2.2.1) and the other for sparse data (Section 2.2.2) in the ambiguous situation of less-sparse data. Specifically, we let (Y_{ijk}, Y_{ijl}) follow a bivariate normal distribution, where marginally $Y_{ijk} \sim N(\lambda_i + x_{ijk}^T\alpha, \phi)$, and chose the bilinear log odds ratio model $\psi(y_{ijk}, y_{ijl}; z_{ijkl}, \beta) = \beta z_{ijkl} y_{ijk} y_{ijl}$, with pairwise covariate $z_{ijkl} = |x_{ijk} - x_{ijl}|$, where $x_{ijk} \sim \text{Unif}(-1, 1)$, and specified $\alpha = 1$, $\phi = 1$, $\beta = 1$, $w_{ij} = 1$, $\lambda_i \sim \text{Unif}(-1, 1)$. We considered the less-sparse situation where $p = 50$, $n_i = 30$ and $m_{ij} = 2$. We considered both extensive intracluster pairwise dependence ($\beta = 3$) and less-extensive intracluster pairwise dependence ($\beta = 1$). As shown in Table 3.4, both estimates of standard error are close to the empirical standard error, demonstrating that both standard error estimators are valid for inference in the ambiguous setting of less-sparse data.

Table 3.3: Results of 1000 simulated samples for inferences on β, α, and ϕ

β	Data	Method	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$se(\hat{\beta})$	$\hat{\alpha}$	$\hat{se}(\hat{\alpha})$	$se(\hat{\alpha})$	$\hat{\phi}$	$\hat{se}(\hat{\phi})$	$se(\hat{\phi})$
1	sparse	CCL	1.0016	0.0712	0.0727	0.9982	0.0348	0.0335	0.9984	0.0344	0.0333
			1.0010	0.0712	0.0727	0.9982	0.0348	0.0335	0.9984	0.0344	0.0333
		GEE	0.8371	0.1172	0.0858	0.9984	0.0301	0.0317	0.8795	0.0294	0.0286
			0.9801	0.1300	0.0930	0.9985	0.0322	0.0326	0.9917	0.0309	0.0305
	less-sparse	CCL	1.0044	0.0695	0.0695	1.0013	0.0320	0.0315	0.9974	0.0317	0.0328
			1.0045	0.0696	0.0696	1.0013	0.0320	0.0315	0.9974	0.0317	0.0328
		GEE	0.9723	0.1285	0.0911	1.0013	0.0312	0.0312	0.9736	0.0308	0.0319
			1.0662	0.1451	0.0995	1.0011	0.0328	0.0333	1.0611	0.0337	0.0345
3	sparse	CCL	2.9928	0.1595	0.1628	0.9980	0.0355	0.0341	0.9983	0.0382	0.0371
			2.9927	0.1595	0.1627	0.9980	0.0355	0.0341	0.9983	0.0382	0.0371
		GEE	2.7597	0.4833	0.2631	0.9982	0.0300	0.0315	0.8595	0.0322	0.0312
			3.0184	0.5191	0.2985	0.9983	0.0322	0.0326	0.9766	0.0337	0.0332
	less-sparse	CCL	3.0032	0.1558	0.1568	1.0014	0.0324	0.0317	0.9972	0.0351	0.0364
			3.0045	0.1558	0.1568	1.0014	0.0324	0.0317	0.9972	0.0351	0.0364
		GEE	2.9723	0.4989	0.2938	1.0014	0.0315	0.0312	0.9695	0.0340	0.0353
			3.1489	0.5570	0.3392	1.0011	0.0330	0.0333	1.0579	0.0367	0.0379

$\hat{se}(\cdot)$: mean of estimated standard errors; $se(\cdot)$: empirical standard error.
sparse data situation: $p = 200$, $n_i = 6$ and $m_{ij} = 2$
less-sparse situation: $p = 40$, $n_i = 30$ and $m_{ij} = 2$
two rows for each approach in each scenario correspond to
two estimation methods for nuisance parameters:
first row, $\hat{\lambda}$; second row, $\tilde{\lambda}$

Table 3.4: Performance of standard error estimators for less sparse data

β	$\hat{se}_1(\hat{\beta})$	$\hat{se}_2(\hat{\beta})$	$se(\hat{\beta})$
1	0.0717	0.0695	0.0695
3	0.1628	0.1558	0.1568

$\hat{se}_1(\cdot)$ and $\hat{se}_2(\cdot)$: means of estimated standard errors
under standard situation and sparse data situation
$se(\cdot)$: empirical standard error

3.3 Water Quality Study

In the analysis of the water quality study, we omitted 8 small geographic areas with zero HCGI episodes. The resulting data set used for analysis consisted of 130 small geographic areas (strata), with 1-36 households (mean=10.2) per stratum, 2-8 individuals (mean=3.92) per household, and 0-36 HCGI episodes (mean=0.67) per individual. Age was classified into 3 categories in accordance with school attendance policies: 2-5 years, 6-17 years, and greater than 17 years, and the category 2-5 years was specified as the reference group. Male was chosen as the reference group for sex, and purified tap water was chosen as the reference group for drinking water type. We used the composite conditional likelihood approach, where we specified $w_{ij} = 1$ in both (2.5) and (2.6), to draw inferences on the covariates that affected the number of HCGI episodes and the aggregation of HCGI episodes within households. Specifically, we assumed that the HCGI episodes marginally followed Poisson log-linear models, $Y_{ijk} \sim \text{Poi}(\exp(\lambda_i + x_{ijk}^T\alpha + T_{ij}))$, where T_{ij} was a known offset for the person log-months on study. We used our proposed odds ratio model (3.4), with $d(z_{ijkl}, \beta) = z_{ijkl} \otimes \beta$, to assess intracluster pairwise associations. When computing $C_{ijkl,iuvw}$, we started with the initial pairwise density function $f(y_1, y_2) = \exp(\beta y_1 y_2/(1+2(y_1 - y_2)^2)$ and conducted five steps of "marginal fittings", which we confirmed was sufficient for the convergence of parameter estimates. We used sandwich-based standard error estimates, as given in Section 2.2.2, to compute the standard errors of $\hat{\alpha}$ and $\hat{\beta}$.

The results, shown in Table 3.5 and Table 3.6, revealed that the risk of HCGI decreased with age, and females were more prone to HCGI than males. Under an intercept-only odds ratio model, there was evidence that HCGI episodes tended to aggregate within households. To further investigate the household aggregation patterns, we fitted a general odds ratio model that included covariates, and found that HCGI episodes tended to aggregate among school-age children, and also between school-age children and adults, whereas HCGI was least likely to aggregate among adults. After adjusting for covariates, the intercept β_0 was significantly different zero, which suggested an unmeasured household-specific environmental effect on HCGI episodes.

Table 3.5: Fitted regression model for the marginal means

	Estimate	*se*	p-value
α_1: unmodified tap water	0.0752	0.1131	0.5061
α_2: tap water with a purge valve	-0.0045	0.1041	0.9655
α_3: bottled plant water	-0.0767	0.1499	0.6089
ref: purified bottled water	0		
α_4: age 6-17	-0.5570	0.0713	<0.0001
α_5: age > 17	-0.7214	0.0563	<0.0001
ref: age < 6	0		
α_6: female	0.1597	0.0431	0.0002
ref: male	0		

Table 3.6: Fitted regression model for intracluster association

	Estimate	*se*	p-value
Intercept-only model			
β_0: intercept	0.0534	0.0172	0.0019
Covariate model			
β_0: intercept	0.0478	0.0090	<0.0001
β_1: unmodified tap water	0.0087	0.0096	0.3648
β_2: tap water with a purge valve	0.0336	0.0431	0.4354
β_3: bottled plant water	0.0038	0.0480	0.9373
ref: purified bottled water	0		
β_4: children(< 6)$-$children(6-17)	0.0104	0.0492	0.8331
β_5: children(< 6)$-$adult	0.0171	0.0136	0.2095
β_6: children(6-17)$-$children(6-17)	0.0433	0.0166	0.0009
β_7: children(6-17)$-$adult	0.0348	0.0157	0.0266
β_8: adult-adult	-0.0295	0.0095	0.0002
ref: children(< 6)$-$children(< 6)	0		

3.4 Appendix

A1. "Marginal fittings" procedure

We consider arbitrary non-empty probability space $(\Omega_{Y_1}, \mathcal{B}_{Y_1}, \pi_{Y_1})$ and $(\Omega_{Y_2}, \mathcal{B}_{Y_2}, \pi_{Y_2})$ as well as their product $(\Omega, \mathcal{B}, \pi)$, i.e., the set $\Omega = \Omega_{Y_1} \times \Omega_{Y_2}$ equipped with the product measure $\pi = \pi_{Y_1} \times \pi_{Y_2}$. Let $\mathcal{P}$ denote the class of probability measures P on $(\Omega, \mathcal{B})$ which have a positive density $f > 0$ with respect to π, i.e. P is dominated by π and dominates $\pi : P \ll \pi \ll P$. Further let $\mathcal{F}$ be the class of corresponding densities, i.e. the Radon-Nikodym derivatives $f = dP/d\pi$ for any $P \in \mathcal{P}$.

The existence theorem in Osius (2004) says that for any ψ that meets the following two conditions, there exists a joint distribution with given marginal distributions π_{Y_1} and π_{Y_2} and with ψ as its log odds ratio function.

(1) $\log q^{Y_1} = \log[\psi^{Y_1}]$ is π_{Y_1} integrable;

(2) $\log q^{Y_2} = \log[\psi^{Y_2}]$ is π_{Y_2} integrable;

where $\psi^{Y_1} = \int \psi d\pi_{Y_1}$, and $\psi^{Y_2} = \int \psi d\pi_{Y_2}$.

To construct an iterative sequence of pairwise densities, we start with any choice of initial pairwise density, such that the corresponding odds ratio is equivalent toψ_{Y_1,Y_2}. A natural choice of initial pairwise density is $f_{Y_1,Y_2} = \psi_{Y_1,Y_2}$. At the tth step, we adjust both marginals of $f^{(t)}_{Y_1,Y_2}$ and obtain $f^{(t+1)}_{Y_1,Y_2}$ by $f^{(t+1)}_{Y_1,Y_2} = (f^{(t)|Y_1}_{Y_1,Y_2})^{|Y_2}$, where $f^{|X}(x,y) = f(x,y)/f^X(x)$. Specifically, under the product measure, the adjustment from $f^{(t)}_{Y_1,Y_2}$

to $f_{Y_1,Y_2}^{(t+1)}$ is given as:

$$\begin{aligned} f_{Y_1,Y_2}^{(t+1)^1} =& f_{Y_1,Y_2}^{(t)} f_{Y_1} / \int f_{Y_1,Y_2} d\pi_{Y_2}; \\ f_{Y_1,Y_2}^{(t+1)} =& f_{Y_1,Y_2}^{(t+1)^1} f_{Y_2} / \int f_{Y_1,Y_2} d\pi_{Y_1} \end{aligned}$$

After each step, the corresponding odds ratio remains equivalent to ψ_{Y_1,Y_2}. The convergence theorem in Osius (2004) ensures the convergence of the "marginal fittings" procedure to the desired density.

Chapter 4

Investigate Optimal Choices of Cluster-specific Weights

4.1 Introduction

Composite likelihood is constructed as a weighted sum of log-likelihoods of low-dimensional margins (Lindsay, 1988). When the weights are all equal, they can be ignored. The selection of unequal weights to improve efficiency in composite likelihood has been discussed in recent literature. In the standard setting, Varin et al. (2011) obtained optimal weights by comparing the Godambe information (Godambe, 1960) of the composite likelihood versus the Fisher information of the full likelihood. However, efficiency comparison cannot be not easily implemented in our proposed composite conditional likelihood. Recall $l^{(1)}(\eta)$ and $l^{(2)}(\beta;\eta,\lambda)$ proposed in Section 2.1, given as below:

$$l^{(1)}(\eta) = \sum_{i=1}^{p} l_i^{(1)}(\eta) = \sum_{i=1}^{p}\sum_{j<u}^{n_i} l_{iju}^{(1)}(\eta) = \sum_{i=1}^{p}\sum_{j<u}^{n_i} w_{ij}w_{iu}\{\sum_{k=1}^{m_{ij}}\sum_{v=1}^{m_{iu}} \log f(y_{ijk}|Y_{ijk}+Y_{iuv};\eta)\}, \tag{4.1}$$

$$l^{(2)}(\beta;\eta,\lambda) = \sum_{i=1}^{p} l_i^{(2)}(\beta;\eta,\lambda_i) = \sum_{i=1}^{p}\sum_{j<u}^{n_i} l_{iju}^{(2)}(\beta;\eta,\lambda_i) \tag{4.2}$$

$$= \sum_{i=1}^{p}\sum_{j<u}^{n_i} w_{ij}w_{iu}\{\sum_{k\neq l}^{m_{ij}}\sum_{v\neq w}^{m_{iu}} \log f(y_{ijl}|Y_{ijk}=y_{ijk}, Y_{iuv}=y_{iuv}, Y_{ijl}+Y_{iuw}=a_{ijl,iuw};\beta,\eta,\lambda_i)\}.$$

Both $l^{(1)}(\eta)$ and $l^{(2)}(\beta;\eta,\lambda)$ are constructed based on a scheme of conditioning on observations drawn from different clusters. The structural difference of our composite conditional likelihood from the standard estimating functions makes it difficult to achieve optimal weights of our composite conditional likelihood by comparing the Godambe information versus the Fisher information. In addition, since both η and β are parameter vectors, both Godambe information and Fisher information are matrices which cannot be compared directly. In Varin et al. (2011), a comparison of diagonal components corresponding to particular parameters of interest is recommended when parameter is a vector. However, in the water quality study, the elements of main effects parameter vector and pairwise association parameter vector are equally important, and we cannot pick a "particular interest" element, so this method is not satisfactory for our proposed composite conditional likelihood. In Lindsay et al. (2011), an optimal weighting strategy based on optimizing the weighted composite score under the least square criterion is discussed, and an optimal weight matrix is constructed which accommodates the parameter vector situation. However, this approach is not applicable to our situation either, as the cluster-specific weights in both $l^{(1)}(\eta)$ and $l^{(2)}(\beta;\eta,\lambda)$ are scalars.

As discussed above, a direct comparison of information matrices is unavailable in our situation. An alternative approach is to compare the maximum composite likelihood estimator versus the maximum likelihood estimator under certain distribution assumptions (Joe and Lee, 2009). Specifically, Joe and Lee (2009) investigated the optimal cluster-specific weights for pairwise likelihood via multivariate normal data with exchangeable correlation structure. Consider $\mathbf{Y}_i = (Y_{i\mathbf{1}},\ldots,Y_{id_i})' \sim N(\mu 1_{d_i}, \Sigma_{d_i})$, for $i = 1,\ldots,n$. Under exchangeable correlation structure with variance η^2 and correlation ρ, the negative log-likelihood is given as:

$$\begin{aligned} -L_0 =& \sum_i \{\frac{1}{2}d_i \log\eta^2 + \frac{1}{2}(d_i-1)\log(1-\rho) + \frac{1}{2}\log[1+(d_i-1)\rho]\} \\ &+ \frac{1}{2\eta^2(1-\rho)}\sum_i [\sum_{j=1}^{d_i}(y_{ij}-\mu)^2 - \frac{\rho}{1+(d_i-1)\rho}(y_{i+}-d_i\mu)^2], \end{aligned} \tag{4.3}$$

and the negative weighted pairwise likelihood is

$$\begin{aligned} -L_1 =& \sum_i w_i\{\frac{d_i(d_i-1)}{2}\log\eta^2 + \frac{1}{4}d_i(d_i-1)\log(1-\rho^2)\} \\ &+ \frac{1}{2\eta^2(1-\rho)}\sum_i w_i \sum_{1\le j<k\le d_i}[(y_{ij}-\mu)^2 + (y_{ik}-\mu)^2 - \frac{\rho}{1+\rho}(y_{ij}+y_{ik}-2\mu)^2]. \end{aligned} \tag{4.4}$$

Under the assumption that ρ is known, the optimal choice of weight is obtained by directly comparing the pairwise likelihood estimator

$$\hat{\mu}_w = \frac{\sum\limits_i w_i(d_i - 1)Y_{i+}}{\sum\limits_i w_i(d_i - 1)d_{i+}}$$

and maximum likelihood estimator

$$\hat{\mu} = \frac{\sum\limits_i [1 + (d_i - 1)\rho]^{-1} Y_{i+}}{\sum\limits_i [1 + (d_i - 1)\rho]^{-1} d_{i+}}.$$

If ρ is known, the optimal w_i is

$$w_i = (d_i - 1)^{-1}[1 + (d_i - 1)\rho]^{-1},$$

which depends on both the cluster size and the correlation. In the above example, the estimators $\hat{\mu}$ and $\hat{\mu}_w$ shared a similar, simple structure and so could be compared directly. By contrast, our composite conditional log-likelihoods, $l^{(1)}(\eta)$ and $l^{(2)}(\beta; \eta, \lambda)$, are constructed by summing over component log-likelihoods for pairs of clusters, and hence the estimators cannot be compared directly with MLE. Moreover, under the sparse data situation where the number of strata is large, we encounter the “infinitely many nuisance parameters“ problem, and hence the maximum likelihood estimator is not valid.

In this chapter, we conduct an exploratory study on optimal choices of cluster-specific weights of our composite conditional likelihoods. Owing to conditioning, the leading term in the asymptotic bias of the estimator vanishes, so the selection of unequal weights to improve efficiency in our composite conditional likelihoods is mainly about achieving estimators with smaller variance. Specifically, we utilize H$\acute{a}$jek projection and adopt the approach in Joe and Lee (2009) by comparing the maximum composite likelihood estimator versus the maximum likelihood estimator under the standard (i.e., non-sparse) situation. For the more challenging sparse data situation, we investigate the cluster-specific weights via simulation studies.

4.2 Optimal Choices of Cluster-specific Weights for Proposed Composite Conditional Likelihood Under Standard Situation

We investigate the optimal choice of cluster-specific weights under two asymptotic schemes. We first investigate the optimal choice of weight under the standard situation with a fixed number of strata, while the number of clusters are large. We consider a single stratum with n clusters with varying cluster sizes m_j, and observations in each cluster follow a multivariate normal distribution, i.e., $\mathbf{Y}_j = (Y_{j\mathbf{1}}, \dots, Y_{jm_j})' \sim N(\mu_j, \sigma^2\Sigma_j)$, for $j = 1, \dots, n$, where $\mu_j = (\mu_{j1}, \dots, \mu_{jm_j})'$, $\mu_{jk} = \lambda + x_{jk}^T\alpha$, σ^2 is the variance parameter, and Σ_j is the correlation matrix for the jth cluster. Without loss of generality, we consider scalar α here. Then $l^{(1)}(\alpha, \sigma^2)$ is given as:

$$l^{(1)}(\alpha, \sigma^2) = \sum_{j<u}^{n} w_j w_u \sum_{k=1}^{m_j}\sum_{v=1}^{m_u}[\log(\pi\sigma^2)^{-1/2} - \frac{(y_{jk} - y_{uv} - (x_{jk} - x_{uv})\alpha)^2}{4\sigma^2}]. \tag{4.5}$$

Refer to A1 in the Appendix in this chapter for detailed derivation of $l^{(1)}(\alpha, \sigma^2)$. Under this multivariate normal distribution assumption, the log-likelihood $l(\alpha, \sigma^2, \rho)$ is given as:

$$l(\alpha, \sigma^2, \Sigma_1, \dots, \Sigma_n) = \sum_{j=1}^{n}[\log(2\pi\sigma^2)^{-\frac{m_j}{2}}|\Sigma_j|^{-\frac{1}{2}} - \frac{(\mathbf{y}_j - \mu_j)^T\Sigma_j^{-1}(\mathbf{y}_j - \mu_j)}{2\sigma^2}]. \tag{4.6}$$

By solving $\partial l^{(1)}(\alpha, \sigma^2)/\partial\alpha = 0$, the composite conditional likelihood estimator is given as:

$$\tilde{\alpha} = \frac{\sum_{j<u}^{n} w_j w_u \sum_{k=1}^{m_j}\sum_{v=1}^{m_u}(y_{jk} - y_{uv})(x_{jk} - x_{uv})}{\sum_{j<u}^{n} w_j w_u \sum_{k=1}^{m_j}\sum_{v=1}^{m_u}(x_{jk} - x_{uv})^2} \tag{4.7}$$

By solving $\partial l(\alpha, \sigma^2, \rho)/\partial\alpha = 0$ assuming ρ and λ are known, the maximum likelihood estimator is given as:

$$\hat{\alpha} = \frac{\sum\limits_{j=1}^{n} \mathbf{x}_j^T \Sigma_j^{-1}(\mathbf{y}_j - \lambda \mathbf{1}_{m_j})}{\sum\limits_{j=1}^{n} \mathbf{x}_j^T \Sigma_j^{-1} \mathbf{x}_j} \tag{4.8}$$

Since $\tilde{\alpha}$ relies on summation over pairs of clusters in both numerator and denominator, it is not directly comparable with $\hat{\alpha}$ given by (4.8). To enhance comparability, we propose conducting Hajek projection to the composite conditional likelihood and comparing the projected composite likelihood estimator with the maximum likelihood estimator. Under the regularity condition $E[\sum\limits_{k=1}^{m_j}\sum\limits_{v=1}^{m_u} \text{log} f(y_{jk}|Y_{jk}+Y_{uv}, \alpha, \sigma^2)]^2 < \infty$,

$$\begin{aligned}
&\sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \text{log} f(y_{jk}|Y_{jk}+Y_{uv}, \alpha, \sigma^2) \\
\approx &\sum_{i=1}^{n} E[\sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \text{log} f(y_{jk}|Y_{jk}+Y_{uv})|\mathbf{Y}_i] - (n-1)E[\sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \text{log} f(y_{jk}|Y_{jk}+Y_{uv})] \\
= &\sum_{i=1}^{n}\{E[\sum_{u>i}^{n}\sum_{k=1}^{m_i}\sum_{v=1}^{m_u} \text{log} f(y_{ik}|Y_{ik}+Y_{uv})|\mathbf{Y}_i]/2 + E[\sum_{j<i}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_i} \text{log} f(y_{hk}|Y_{jk}+Y_{iv})|\mathbf{Y}_i]/2\},
\end{aligned}$$

Refer to A2 in Appendix in this chapter for detailed derivation.

Then $l^{(1)}(\alpha, \sigma^2)$ can be approximated as:

$$\begin{aligned}
&l^{(1)}(\alpha, \sigma^2) = \sum_{j<u}^{n} w_j w_u \sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \text{log} f(y_{jk}|Y_{jk}+Y_{uv}, \alpha, \sigma^2) \\
\approx &\sum_{i=1}^{n} w_i\{E[\sum_{u>i}^{n}\sum_{k=1}^{m_i}\sum_{v=1}^{m_u} \text{log} f(y_{ik}|Y_{ik}+Y_{uv})|\mathbf{Y}_i]/2 + E[\sum_{j<i}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_i} \text{log} f(y_{jk}|Y_{jk}+Y_{iv})|\mathbf{Y}_i]/2\}
\end{aligned}$$

Under multivariate normal distribution assumption, we have

$$\begin{aligned} & l^{(1)}(\alpha, \sigma^2) \qquad (4.9) \\ \approx & \sum_{i=1}^{n} w_i \{ \sum_{u>i}^{n} \sum_{k=1}^{m_i} \sum_{v=1}^{m_u} \frac{(y_{ik} - x_{ik}\alpha - \lambda)^2}{8\sigma^2} + \sum_{j<i}^{n} \sum_{k=1}^{m_j} \sum_{v=1}^{m_i} \frac{(y_{iv} - x_{iv}\alpha - \lambda)^2}{8\sigma^2} \} + C \\ = & \sum_{i=1}^{n} w_i \{ \sum_{j=1, j\neq i}^{n} m_j \sum_{k=1}^{m_i} \frac{(y_{ik} - x_{ik}\alpha - \lambda)^2}{8\sigma^2} \} + C, \end{aligned}$$

where C is free of α.

We obtain a projected composite conditional likelihood estimator $\tilde{\alpha}_s$ by maximizing the right-hand size of equation (4.9), giving:

$$\tilde{\alpha}_s = \frac{\sum_{i=1}^{n} w_j \sum_{j=1, j\neq i}^{n} m_j [\sum_{k=1}^{m_i} x_{ik}(y_{ik} - \lambda)]}{\sum_{i=1}^{n} w_j \sum_{j=1, j\neq i}^{n} m_j [\sum_{k=1}^{m_i} x_{ik}^2]}, \qquad (4.10)$$

which is now comparable to the MLE, since it has the same form of summation with MLE in both numerator and denominator. Following the basic approach in Joe and Lee (2009), we can compare $\tilde{\alpha}_s$ and $\hat{\alpha}$ to yield insights into the optimal weights. These cluster-specific weights generally depend on cluster sizes, as well as the correlation structure and covariates. So we consider some choices of weights under different assumptions.

First, considering weights that only depend on cluster sizes which is our primary interest, a comparison of $\tilde{\alpha}_s$ and $\hat{\alpha}$ suggests we assign weight $w_j^{(1)} = \frac{1}{M - m_j}$, where $M = \sum_{j=1}^{n} m_j$, to the jth cluster. Under this choice of weights, $\hat{\alpha}$ is equivalent to $\tilde{\alpha}_s$ when we assume independent correlation structure. More generally, this choice of weights, based on Hajek projection, does not actually depend on the multivariate normal assumption, and so it is still appropriate when the outcome is not normally distributed (Refer to A3 in the Appendix in this chapter for a discussion on the robustness of this choice of weights to the multivariate normal distribution assumption). It is notable that, unlike standard composite likelihoods that do not rely on conditioning, where it is common to use the weights $w_j = \frac{1}{m_j - 1}$ (Varin et al., 2011), under our pairwise conditioning setting the more appropriate weights $w_j^{(1)} = \frac{1}{M - m_j}$ increase as the cluster size increases. Under the standard situation where n is large and m_j is uniformly bounded, we can expect this choice of weights to be close to equal weights.

Second, we consider weights that depend on both cluster sizes and correlation structure. We consider exchangeable correlation structure with intracluster correlation ρ. Under this assumption, $\hat{\alpha}$ is given by:

$$\hat{\alpha} = \frac{\sum_{j=1}^{n}[\sum_{k=1}^{m_j}\{x_{jk}(y_{jk}-\lambda)[1+(m_j-2)\rho] - \sum_{l\neq k} x_{jk}(y_{jl}-\lambda)\rho\}]\frac{1}{1+(m_j-1)\rho}}{\sum_{j=1}^{n}[\sum_{k=1}^{m_j}\{x_{jk}^2[1+(m_j-2)\rho] - \sum_{l\neq k} x_{jk}x_{jl}\rho\}]\frac{1}{1+(m_j-1)\rho}} \tag{4.11}$$

When $\rho \neq 0$, we cannot assign weights such that $\tilde{\alpha}_s$ has the same form as $\hat{\alpha}$. but a comparison suggests that we consider assigning weight $w_j^{(2)} = \frac{1}{M-m_j}\frac{1+(m_j-2)\rho}{1+(m_j-1)\rho}$ to the jth cluster. When m_j is large, $w_j^{(2)}$ is close to $w_j^{(1)}$. Unlike $w_j^{(1)}$, $w_j^{(2)}$ depends on the distribution assumption, so $w_j^{(2)}$ should be adjusted accordingly when the observations are not multivariate normal.

Third, we consider weights that depend on cluster sizes, correlation structure and covariates. We consider a special case where $x_{jk} \sim$ Bernoulli (0.5). Without loss of generality, we assume that the observations in each cluster are sorted so that the first half are 1 and the latter half are 0. Under this assumption, ignoring whether m_j is odd or even, $\hat{\alpha}$ is given by:

$$\hat{\alpha} = \frac{\sum_{j=1}^{n}[\sum_{k=1}^{m_j/2}\{(y_{jk}-\lambda)[1+(m_j-2)\rho] - \sum_{l\neq k}(y_{jl}-\lambda)\rho\}]\frac{1}{1+(m_j-1)\rho}}{\sum_{j=1}^{n}\frac{m_j}{2}[1+(m_j/2-1)\rho]\frac{1}{1+(m_j-1)\rho}}, \tag{4.12}$$

and $\tilde{\alpha}$ is given by:

$$\tilde{\alpha} = \frac{\sum_{j=1}^{n} w_i \sum_{j=1,j\neq i}^{n} m_j[\sum_{k=1}^{m_i/2}(y_{ik}-\lambda)]}{\sum_{j=1}^{n} w_i \sum_{j=1,j\neq i}^{n} m_j(\frac{m_i}{2})} \tag{4.13}$$

By comparing these two estimators, we consider assigning weight $w_j^{(3)} = \frac{1}{M-m_j}\frac{1+(m_j/2-1)\rho}{1+(m_j-1)\rho}$ to the jth cluster. Under this choice of weights, the denominators of $\hat{\alpha}$ and $\tilde{\alpha}_s$ are the same, and the numerators are close given than y_{jk} has mean λ when x_{jk} is 0, so the two estimators are approximately equivalent. Here, the $w_j^{(3)}$ depend on the distribution assumption, so $w_j^{(3)}$ should be adjusted accordingly under other distribution assumption.

To summarize, under the standard (i.e., non-sparse) situation with single stratum, the following weights are suggested for $l^{(1)}(\alpha, \sigma^2)$:

(a) when cluster-specific weights are allowed to depend on cluster sizes only, we suggest assigning weights $w_j^{(1)} = \frac{1}{M-m_j}$ to the jth cluster;

(b) when cluster-specific weights are allowed to depend on cluster sizes and correlation structure, under exchangeable multivariate normal distribution assumption, we suggest assigning weights $w_j^{(2)} = \frac{1}{M-m_j}\frac{1+(m_j-2)\rho}{1+(m_j-1)\rho}$ to the jth cluster

(c) when cluster-specific weights are allowed to depend on cluster sizes, correlation structure and covariates, under exchangeable multivariate normal distribution assumption with $x_{jk} \sim$ Bernoulli (0.5), we suggest assigning weight $w_j^{(3)} =$

$\frac{1}{M-m_j}\frac{1+(m_j/2-1)\rho}{1+(m_j-1)\rho}$ to the jth cluster.

For (b) and (c), the choices of weights depend on the distribution assumption. Under other distribution assumptions, the weights should be adjusted accordingly.

We apply the same strategy to $l^{(2)}(\beta;\eta,\lambda)$ to investigate the optimal choice of cluster-specific weights. Specifically, we first conduct Hajek projection to $l^{(2)}(\rho)$ as:

$$\begin{aligned}
l^{(2)}(\rho) =& \sum_{j<u}^{n} w_j w_u \sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_u} \log\, f(y_{jl}|Y_{jk}=y_{jk}, Y_{uv}=y_{uv}, Y_{jl}+Y_{uw}=a_{jl,uw};\rho)\\
&\approx \sum_{i=1}^{n} w_i\{E[\sum_{u>i}^{n}\sum_{k\neq l}^{m_i}\sum_{v\neq w}^{m_u} \log\, f(y_{il}|Y_{ik}=y_{ik}, Y_{uv}=y_{uv}, Y_{il}+Y_{uw}=a_{il,uw})|\mathbf{Y}_i]/2\\
&\quad + E[\sum_{j<i}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_i} \log\, f(y_{jl}|Y_{jk}=y_{jk}, Y_{iv}=y_{iv}, Y_{jl}+Y_{iw}=a_{jl,iw})|\mathbf{Y}_i]/2\}
\end{aligned}$$

Refer to A4 in Appendix in this chapter for more details.

After conducting Hajek projection, we investigate the optimal choices of weights under exchangeable multivariate normal distribution assumption with no covariates. Under this assumption, the log-likelihood $l(\rho)$ is given as:

$$\begin{aligned}
l(\rho) &= \sum_{j=1}^{n}\{\log[(2\pi\sigma^2)^{-\frac{m_j}{2}}|\Sigma_j|^{-\frac{1}{2}}] - \frac{(\mathbf{y}_j-\lambda\mathbf{1})^T\Sigma_j^{-1}(\mathbf{y}_j-\lambda\mathbf{1})}{2\sigma^2}\}\\
&= \sum_{j=1}^{n}\{-\frac{m_j}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|\Sigma_j| - \frac{(\mathbf{y}_j-\lambda\mathbf{1})^T\Sigma_j^{-1}(\mathbf{y}_j-\lambda\mathbf{1})}{2\sigma^2}\}
\end{aligned}$$

and $l^{(2)}(\rho)$ is given as:

$$\begin{aligned}
l^{(2)}(\rho) &\approx \sum_{i=1}^{n} w_i \sum_{j=1,j\neq i}^{n} m_j(m_j-1)\{\sum_{k\neq l}^{m_i}[-\frac{1}{2}\log\frac{2\pi\sigma^2(1-\rho^2)}{2} - \frac{(y_{il}-y_{ik})^2-2\rho\sigma^2}{4(1-\rho^2)\sigma^2}]\}\\
&= \sum_{i=1}^{n} w_i \sum_{j=1,j\neq i}^{n} m_j(m_j-1)\{-\frac{m_j(m_j-1)}{2}\log(2\pi\sigma^2) - \sum_{k\neq l}^{m_i}[\frac{1}{2}\log\frac{(1-\rho^2)}{2} + \frac{(y_{il}-y_{ik})^2-2\rho\sigma^2}{4(1-\rho^2)\sigma^2}]\}
\end{aligned}$$

Refer to A5 in Appendix in this chapter for more details.

Owing to varying cluster sizes, we need to reduce n fractions to a common fraction to obtain the analytical form of $\hat{\rho}$ even under this simple situation where there are no covariates, so we cannot easily follow the approach in Joe and Lee (2009) to determine choices of weights by comparing estimators. As an alternative strategy, we compare $l(\rho)$ versus $l^{(2)}(\rho)$. We consider cluster-specific weights that depend on cluster sizes only, and it is natural that we assign weight $w_j^{(4)} = \frac{1}{(m_j-1)\sum_{i=1,i\neq j}^{n} m_i(m_i-1)}$ to the jth cluster. Moreover, this choice of weights does not depend on the distribution assumption, so it is still appropriate when the outcome is not normally distributed. Under the standard situation where n is large and m_j is uniformly bounded, we can expect this choice of weights to be close to $\frac{1}{(m_j-1)}$.

The optimal weights under the single stratum situation can be directly generalized to multiple strata situation. Under the standard situation where the number of strata is fixed and the number of clusters per-stratum is large, we apply Hajek projection to each stratum of $l^{(1)}(\alpha)$ and $l^{(2)}(\rho)$, and the sum of the stratum-level remainders is still negligible. Therefore, the optimal cluster-specific weights under the single stratum situation $w_j^{(1)}$, $w_j^{(2)}$, $w_j^{(3)}$ and $w_j^{(3)}$, $j = 1, \ldots, n$ become: 1) $w_{ij}^{(1)} = \frac{1}{\sum_{k=1,k\neq j}^{n_j} m_{ik}}$; 2) $w_{ij}^{(2)} = \frac{1}{\sum_{k=1,k\neq j}^{n_j} m_{ik}} \frac{1+(m_{ij}-2)\rho}{1+(m_{ij}-1)\rho}$; 3) $w_{ij}^{(3)} = \frac{1}{\sum_{k=1,k\neq j}^{n_j} m_{ik}} \frac{1+(m_{ij}/2-1)\rho}{1+(m_{ij}-1)\rho}$; and 4) $w_{ij}^{(4)} = \frac{1}{(m_{ij}-1)\sum_{k=1,k\neq j}^{n} m_{ik}(m_{ik}-1)}$.

We conducted simulation studies to assess the performances of $l^{(1)}(\alpha)$ and $l^{(2)}(\rho)$ under different weights. We first conducted a simulation study for $l^{(1)}(\alpha)$. We let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijm_{ij}})'$ follow an exchangeable multivariate normal distribution with intracluster correlation ρ, where marginally $Y_{ijk} \sim N(\lambda_i + x_{ijk}^T\alpha, \sigma^2)$, where $x_{ijk} \sim$ Bernoulli (0.5), and specified $\alpha = 2$, $\phi = 1$, $\lambda_i \sim$ Unif$(-1, 1)$. We obtained $\hat{\alpha}$ by maximizing $l^{(1)}(\alpha)$. We considered non-sparse situation, where $p = 4$, $n_j = 50$, and m_{ij} were between 2 to 10 with equal probability. We considered different levels of intracluster correlation where $\rho = 0.1, 0.3, 0.5, 0.7, \text{or } 0.9$. We considered five choices of weights: 1) $w_{ij}^{(1)} = \frac{1}{\sum\limits_{k=1,k\neq j}^{n_j} m_{ik}}$; 2) $w_{ij}^{(2)} = \frac{1}{\sum\limits_{k=1,k\neq j}^{n_j} m_{ik}} \frac{1+(m_{ij}-2)\rho}{1+(m_{ij}-1)\rho}$; 3) $w_{ij}^{(3)} = \frac{1}{\sum\limits_{k=1,k\neq k}^{n_j} m_{ij}} \frac{1+(m_{ij}/2-1)\rho}{1+(m_{ij}-1)\rho}$; 4) equal weights $w_{ij}^{(5)} = 1$; and 5) commonly used weights for composite likelihood $w_{ij}^{(6)} = \frac{1}{m_{ij}-1}$ (Varin et al., 2011). Results for inferences on α are shown in Table 4.1. As expected, $w_{ij}^{(1)}$, $w_{ij}^{(2)}$ and $w_{ij}^{(3)}$ had similar performance with equal weights, with our recommended weights $w_{ij}^{(3)}$ performing the best in terms of achieving the estimator with the smallest empirical standard deviation, and the smallest mean square error under all levels of ρ, while the commonly used weights for composite likelihood $w_{ij}^{(6)}$ performed worst.

We also conducted a simulation study for $l^{(2)}(\beta)$. We considered a single stratum with n clusters with varying cluster sizes m_j, $j = 1, \ldots, n$, and observations in each cluster follow a exchangeable multivariate normal distribution with intracluster correlation ρ and no covariates, where marginally $Y_{jk} \sim N(\lambda, \sigma^2)$, $k = 1, \ldots, m_j$, and specified $\phi = 1$, $\lambda \sim \text{Unif}(-1, 1)$. We obtained $\hat{\rho}$ by maximizing $l^{(2)}(\rho)$. We considered non-sparse situation where $p = 4$, $n = 50$, and m_j were between 2 to 10 with equal probability. We considered different levels of intracluster correlation where $\rho = 0.1, 0.3, 0.5, 0.7, \text{or } 0.9$. We considered four choices of cluster-specific weights: 1) $w_j^{(4)}$; 2) equal weights $w_j^{(5)} = 1$; 3) commonly used weights for composite likelihood $w_j^{(6)} = \frac{1}{m_j-1}$ (Varin et al., 2011); 4) commonly used weights for composite pairwise likelihood $w_j^{(7)} = \frac{1}{m_j(m_j-1)}$ (Varin et al., 2011). Results for inferences on ρ are shown in Table 4.2. As expected, all the results were only mildly sensitive to the choice of weights, with our recommended weights $w_j^{(4)}$ performing the best in terms of achieving the estimator with the smallest empirical standard deviation, and the smallest mean square error under all levels of ρ, followed by $w_j^{(6)}$ which $w_j^{(4)}$ is close to under the standard situation, and then equal weight, while the commonly used weights for composite pairwise likelihood $w_j^{(7)}$ performed worst.

For the more challenging sparse data situation, where MLE is not consistent, it would be inappropriate to investigate optimal choices of weights by comparing the composite conditional likelihood estimator versus MLE. Moreover, when cluster sizes vary, the minimum sufficient statistics for λ, when other parameters are known, generally depends on the known parameters, and so we cannot easily compare the composite conditional likelihood estimator versus the consistent maximum conditional likelihood estimator. In addition, when the number of strata is large, the sum of the stratum-level remainders of our composite conditional likelihood after conducting Hajek projection might not be negligible. Therefore, when the data are sparse, we investigate optimal choices of cluster-specific weights for our proposed composite conditional likelihood via simulation studies, as detailed in the next Section.

Table 4.1: Inference on α under different choices of weights based on 1000 simulated samples

weight		ρ				
		0.1	0.3	0.5	0.7	0.9
1	EmpBias	0.000887	0.001985	0.002933	0.003835	0.004805
	EmpSD	0.057952	0.058261	0.058710	0.05929	0.060071
	MSE	0.003359	0.003398	0.003456	0.003530	0.003632
2	EmpBias	0.000881	0.001979	0.002955	0.003908	0.004953
	EmpSD	0.057955	0.058281	0.058763	0.059386	0.060218
	MSE	0.003360	0.003401	0.003462	0.003542	0.003651
3	EmpBias	0.000917	0.001997	0.002918	0.003811	0.004792
	EmpSD	0.058045	0.058424	0.058821	0.059343	0.060083
	MSE	0.003370	0.003417	0.003468	0.003536	0.003633
4	EmpBias	0.001023	0.002100	0.003024	0.003899	0.004831
	EmpSD	0.058002	0.058333	0.058785	0.05935	0.060088
	MSE	0.003365	0.003407	0.003465	0.003538	0.003634
5	EmpBias	0.001590	0.002276	0.002847	0.003364	0.003871
	EmpSD	0.074321	0.074512	0.074808	0.075224	0.075841
	MSE	0.005526	0.005557	0.005604	0.005670	0.005767

weight:1. $w_{ij}^{(1)}$; 2. $w_{ij}^{(2)}$; 3. $w_{ij}^{(3)}$; 4. $w_{ij}^{(5)}$; and 5. $w_{ij}^{(6)}$

Table 4.2: Inference on ρ under different choices of weights based on 1000 simulated samples

		ρ				
		0.1	0.3	0.5	0.7	0.9
1	EmpBias	0.005796	0.005968	0.005158	0.003493	0.001128
	EmpSD	0.048977	0.055043	0.043839	0.026546	0.008753
	MSD	0.002432	0.003065	0.001948	0.000717	7.80E-05
2	EmpBias	0.007916	0.007387	0.006108	0.004105	0.001303
	EmpSD	0.049638	0.058906	0.04714	0.028363	0.009322
	MSD	0.002527	0.003525	0.002259	0.000821	8.90E-05
3	EmpBias	0.005718	0.005907	0.005116	0.003468	0.001123
	EmpSD	0.049123	0.055059	0.043857	0.026573	0.008763
	MSD	0.002446	0.003066	0.00195	0.000718	7.80E-05
4	EmpBias	0.003679	0.002484	0.002704	0.00208	0.000719
	EmpSD	0.064258	0.067682	0.052888	0.032301	0.010688
	MSD	0.004143	0.004587	0.002804	0.001048	0.000115

weight:1. $w_j^{(4)}$; 2. $w_j^{(5)}$; 3. $w_j^{(6)}$; and 4. $w_j^{(7)}$

4.3 Optimal Choices of Cluster-specific Weights for Proposed Composite Conditional Likelihood Under Sparse Data Situation

As discussed in Section 4.2, when the data are sparse it is extremely challenging to derive optimal weights for composite conditional likelihood analytically. We instead investigate the optimal choices of cluster-specific weights for sparse data via simulation studies.

We first investigate the choices of weights for $l^{(1)}(\alpha)$. We let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijm_{ij}})'$ follow an exchangeable multivariate normal distribution with intracluster correlation ρ, where marginally $Y_{ijk} \sim N(\lambda_i + x_{ijk}^T\alpha, \sigma^2)$, where $x_{ijk} \sim$ Bernoulli (0.5), and specify $\alpha = 2$, $\phi = 1$, $\lambda_i \sim$ Unif$(-1, 1)$. We obtained $\hat{\alpha}$ by maximizing $l^{(1)}(\alpha)$. We considered sparse data, where $p = 50$, $n_j = 4$, and m_{ij} were between 2 to 10 with equal probability. We considered different levels of intracluster correlation where $\rho = 0.1, 0.3, 0.5, 0.7, \text{or } 0.9$. We first considered the five weights compared in the simulation study in the simpler non-sparse setting: 1) $w_{ij}^{(1)} = \frac{1}{\sum\limits_{k=1,k\neq j}^{n_j} m_{ik}}$; 2) $w_{ij}^{(2)} = \frac{1}{\sum\limits_{k=1,k\neq j}^{n_j} m_{ik}} \frac{1+(m_{ij}-2)\rho}{1+(m_{ij}-1)\rho}$; 3) $w_{ij}^{(3)} = \frac{1}{\sum\limits_{k=1,k\neq k}^{n_j} m_{ij}} \frac{1+(m_{ij}/2-1)\rho}{1+(m_{ij}-1)\rho}$; 4) equal weights $w_{ij}^{(5)} = 1$; and 5) commonly used weights for composite likelihood $w_{ij}^{(6)} = \frac{1}{m_{ij}-1}$ (Varin et al., 2011). However, as shown in Table 4.3, equal weights performed better than $w_{ij}^{(1)}$, $w_{ij}^{(3)}$ and $w_{ij}^{(5)}$ in terms of achieving the estimator with smaller empirical standard deviation and smaller mean square error under all levels of ρ, and performed better than $w_{ij}^{(2)}$ under weak intracluster correlation.

The unsatisfactory performance of the above weights which performed well in the non-sparse setting is highly likely related to the large number of strata under the sparse data situation. Therefore, we consider adjusting the weights in the sparse data situation as: 1) $w_{ij}^{(1*)} = \frac{1}{\sum\limits_{s=1}^{p}\sum\limits_{k=1,k\neq j}^{n_j} m_{sk}}$; 2) $w_{ij}^{(2*)} = \frac{1}{\sum\limits_{s=1}^{p}\sum\limits_{k=1,k\neq j}^{n_j} m_{sk}}\frac{1+(m_{ij}-2)\rho}{1+(m_{ij}-1)\rho}$; 3) $w_{ij}^{(3*)} = \frac{1}{\sum\limits_{s=1}^{p}\sum\limits_{k=1,k\neq k}^{n_j} m_{sj}}\frac{1+(m_{ij}/2-1)\rho}{1+(m_{ij}-1)\rho}$. Under the same simulation setting, we compare these three adjusted weights with 4) equal weights $w_{ij}^{(5)} = 1$; and 5) commonly used weights for composite likelihood $w_{ij}^{(6)} = \frac{1}{m_{ij}-1}$. As shown in Table 4.4, our composite conditional likelihood was only mildly sensitive to the choice of weights, with $w_{ij}^{(3*)}$ performing the best in terms of achieving the estimator with the smallest empirical standard deviation and the smallest mean square error under all levels of ρ, followed by $w_{ij}^{(1*)}$, and then equal weights, while the commonly used weights for composite likelihood $w_{ij}^{(6)}$ performed worst. Therefore, in the sparse data situation, we suggest:1) assigning weights $w_{ij}^{(1*)} = \frac{1}{\sum\limits_{s=1}^{p}\sum\limits_{k=1,k\neq j}^{n_j} m_{sk}}$ to the ith stratum, jth cluster for $l^{(1)}(\alpha)$ when cluster-specific weights are allowed to depend on cluster size only; and 2) assigning weights $w_{ij}^{(3*)} = \frac{1}{\sum\limits_{s=1}^{p}\sum\limits_{k=1,k\neq k}^{n_j} m_{sj}}\frac{1+(m_{ij}/2-1)\rho}{1+(m_{ij}-1)\rho}$ to the ith stratum, jth cluster under the above distribution assumption when cluster-specific weights are allowed to depend on cluster size, correlation structure and covariates.

We also investigated otpimal choices of weights for $l^{(2)}(\rho)$. We let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijm_{ij}})'$ follow a exchangeable multivariate normal distribution with intracluster correlation ρ and no covariates, where marginally $Y_{ijk} \sim N(\lambda_i, \sigma^2)$, and specified $\phi = 1$, $\lambda_i \sim \text{Unif}(-1, 1)$. We obtained $\hat{\rho}$ by maximizing $l^{(2)}(\rho)$. We considered sparse data situation where we specified $p = 50$, $n_j = 4$, and m_{ij} were between 2 to 10 with equal probability. We considered different levels of intracluster correlation where $\rho = 0.1, 0.3, 0.5, 0.7, \text{or } 0.9$. Based on the results for $l^{(2)}(\rho)$ in the non-sparse setting and the results for $l^{(1)}(\alpha)$ in the sparse setting, we considered five choices of cluster-specific weights: 1) the optimal weights in the non-sparse setting

$w_{ij}^{(4)} = \frac{1}{(m_{ij}-1)\sum\limits_{k=1,k\neq j}^{n} m_{ik}(m_{ik}-1)}$; 2) adjusted $w_{ij}^{(4)}$, $w_{ij}^{(4*)} = \frac{1}{(m_{ij}-1)\sum\limits_{s=1}^{p}\sum\limits_{k=1,k\neq j}^{n} m_{sk}(m_{sk}-1)}$; 3) equal weights $w_{ij}^{(5)} = 1$; and 4) commonly used weights for composite pairwise likelihood $w_{ij}^{(7)} = \frac{1}{m_{ij}(m_{ij}-1)}$. Results for inferences on ρ are shown in Table 4.5. Similar to $l^{(1)}(\alpha)$, the optimal weights $w_{ij}^{(4)}$ which performed well in the non-sparse setting performed worse than equal weights, but the adjusted weights $w_{ij}^{(4*)}$ achieved the estimator with the smallest empirical standard deviation and smallest mean square error under all levels of ρ. Therefore, under sparse data situation, we suggest assigning weights $w_{ij}^{(4*)} = \frac{1}{(m_{ij}-1)\sum\limits_{s=1}^{p}\sum\limits_{k=1,k\neq j}^{n} m_{sk}(m_{sk}-1)}$ to the ith stratum, jth cluster for $l^{(2)}(\rho)$.

Table 4.3: Inference on α under different choices of weights based on 1000 simulated samples

weight		ρ				
		0.1	0.3	0.5	0.7	0.9
1	EmpBias	0.000625	0.000839	0.001009	0.001154	0.001281
	EmpSD	0.06234	0.06432	0.066258	0.06817	0.070086
	MSE	0.003887	0.004138	0.004391	0.004648	0.004914
2	EmpBias	0.000664	0.000979	0.001217	0.001399	0.001526
	EmpSD	0.062136	0.063619	0.065296	0.067138	0.069138
	MSE	0.003861	0.004048	0.004265	0.00451	0.004782
3	EmpBias	0.000422	0.000593	0.000791	0.000948	0.001052
	EmpSD	0.063632	0.065838	0.067292	0.068691	0.070178
	MSE	0.004049	0.004335	0.004529	0.004719	0.004926
4	EmpBias	0.002279	0.002334	0.00232	0.002236	0.002026
	EmpSD	0.061567	0.063559	0.065555	0.067564	0.069633
	MSE	0.003796	0.004045	0.004303	0.00457	0.004853
5	EmpBias	0.000481	0.000647	0.00077	0.000863	0.000921
	EmpSD	0.080326	0.082497	0.084493	0.086345	0.08802
	MSE	0.006453	0.006806	0.00714	0.007456	0.007748

weight:1. $w_{ij}^{(1)}$; 2. $w_{ij}^{(2)}$; 3. $w_{ij}^{(3)}$; 4. $w_{ij}^{(5)}$; and 5. $w_{ij}^{(6)}$

Table 4.4: Inference on α under different choices of weights based on 1000 simulated samples

		ρ				
		0.1	0.3	0.5	0.7	0.9
1	EmpBias	0.002236	0.002284	0.002264	0.002175	0.001963
	EmpSD	0.061519	0.063517	0.065521	0.067538	0.069619
	MSE	0.00379	0.00404	0.004298	0.004566	0.004851
2	EmpBias	0.002261	0.002354	0.002343	0.002236	0.001977
	EmpSD	0.06156	0.063733	0.065984	0.068308	0.070772
	MSE	0.003795	0.004067	0.004359	0.004671	0.005013
3	EmpBias	0.002101	0.00216	0.002196	0.00215	0.00196
	EmpSD	0.061461	0.063453	0.065404	0.067415	0.069558
	MSE	0.003782	0.004031	0.004283	0.004549	0.004842
4	EmpBias	0.002279	0.002334	0.00232	0.002236	0.002026
	EmpSD	0.061567	0.063559	0.065555	0.067564	0.069633
	MSE	0.003796	0.004045	0.004303	0.00457	0.004853
5	EmpBias	0.000481	0.000647	0.00077	0.000863	0.000921
	EmpSD	0.080326	0.082497	0.084493	0.086345	0.08802
	MSE	0.006453	0.006806	0.00714	0.007456	0.007748

weight:1. $w_{ij}^{(1*)}$; 2. $w_{ij}^{(2*)}$; 3. $w_{ij}^{(3*)}$; 4. $w_{ij}^{(5)}$; and 5. $w_{ij}^{(6)}$

Table 4.5: Inference on ρ under different choices of weights based on 700 simulated samples

		ρ				
		0.1	0.3	0.5	0.7	0.9
1	EmpBias	-0.004111	-0.004991	-0.003012	-0.001025	-0.000103
	EmpSD	0.050218	0.051395	0.039707	0.023874	0.007807
	MSE	0.002539	0.002666	0.001586	0.000571	6.10E-05
2	EmpBias	-0.002272	-0.00278	-0.001837	-0.00065	-6.70E-05
	EmpSD	0.027987	0.031162	0.024399	0.014263	0.004532
	MSE	0.00079	0.000981	6.00E-04	0.000204	2.10E-05
3	EmpBias	-0.002181	-0.003169	-0.00221	-0.000806	-0.000108
	EmpSD	0.029887	0.035275	0.027816	0.016116	0.005092
	MSE	0.000898	0.001254	0.000779	0.00026	2.60E-05
4	EmpBias	-0.002274	-0.002453	-0.001476	-0.000449	-9.00E-06
	EmpSD	0.03634	0.036043	0.027842	0.016668	0.005426
	MSE	0.001326	0.001305	0.000777	0.000278	2.90E-05

weight:1. $w_{ij}^{(4)}$; 2. $w_{ij}^{(4*)}$; 3. $w_{ij}^{(5)}$; and 4. $w_{ij}^{(7)}$

4.4 Water Quality Study Reanalysis

We have analyzed the water quality data in Section 3.2 under equal cluster-specific weights. In this Section, we re-analyze the water quality data allowing for unequal weights. Based on the discussion in Section 4.2 for optimal choices of weights under the sparse data situation, we re-analyze the data by assigning $w_{ij}^{(1*)} = \frac{1}{\sum_{s=1}^{p} \sum_{k=1,k\neq j}^{n_j} m_{sk}}$ to the ith stratum jth cluster for $l^{(1)}(\alpha)$ and $w_{ij}^{(4*)} = \frac{1}{(m_{ij}-1) \sum_{s=1}^{p} \sum_{k=1,k\neq j}^{n} m_{sk}(m_{sk}-1)}$ to the ith stratum jth cluster for $l^{(2)}(\beta)$ (Figure **??**, Figure 4.2, Figure 4.3). We followed the distribution assumption and estimation method in Section 3.3. Specifically, the HCGI episodes marginally followed Poisson log-linear models, $Y_{ijk} \sim \text{Poi}(\exp(\lambda_i + x_{ijk}^T \alpha + T_{ij}))$, where T_{ij} was a known offset for the person log-months on study. We used the general odds ratio model (3.4), with $d(z_{ijkl}, \beta) = z_{ijkl} \otimes \beta$, to assess intracluster pairwise associations. When computing $C_{ijkl,iuvw}$, we started with the initial pairwise density function $f(y_1, y_2) = \exp(\beta y_1 y_2/(1 + 2(y_1 - y_2)^2) - \beta(y_1^2 + y_2^2)/2)$ and conducted five steps of "marginal fittings".

The results are summarized in Table 4.6 and Table 4.7. Under assigned unequal weights, the effects of covariates on the number of HCGI episodes were very close to the results under equal weights. The household aggregation patterns were slightly different than the results under equal weights. Under an intercept-ony odds ratio model, we still concluded that HCGI episodes tended to aggregate within households. For the regression incorporating covariates, the results indicated that HCGI episodes tended to aggregate among school-age children, but was least likely to aggregate among adults. After adjusting for covariates, the intercept β_0 was significantly different zero, which suggested an unmeasured household-specific environmental effect on HCGI episodes.

The different results regarding household aggregation of HCGI using unequal weights from equal weights results might be driven by the households to which extremely large weights are assigned (Figure 4.1). Among the 84 households to which large weights are assigned, 32 of them belong to the "unmodified tap water" group. The high proportion of "unmodified tap water" group in households with large weights explains the difference results regarding aggregation pattern in this particular drinking water group using unequal versus equal weights. To reduce the influence of those extremely large weights, we consider Gaussian-based shrinkage weights, $w_{sij} = (w_{ij}^{(4*)} + \sum_{i'}\sum_{j'} w_{i'j'}/N)/2$ (Figure 4.4). Table 4.8 shows the results of covariate model under this smoothed weights. We found that HCGI episodes tended to aggregate among school-age children, but was least likely to aggregate among adults. These results suggest that there might be a tradeoff between robustness and efficiency when using unequal weights. One need to be cautious about using unequal weights when covariates are not balanced in clusters with extreme large or small weights.

Table 4.6: Fitted regression model for the marginal means

	unequal weights			equal weights		
	Estimate	*se*	p-value	Estimate	*se*	p-value
α_1: unmodified tap water	0.0809	0.1123	0.4716	0.0752	0.1131	0.5061
α_2: tap water with a purge valve	-0.0050	0.1046	0.9616	-0.0045	0.1041	0.9655
α_3: bottled plant water	-0.07726	0.1487	0.6036	-0.0767	0.1499	0.6089
ref: purified bottled water	0			0		
α_4: age 6-17	-0.5559	0.0715	<0.0001	-0.5570	0.0713	<0.0001
α_5: age > 17	-0.7227	0.0566	<0.0001	-0.7214	0.0563	<0.0001
ref: age < 6	0			0		
α_6: female	0.1602	0.0433	0.0001	0.1597	0.0431	0.0002
ref: male	0			0		

Table 4.7: Fitted regression model for intracluster association

	unequal weights			equal weights		
	Estimate	*se*	p-value	Estimate	*se*	p-value
Intercept-only model						
β_0: intercept	0.0510	0.0176	0.0038	0.0534	0.0172	0.0019
Covariate model						
β_0: intercept	0.0534	0.0094	<0.0001	0.0478	0.0090	<0.0001
β_1: unmodified tap water	0.0126	0.0032	0.0001	0.0087	0.0096	0.3648
β_2: tap water with a purge valve	0.0355	0.0217	0.1022	0.0336	0.0431	0.4354
β_3: bottled plant water	0.0175	0.0161	0.2775	0.0038	0.0480	0.9373
ref: purified bottled water	0			0		
β_4: children(<6)$-$children(6-17)	0.0041	0.0328	0.9005	0.0104	0.0492	0.8331
β_5: children(<6)$-$adult	0.0176	0.0270	0.5145	0.0171	0.0136	0.2095
β_6: children(6-17)$-$children(6-17)	0.0163	0.0008	<0.0001	0.0433	0.0166	0.0091
β_7: children(6-17)$-$adult	0.0277	0.0209	0.1847	0.0348	0.0157	0.0266
β_8: adult-adult	-0.0356	0.0111	0.0013	-0.0295	0.0095	0.0020
ref: children(<6)$-$children(<6)	0			0		

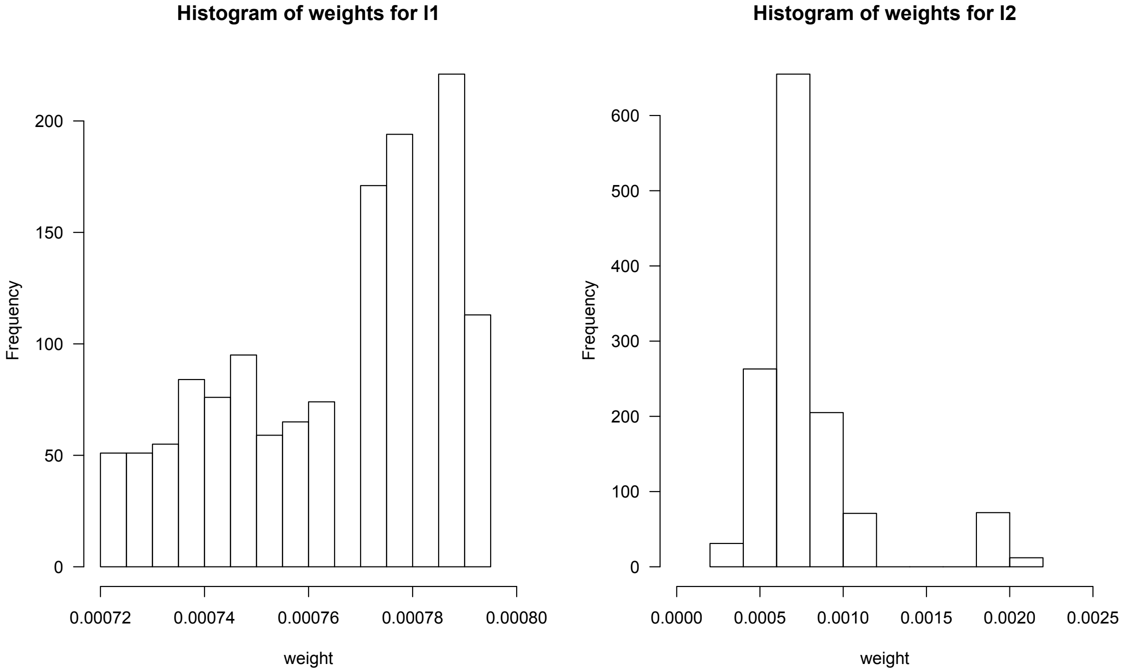

Figure 4.1

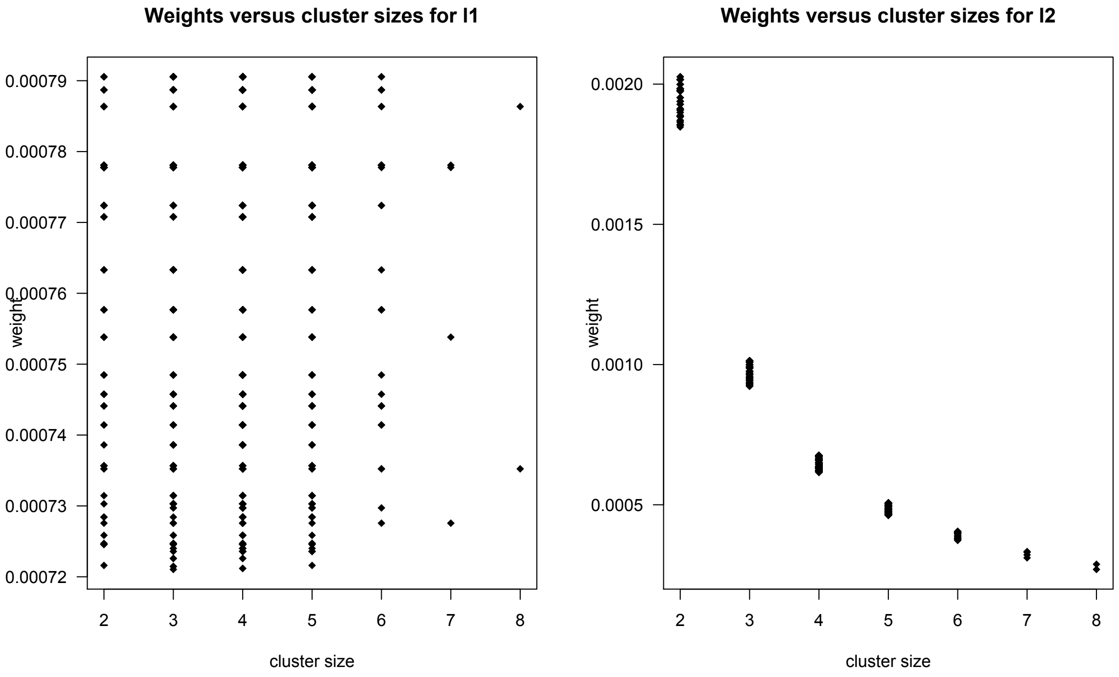

Figure 4.2

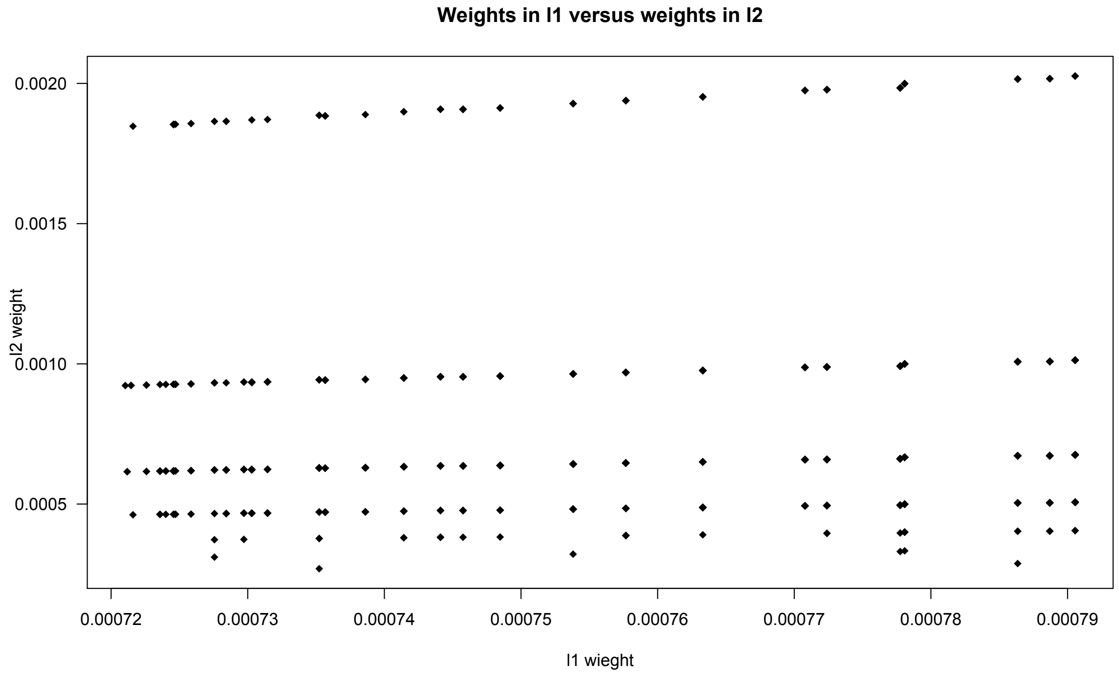

Figure 4.3

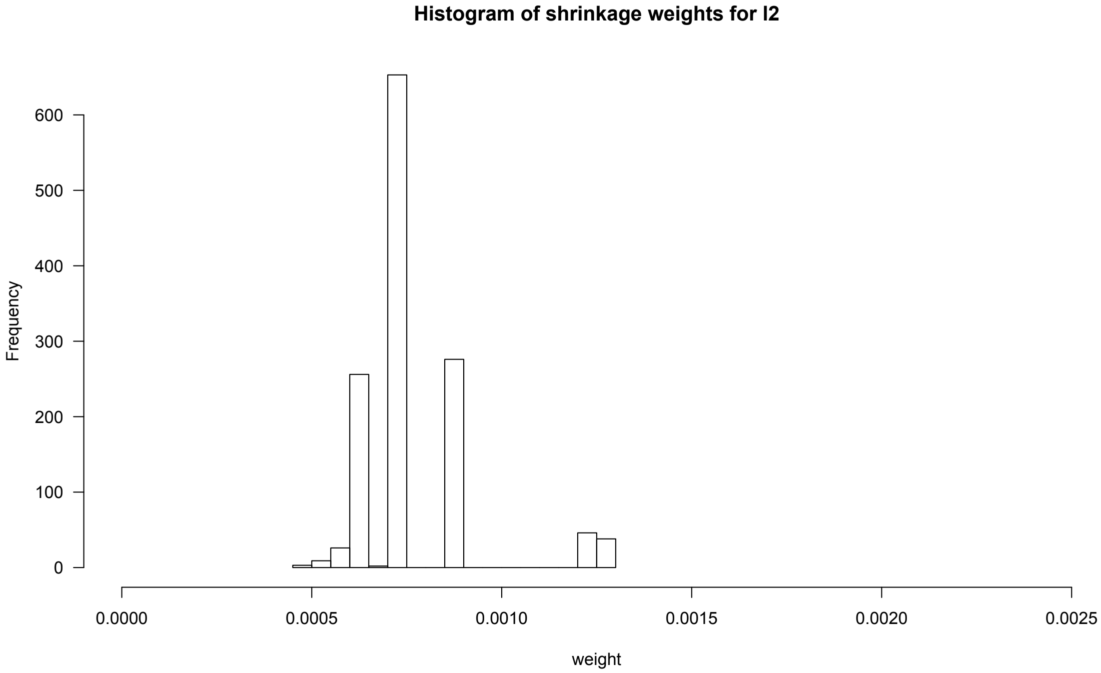

Figure 4.4

Table 4.8: Fitted covariate model for intracluster association under shrinkage weights

	Estimate	*se*	p-value
β_0: intercept	0.0459	0.0108	<0.0001
β_1: unmodified tap water	0.0098	0.0090	0.2725
β_2: tap water with a purge valve	0.0315	0.0398	0.4283
β_3: bottled plant water	0.0013	0.0092	0.8959
ref: purified bottled water	0		
β_4: children(<6)$-$children(6-17)	0.0093	0.0368	0.8016
β_5: children(<6)$-$adult	0.0150	0.0147	0.3068
β_6: children(6-17)$-$children(6-17)	0.0386	0.0132	0.0033
β_7: children(6-17)$-$adult	0.0396	0.0544	0.4675
β_8: adult-adult	-0.0267	0.0093	0.0041
ref: children(<6)$-$children(<6)	0		

4.5 Appendix

A1.

Let $\mathbf{Y}_j = (Y_{j\mathbf{1}}, \ldots, Y_{jm_j})' \sim N(\mu_j, \sigma^2\Sigma_j)$, for $j = 1, \ldots, n$, where $\mu_j = (\mu_{j1}, \ldots, \mu_{j2})'$, $\mu_{jk} = \lambda + x_{jk}^T\alpha$. Given the assumption that clusters are independent, we have

$$\begin{pmatrix} Y_{jk} \\ (Y_{jk}+Y_{uv}) \end{pmatrix} \sim N(\begin{pmatrix} \mu_{jk} \\ (\mu_{jk}+\mu_{uv}) \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}).$$

By the property of conditional multivariate normal distribution, we have

$$Y_{jk}|(Y_{jk}+Y_{uv}=a_{jk,uv}) \sim N(\frac{1}{2}(y_{jk}+y_{uv}+(x_{jk}+x_{uv})\alpha), \frac{1}{2}\sigma^2),$$

where $a_{jk,uv} = y_{jk} + y_{uv}$. Then $l^{(1)}(\alpha, \sigma^2)$ is given as:

$$\begin{aligned} l^{(1)}(\alpha,\sigma^2) =& \sum_{j<u}^{n} w_j w_u \sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv},\alpha,\sigma^2) \\ =& \sum_{j<u}^{n} w_j w_u \sum_{k=1}^{m_j}\sum_{v=1}^{m_u} [\log\frac{1}{\sqrt{\pi\sigma^2}} - \frac{(y_{jk}-y_{uv}-(x_{jk}-x_{uv})\alpha)^2}{4\sigma^2}]. \end{aligned}$$

A2.

Under the regularity condition $E[\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv},\alpha,\sigma^2)]^2 < \infty$, by Hajek projection,

$$\begin{aligned} & \sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv},\alpha,\sigma^2) \qquad (4.14) \\ \approx & \sum_{i=1}^{n} E[\sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv})|\mathbf{Y}_i] - (n-1)E[\sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv})]. \end{aligned}$$

Given the assumption that clusters are independent, equation (4.14) can be expressed as:

$$\begin{aligned}
&\sum_{i=1}^{n} E[\sum_{u>i}^{n}\sum_{k=1}^{m_i}\sum_{v=1}^{m_u} \log f(y_{ik}|Y_{ik}+Y_{uv})|\mathbf{Y}_i] \\
&+\sum_{i=1}^{n} E[\sum_{j<i}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_i} \log f(y_{hk}|Y_{jk}+Y_{iv})|\mathbf{Y}_i] \\
&+\sum_{i=1}^{n} E[\sum_{j<u,j\neq i,u\neq i}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv})] \\
&+(n-1)E[\sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv})]
\end{aligned}$$

Note that the difference of the last two term in the above equation is $-E[\sum\limits_{j<u}^{n}\sum\limits_{k=1}^{m_j}\sum\limits_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv})]$, so approximately we have

$$\begin{aligned}
&\sum_{j<u}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv},\alpha,\sigma^2) \\
&\approx \sum_{i=1}^{n}\{E[\sum_{u>i}^{n}\sum_{k=1}^{m_i}\sum_{v=1}^{m_u} \log f(y_{ik}|Y_{ik}+Y_{uv})|\mathbf{Y}_i]/2 + E[\sum_{j<i}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_i} \log f(y_{hk}|Y_{jk}+Y_{iv})|\mathbf{Y}_i]/2\}
\end{aligned}$$

A3.

It is clear to see the rational for assigning weight $w_j^{(1)} = \frac{1}{M-m_j}$, where $M = \sum\limits_{j=1}^{n} m_j$, to the jth cluster via multivariate normal distribution. More generally, this choice of weights, based on Hajek projection, does not actually depend on the multivariate normal assumption. As shown in Section 4.2, $l^{(1)}(\alpha,\sigma^2)$ is approximated as:

$$\begin{aligned}
&l^{(1)}(\alpha,\sigma^2) = \sum_{j<u}^{n} w_j w_u \sum_{k=1}^{m_j}\sum_{v=1}^{m_u} \log f(y_{jk}|Y_{jk}+Y_{uv},\alpha,\sigma^2) \\
&\approx \sum_{i=1}^{n} w_i\{E[\sum_{u>i}^{n}\sum_{k=1}^{m_i}\sum_{v=1}^{m_u} \log f(y_{ik}|Y_{ik}+Y_{uv})|\mathbf{Y}_i]/2 + E[\sum_{j<i}^{n}\sum_{k=1}^{m_j}\sum_{v=1}^{m_i} \log f(y_{jk}|Y_{jk}+Y_{iv})|\mathbf{Y}_i]/2\}
\end{aligned}$$

The conditional expectation $E[\sum\limits_{u>i}^{n}\sum\limits_{k=1}^{m_i}\sum\limits_{v=1}^{m_u}\log f(y_{ik}|Y_{ik}+Y_{uv})|\mathbf{Y}_i]$ depends on the observed data in the ith cluster only, so it can be further simplified as $E[\sum\limits_{u>i}^{n} m_u \sum\limits_{k=1}^{m_i}\log f(y_{ik}|Y_{ik}+Y_{u1})|\mathbf{Y}_i]$. Similarly, $E[\sum\limits_{j<i}^{n}\sum\limits_{k=1}^{m_j}\sum\limits_{v=1}^{m_i}\log f(y_{jk}|Y_{jk}+Y_{iv})|\mathbf{Y}_i]$ can be expressed as $E[\sum\limits_{j<i}^{n} m_j \sum\limits_{v=1}^{m_i}\log f(y_{j1}|Y_{j1}+Y_{iv})|\mathbf{Y}_i]$. Therefore, the above approximation of $l^{(1)}(\alpha,\sigma^2)$ can be expressed as:

$$l^{(1)}(\alpha,\sigma^2) \approx \sum_{i=1}^{n} w_i\{\sum_{j=1,j\neq i}^{n} m_j E[\sum_{v=1}^{m_i}\log f(y_{j1}|Y_{j1}+Y_{iv})|\mathbf{Y}_i]/2\}. \tag{4.15}$$

The closed form of this approximation can be obtained under particular distribution assumptions of the outcome. When we consider cluster-specific weights that only depend on cluster sizes, we ignore the distribution assumption of the outcome, so it is natural that we assign weight $w_j^{(1)} = \frac{1}{M-m_j}$, where $M = \sum\limits_{j=1}^{n} m_j$, to the jth cluster such that the term $w_i \sum\limits_{j=1,j\neq i}^{n} m_j$ in the right hand side of equation (4.15) cancels out. Therefore, we can still consider this choice of weights when the outcome is not normally distributed

A4.

Under the regularity condition $E[\sum\limits_{k\neq l}^{m_j}\sum\limits_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk}, Y_{uv}=y_{uv}, Y_{jl}+Y_{uw}=$

$a_{jl,uw};\rho)]^2 < \infty$, by Hajek projection,

$$
\begin{aligned}
&\sum_{j<u}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{uv}=y_{uv},Y_{jl}+Y_{uw}=a_{jl,uw};\rho)\\
\approx&\sum_{i=1}^{n}E[\sum_{j<u}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{uv}=y_{uv},Y_{jl}+Y_{uw}=a_{jl,uw})|\mathbf{Y}_i]\\
&-(n-1)E[\sum_{j<u}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{uv}=y_{uv},Y_{jl}+Y_{uw}=a_{jl,uw})]\\
=&\sum_{i=1}^{n}E[\sum_{u>i}^{n}\sum_{k\neq l}^{m_i}\sum_{v\neq w}^{m_u}\log\, f(y_{il}|Y_{ik}=y_{ik},Y_{uv}=y_{uv},Y_{il}+Y_{uw}=a_{il,uw})|\mathbf{Y}_i]\\
&+\sum_{i=1}^{n}E[\sum_{j<i}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_i}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{iv}=y_{iv},Y_{jl}+Y_{iw}=a_{jl,iw})|\mathbf{Y}_i]\\
&+\sum_{i=1}^{n}E[\sum_{j<u,j\neq i,u\neq i}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{uv}=y_{uv},Y_{jl}+Y_{uw}=a_{jl,uw})]\\
&-(n-1)E[\sum_{j<u}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{uv}=y_{uv},Y_{jl}+Y_{uw}=a_{jl,uw})].
\end{aligned}
$$

Note that the difference of the last two term in the above equation is $-E[\sum\limits_{j<u}^{n}\sum\limits_{k\neq l}^{m_j}\sum\limits_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{uv}=y_{uv},Y_{jl}+Y_{uw}=a_{jl,uw})]$, so approximately we have

$$
\begin{aligned}
&\sum_{j<u}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_u}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{uv}=y_{uv},Y_{jl}+Y_{uw}=a_{jl,uw};\rho)\\
\approx&\sum_{i=1}^{n}\{E[\sum_{u>i}^{n}\sum_{k\neq l}^{m_i}\sum_{v\neq w}^{m_u}\log\, f(y_{il}|Y_{ik}=y_{ik},Y_{uv}=y_{uv},Y_{il}+Y_{uw}=a_{il,uw})|\mathbf{Y}_i]/2\\
&+E[\sum_{j<i}^{n}\sum_{k\neq l}^{m_j}\sum_{v\neq w}^{m_i}\log\, f(y_{jl}|Y_{jk}=y_{jk},Y_{iv}=y_{iv},Y_{jl}+Y_{iw}=a_{jl,iw})|\mathbf{Y}_i]/2\}
\end{aligned}
$$

Then we can approximate $l^{(2)}(\rho)$ as:

$$\begin{aligned} l^{(2)}(\rho) =& \sum_{j<u}^{n} w_j w_u \sum_{k\neq l}^{m_j} \sum_{v\neq w}^{m_u} \log\, f(y_{jl}|Y_{jk}=y_{jk}, Y_{uv}=y_{uv}, Y_{jl}+Y_{uw}=a_{jl,uw};\rho) \\ \approx& \sum_{i=1}^{n} w_i \{E[\sum_{u>i}^{n} \sum_{k\neq l}^{m_i} \sum_{v\neq w}^{m_u} \log\, f(y_{il}|Y_{ik}=y_{ik}, Y_{uv}=y_{uv}, Y_{il}+Y_{uw}=a_{il,uw})|\mathbf{Y}_i]/2 \\ &+ E[\sum_{j<i}^{n} \sum_{k\neq l}^{m_j} \sum_{v\neq w}^{m_i} \log\, f(y_{jl}|Y_{jk}=y_{jk}, Y_{iv}=y_{iv}, Y_{jl}+Y_{iw}=a_{jl,iw})|\mathbf{Y}_i]/2\} \end{aligned}$$

A5.

Let $\mathbf{Y}_j = (Y_{j\mathbf{1}}, \ldots, Y_{jm_j})' \sim N(\mu_j, \sigma^2\Sigma_j)$, for $j = 1, \ldots, n$, where $\mu_j = (\mu_{j1}, \ldots, \mu_{j2})'$, $\mu_{jk} = \lambda + x_{jk}^T\alpha$, and Σ_j has an exchangeable structure. When there is no covariates, $\mu_j = \lambda\mathbf{1}$. Then we have

$$Y_{jl}|(Y_{jk}, Y_{uv}, Y_{jl}+Y_{uw}=a_{jl,uw}) \sim N(\mu_{jkl,uvw}, \tau^2),$$

where $\mu_{jkl,uvw} = \frac{1}{2}\rho(y_{jk} - y_{uv}) + \frac{1}{2}a_{jl,uw}$, and $\tau^2 = \frac{1}{2}(1-\rho^2)\sigma^2$. Therefore, after conducting Hajek projection, $l^{(2)}(\rho)$ can be approximated as:

$$l^{(2)}(\rho) \approx \sum_{i=1}^{n} w_i \sum_{j=1,j\neq i}^{n} m_j(m_j-1)\{\sum_{k\neq l}^{m_i}[-\frac{1}{2}\log\frac{2\pi\sigma^2(1-\rho^2)}{2} - \frac{(y_{il}-y_{ik})^2 - 2\rho\sigma^2}{4(1-\rho^2)\sigma^2}]\} \tag{4.16}$$

Chapter 5

Summary and Future Direction of Study

5.1 Summary

Our composite conditional likelihood approach is insensitive to nuisance parameters, and provides robust and flexible inference for sparse clustered data. One limitation, however, is that the marginal univariate distributions must follow a generalized linear model. The general odds ratio function (1.5) can be used to assess the intracluster pairwise associations and is particularly well-suited for use with the composite conditional likelihood approach; this measure of pairwise association has several attractive features: it accommodoates responses of any type, is invariant under prospective or retrospective proband-based sampling design, is unconstrained by the marginal univariate distributions of the responses, and it completely characterizes the pairwise association. Specifically, we have proposed the odds ratio model (3.4) for use when responses are all on the same scale, and we desire to investigate the familial aggregation patterns of disease. Moreover, as an exploratory study, we investigate the optimal choices of cluster-specific weights for our proposed composite conditional likelihood. Under our choices of weights, the efficiency of estimation improves.

5.2 Future Direction of Study

The third aim of my dissertation is an exploratory study, and hence needs more future work. One future direction of study is to further investigate the choices of cluster-specific weights which are allowed to depend on cluster sizes as well as correlation structure and covariates. Moreover, this exploratory study on selection of weights is under the i.i.d cluster size assumption, and another future direction of study is to investigate optimal choices of weights when cluster sizes are not identically distributed.

Furthermore, my dissertation work on composite conditional likelihood focuses on two asymptotic schemes: standard situation and sparse data situation. Future work could be done on the composite conditional likelihood under the rectangular array asymptotic scheme.

In my dissertation, we assume that individual outcomes marginally are distributed according to a generalized linear model, and we model the general odds ratio function. One future direction of study is to investigate if there is a way to relax the generalized linear model assumption on marginal outcomes and model the general odds ratio function only.

Bibliography

Brent, R. P. (1973). *Algorithms for minimization without derivatives.* Courier Dover Publications.

Chen, H. Y. (2003). A note on the prospective analysis of outcome-dependent samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65,** 575–584.

Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association* **99,** 1176–1189.

Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics* **63,** 413–421.

Datta, S. and Satten, G. A. (2008). A signed-rank test for clustered data. *Biometrics* **64,** 501–507.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* **31,** 1208–1211.

Hanfelt, J. J. (2004). Composite conditional likelihood for sparse clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66,** 259–273.

Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis* **100,** 670–685.

Lakshminarayana, J., Pandit, S., and Srinivasa Rao, K. (1999). On a bivariate poisson distribution. *Communications in Statistics-Theory and Methods* **28,** 267–276.

Le Cessie, S. and Van Houwelingen, J. (1994). Logistic regression for correlated binary data. *Applied Statistics* pages 95–108.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 3–40.

Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **15,** 211–225.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80,** 221–39.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica* pages 71–105.

Madsen, L. and Dalthorp, D. (2007). Simulating correlated count data. *Environmental and Ecological Statistics* **14,** 129–148.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society* pages 1–32.

Osius, G. (2004). The association between two random elements: A complete characterization and odds ratio models. *Metrika* **60,** 261–277.

Payment, P., Siemiatycki, J., Richardson, L., Renaud, G., Franco, E., and Prevost, M. (1997). A prospective epidemiological study of gastrointestinal health effects due to the consumption of drinking water. *International Journal of Environmental Health Research* **7,** 5–31.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Van Der Linde, A. (2003). Dimension reduction with linear discriminant functions based on an odds ratio parameterization. *International statistical review* **71,** 629–666.

Van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory* **15,** 228–237.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* pages 5–42.

Yeo, I.-K. and Johnson, R. A. (2001). A uniform strong law of large numbers for¡ i¿ u¡/i¿-statistics with application to transforming to near symmetry. *Statistics & probability letters* **51,** 63–69.