The Pennsylvania State University
The Graduate School
Department of Statistics

COMPOSITE LIKELIHOOD IN LONG SEQUENCE DATA

A Dissertation in
Statistics

by

Jianping Sun

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2011

The dissertation of Jianping Sun was reviewed and approved* by the following:

Bruce Lindsay
Willaman Professor of Statistics
Dissertation Advisor
Chair of Committee

Debashis Ghosh
Professor of Statistics

Bing Li
Professor of Statistics

Rongling Wu
Professor of Public Health Sciences

Runze Li
Professor of Statistics
Graduate Chair of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Composite Likelihood in long sequence data

The primary motivation of this thesis is to develop richer models and statistical methodologies for sequence data now generated in biology, such as SNP data. Suppose we have observed n current descendant sequences of length L. One interesting question is how to estimate the unknown ancestral distribution from the observed descendants, considering realistic biology complexities such as mutation and recombination.

We developed a statistical model by extending the ancestor mixture model (Chen and Lindsay (2006)) with both mutation and recombination to estimate the ancestral distribution. However, though we can write out the full likelihood for the ancestral distribution explicitly, there is an enormous computation challenge in the actual computation due to an enormous number of recombination possibilities. This number grows exponentially in sequence length. Therefore, we proposed to develop composite likelihood as a methodology that makes the computation feasible.

In this thesis, we first introduce our developed statistical model. We then discuss issues in the construction of composite likelihoods from the view of statistical efficiency. Then, a Markov chain composite likelihood (MCCL) method is proposed and applied to our statistical model, now considering only the more challenging recombination factor. Two estimators, a left to right estimator and a hierarchical estimator, are developed consequently to estimate both marginal and

joint ancestral distributions through MCCL. Finally, some simulation results are shown to investigate the performance of two estimators.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to express my deepest gratitude to my advisor, Prof. Bruce G. Lindsay, whose inspiration, guidance, patience and support from the initial to the final level enabled me to develop an understanding of the subject. He spent a lot of his precious time guiding my research work, without which this thesis could not come into existence. His advice, encouragement, and research attitude also direct my future career.

I would also like to send my thanks to my committee members, Prof.Debashis Ghosh, Prof.Bing Li, and Prof.Rongling Wu, for their invaluable feed backs and suggestions during this thesis development. Thanks go to all the faculty members in the Statistics Department at Penn State University as well, for the greatest education I received, especially to Prof. Naomi Altman, for her helpful advice and support of my future career.

In addition, I would like to take this opportunity to heartily gratitude to my family. Thanks to my parents, Youbin Sun, Baoyan Fu, and my husband, Xianming Tan. Their unconditional love and kindness support me through to the end of this thesis.

Finally, I need to say thanks to my friends, Lu Zhang, Zhe Chen, Yijia Feng, Hui-Wen Teng, and Tao Yang. They made my experience at Penn State memorable. Special thanks to my fellow grad, Xianyun Mao, for the frequent discussion on both statistical and genetic problems related with this thesis.

Chapter 1

MOTIVATION

The goals of this thesis are to develop richer models and statistical methodologies for sequence data. The primary motivation will be the analysis of the types of sequence data now generated in biology. One example of biological sequence data is SNP data. A single nucleotide polymorphism (SNP) is a DNA sequence variation in which a single nucleotide (A, T, C or G) base differs from the usual base at a specific position on the DNA. Variations in the human DNA sequences can affect the development of diseases, and SNPs can help determine the likelihood that people will develop a particular disease. Scientists believe that mapping out these SNPs will help people identify the many genes associated complex diseases, such as cancer and diabetes.

In addition, after the building of the map of SNPs from human genome, scientists can also construct hierarchical trees from the map of SNPs. Hierarchical trees play a critical role in genome research. For example, it helps people finding the evolutionary history, identifying and mapping the potential genome which is susceptible to cause disease.

In this chapter, I will introduce the mathematical model used to construct a hierarchical tree in Chen and Lindsay (2006), and then a model extension will be discussed thereafter.

1.1 Ancestral Mixture Model

In the majority of SNP data, only two possible character states exist at each of the positions. That is, one finds A and G at a purine site or T and C at a pyrimidine site. Thus, we can use binary sequences to represent SNP data. We will use $\{0, 1\}$ coding for these two states. In Chen and Lindsay (2006), the authors developed a novel mathematical model, the ancestral mixture model, to building hierarchical tree from binary sequence data. The model is summarized as follows.

The simplest model has one ancestor vector $\mu \equiv (\mu_1, ..., \mu_L)$, and one observation vector $X \equiv (X_1, ..., X_L)$, where both X_s and μ_s are binary. Suppose only mutation is considered. Let $x = (x_1, x_2, ..., x_L)$ be an observed binary sequence of length L, where the variable x_s at site s is generated from an unknown ancestor $\mu_s \in \{0, 1\}$ with *mutation probability* p, where

$$p = P(x_s \neq \mu_s) = 1 - P(x_s = \mu_s)$$

The occurrence of mutation on each site is assumed to be independent, so that the observation $x = (x_1, x_2, ..., x_L)$ can be modeled as independent Bernoulli trials. As a result, the density of x can be written as

$$\begin{aligned} \kappa(x|\mu, p) &= \Pi_{s=1}^{L} p^{(x_s-\mu_s)^2}(1-p)^{1-(x_s-\mu_s)^2} \\ &= p^{D(x,\mu)}(1-p)^{L-D(x,\mu)} \end{aligned}$$

where

$$D(x, \mu) = \sum_{s=1}^{L}(x_s - \mu_s)^2$$

is the number of disagreements between the elements of x and the elements of μ (We are using the fact that $x_s = \mu_s$ if $(x_s - \mu_s)^2 = 0$ and $x_s \neq \mu_s$ if $(x_s - \mu_s)^2 = 1$). The function $\kappa(x|\mu, p)$ is called mutation kernel. The above model is called as single ancestral model with the unknown parameter μ as the ancestral sequence.

The single ancestral model can be extended to an ancestral mixture model as following. Suppose a random variable ϑ takes value from the μ-parameter space under an *ancestral distribution* Q, where Q is a discrete distribution with k support points $\mu_1, \mu_2, ..., \mu_k$ and corresponding probabilities

$$P(\vartheta = \mu_i) = \pi_i$$

where $\pi_i \geq 0$ and $\sum_{i=1}^{k} \pi_i = 1$.

The descendant sequence, a random vector X_i, is generated hierarchically by first generating an ancestor sequence $\vartheta = \mu_i$ from Q, then generating $X_i = x_i$ from $\kappa(x|\mu, p)$. Here ϑ is unobserved, and so is called a *latent variable*. The descendant sequence X is said to follow the *ancestral mixture model*, denoted by $X \sim A(Q, p)$ if it is generated in this scheme. If so the density of X is

$$f(x; Q, p) = \sum_{i=1}^{k} \pi_i \kappa(x|\mu_i, p) \tag{1.1}$$

Just as in the normal mixture model, the ancestral mixture model has nested structure. By this we mean that for the normal mixture model, any mixture $N(Q, \sigma^2)$ can be represented as another normal mixture by $N(Q^*, \sigma^2 - \sigma_1^2)$, where Q^* is the convolution of Q and $N(0, \sigma_1^2)$ (Lindsay 1995). Similarly, in ancestral mixture model, suppose an ancestor μ goes through T generation of mutation, and the mutation rate for each generation is fixed at $p^* = 0.5 - r$. Then, the distribution of descendant X after T generation is equivalent to the distribution of descendant only after one generation but with mutation ration as $p = \frac{1}{2} - \frac{1}{2}(2r)^T$. That is,

$$X \sim A(Q, p)$$

where $-\log(1-2p) = T \times [-\log(1-2p^*)]$. Hence, $\eta = -\log(1-2p)$ is called time parameter.

The parameters (Q, p) in the ancestral mixture model are not jointly identifiable. However, they are identifiable in $\Omega = \mathcal{P} \times \{p\}$, where p is any fixed value in $[0, \frac{1}{2})$.

To estimate the unknown parameters π and μ, Chen and Lindsay (2006) developed the nonparametric maximum likelihood method. At each fixed value p, they used K-component Expectation Maximization (EM) algorithm to find the MLE of Q. They estimated the MLE of Q on a grid of p values, starting at $p = 0$. When $p = 0$, then no mutations exist in the model. Hence, the distinct sequences in current samples become the estimated ancestors for these samples. As p increases, i.e. looking back further in time to identify the ancestors, we can use the estimated ancestors at p as starting values in an EM algorithm which is used to updated the ancestors at new p.

At each stage, it is possible that some of the estimated ancestors at the previous p merge into one sequence under the new p. Thus, the number of support points of Q decreases as p increases. As p increases to $1/2$, the mutation model looks just like a fair coin toss. In fact, the number of support points of Q decreases to 1 as p becomes close to $1/2$. Hence, those various estimators of Q at different value of p, or at different passing time, form a tree structure. The single sequence at the top of the tree is simply the "majority rule" ancestor, where the most common allele at each site is used.

Chen and Lindsay (2006) used the mitochondrial DNA sequences of Griffiths and Tavaré (1991) as an example to illustrate the performance of the method.

1.2 Extension of Ancestral Mixture Model

The ancestral mixture model in previous section is designed for binary sequences that contain only mutations. However, to deal with the existence of genetic complexities, other realistic biological factors, for example, recombination, could

be added into the model. In this section, we will consider how to extend the ancestral mixture model to include other evolutionary factors.

Firstly, notice that the mutation kernel can also be described as diffusion kernel from a continuous time Markov Chain (CTMC). To see this, let the binary observation $X = X(\eta)$ arise from a CTMC with state space $\{0, 1\}$, and let the ancestor μ be $X(0)$. If the rates of transition from 0 to 1 and 1 to 0 are equal to $\frac{1}{2}$. Then the probability transition matrix in η time units is

$$P_\eta = \begin{pmatrix} \frac{1+\exp(-\eta)}{2} & \frac{1-\exp(-\eta)}{2} \\ \frac{1-\exp(-\eta)}{2} & \frac{1+\exp(-\eta)}{2} \end{pmatrix}. \tag{1.2}$$

Hence, the probability of mutation, which is the same as probabilities of going from state 0 to state 1 or from state 1 to state 0 in η time units, is $p = \frac{1-\exp(-\eta)}{2}$. Thus, the mutation kernel, $\kappa(x|\mu, \eta)$, exactly represents the probability of transiting from ancestor μ to descendant x in η time units.

The following discussion about the model extension will be based on this framework.

1.2.1 Recombination sequences with length 2

For the reason of clarity, let's first consider recombination in sequences with length as 2. In such situation, there are four possible types of ancestors: $\mu_1 = (0, 0)$, $\mu_2 = (0, 1)$, $\mu_3 = (1, 0)$, and $\mu_4 = (1, 1)$. We let the population frequency for the four types of ancestors be π_1, π_2, π_3, and π_4 (for convenience, these weights can also be denoted as $\pi_{0,0}$, $\pi_{0,1}$, $\pi_{1,0}$, and $\pi_{1,1}$). The sum of weights is 1. The observed descendants also include the above four types of sequences with counts n_1, n_2, n_3, and n_4 (for convenience, these counts can also be denoted as $n_{0,0}$, $n_{0,1}$, $n_{1,0}$, and $n_{1,1}$). The sum of all the counts is equal to n, the number of observations.

Firstly, we build a model for recombination. Let (A_1, A_2) be independent draws from the ancestor distribution with $P(A_1 = k) = \pi_k$, $k = 1, 2, 3, 4$. Suppose

the probability of recombination occurring between site 1 and site 2 in η time units is q, where $q = 1 - e^{-\eta}$. Let $R = 1$ if a recombination occurs. Given $(A_1 = k, A_2 = l, R = 1)$, define the realized ancestral sequence to be $(B_1 = k, B_2 = l)$. Given $(A_1 = k, A_2 = l, R = 0)$, let the realized ancestral sequence to be $(B_1 = k, B_2 = k)$. That is, if $R = 0$ (no recombination), then the ancestor for site 2 is the one chosen for site 1. If $R = 1$, we choose a new ancestor for site 2. We specify further that

$$\begin{aligned} P(X = x|A_1, A_2, R) &= P(X = x|B_1, B_2) \\ &= \mathbb{I}\{X(1) = \mu_{B_1}(1)\}\mathbb{I}\{X(2) = \mu_{B_2}(2)\} \end{aligned}$$

where $\mathbb{I}$ is an indicator function. That is, the realized X sequence has at each site the value of the corresponding realized ancestor. Then the probability of $X = (i, j)$, where i and j equal 0 or 1, is

$$\begin{aligned} P(X = (i,j)) &= P(X = (i,j)|R = 0)P(R = 0) + P(X = (i,j)|R = 1)P(R = 1) \\ &= (1 - q)\pi_{i,j} + q\pi_{i,+}\pi_{+,j} \end{aligned}$$

where $\pi_{i,+} = \pi_{i,0} + \pi_{i,1}$ and $\pi_{+,j} = \pi_{0,j} + \pi_{1,j}$. That is, in this model, if a recombination occurs, the ancestor at site 1 and 2 are drawn independently from the ancestral model. This occurs with probability q. If recombination does not occur, then both sites are drawn from the same ancestor. Thus, the log-likelihood for the observed data is $\sum_{i=0}^{1}\sum_{j=0}^{1} n_{i,j} \log P(X = (i, j))$.

We can easily add mutation into the above recombination model. We can consider the observed sequence, Y, to be a mutated version of X, where X is the sequence generated from ancestors in η time units with only potential recombination. Thus, the probability that we observe $Y = (i, j)$, where i and j equal 0 or 1, is

$$\begin{aligned} P(Y = (i,j)) = \sum_{k=1}^{4}\sum_{l=1}^{4}\sum_{r=0}^{1} \quad & [P(Y = (i,j)|A(1) = k, A(2) = l, R = r) \\ & \times P(A(1) = k)P(A(2) = l)P(R = r)] \end{aligned}$$

where $A(1)$ and $A(2)$ represent the hypothetical ancestors for site 1 and site 2, $k, l \in \{1, 2, 3, 4\}$, and R equals 1 or 0, indicating there is recombination or not respectively. Hence, $P(A(i) = l) = \pi_l$, and $P(R = 1) = 1 - P(R = 0) = q$.

If $R = 1$, the ancestors for two sites of current observed Y are different, and $P(Y = (i, j)|A(1) = k, A(2) = l, R = 1) = \kappa((i, j)|(\mu_k(1), \mu_l(2)), \eta)$. Otherwise, the ancestors for two sites of the current observation are the same, and $P(Y = (i, j)|A(1) = k, A(2) = l, R = 0) = \kappa((i, j)|(\mu_k(1), \mu_k(2)), \eta)$.

1.2.2 Recombination sequences with length L

The same idea can be applied to more general sequences with length L and generated from K different unknown ancestors, $\mu_1, ..., \mu_K$. The corresponding weights for the ancestors are $\pi_1, ..., \pi_K$.

Let $A(i) = l$ the ancestor at site i is μ_l, where $i = 1, ..., L$ and $l \in 1, ..., K$. Then $P(A(i) = l) = \pi_l$. Let R_j, $j = 1, ..., L-1$, be the results of $L-1$ independent Bernoulli trials with probability q, where $R_j = 1$ means there is recombination between site j and site $j + 1$. Then, $P(R_j = 1) = 1 - P(R_j = 0) = q = 1 - e^{-\eta}$ is just the probability of recombination between site j and site $j + 1$ in η time units. (The model can easily be modified to let q depends on j.)

Suppose the realized ancestor sequence, $(B_1, ..., B_L)$, depends on the recombination indicator R and the hypothetical ancestor sequence $A_1, ..., A_L$ as follows. First, $B_1 = A_1$. Then $B_2 = A_1$ if $R_1 = 0$, and $B_2 = A_2$ if $R_1 = 1$. To determine B_3, let $B_3 = B_2$ if $R_2 = 0$, and $B_3 = A_3$ else. Continue in this fashion, let $B_k = B_{k-1}$ if $R_{k-1} = 0$, and $B_k = A_k$ else. The probability $P(A(1) = i_1, ..., A(L) = i_L) = \prod_{s=1}^{L} \pi_{i_s}$ is the probability of independently drawing the corresponding ancestor sequence, and $P(R_1 = r_1, ..., R_{L-1} = r_{L-1}) = \prod_{s=1}^{L-1} q^{r_s}(1-q)^{1-r_s}$ is the probability of independent Bernoulli trials with success probability (i.e. recombination rate)

q. Similar to $L = 2$, we can write

$$\begin{aligned} & P(Y = (y_1, ..., y_L)) \\ = & \sum_{i_1,...,i_L} \sum_{r_1,...,r_{L-1}} [P(Y = (y_1, ..., y_L) | A(1) = i_1, ..., A(L) = i_L, R_1 = r_1, ..., R_{L-1} = r_{L-1}) \\ & \times P(A(1) = i_1, ..., A(L) = i_L) P(R_1 = r_1, ..., R_{L-1} = r_{L-1})], \end{aligned}$$

where the first factor in the summand is

$$\begin{aligned} & P(Y = (y_1, ..., y_L) | A(1) = i_1, ..., A(L) = i_L, R_1 = r_1, ..., R_{L-1} = r_{L-1}) \\ = & \kappa((y_1, ..., y_L) \mid (\mu_{B_1}(1), ..., \mu_{B_L}(L)), \eta). \end{aligned}$$

Although we can write the above model with both mutation and recombination explicitly, there are significant computation challenges when applying it on data due to a enormous number of recombination possibilities when L is large. Note there are 2^{L-1} possible recombination sequences $B_1, ..., B_{L-1}$.

To illustrate the challenge, a simulation result by Yijia Feng (Table 1.1) showed that it take thousands of steps for the EM algorithm to converge even for the simple case $L = 3$.

The above computation challenge calls out for some other methods to reduce the complexity. Composite likelihood is the possibility we will explore in this thesis. We also plan to come to a better understanding of this model more generally.

Table 1.1: Binary sequence with length 3, mutation rate $p = \frac{1}{2}(1 - e^{-\eta})$, and recombination rate $q = 1 - e^{-\eta}$.

$\pi = (0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)$	
passing time η=1, and $n = 10000$	
Data	(1266,1274,1221,1244,1247,1256,1220,1272)
π^0	(0.125,0.125,0.125,0.125,0.125,0.125,0.125,0.125)
$\hat{\pi}$	(0.1465,0.1361,0.1043,0.1204,0.1215,0.1135,0.1152,0.1484)
Steps	1176
π^0	(0.05,0.05,0.2,0.2,0.1,0.1,0.15,0.15)
$\hat{\pi}$	(0.1491,0.1238,0.1055,0.1229,0.1252,0.1136,0.1077,0.1522)
Steps	2134
$\pi = (0.05, 0.1, 0.05, 0.3, 0.3, 0.05, 0.1, 0.05)$	
Data	(1228,1218,1245,1355,1290,1195,1300,1169)
π^0	(0.125,0.125,0.125,0.125,0.125,0.125,0.125,0.125)
$\hat{\pi}$	(0.0335,0.1391,0.0403,0.2991,0.2693,0.0393,0.1736,0.0058)
Steps	7666
π^0	(0.05,0.05,0.2,0.2,0.1,0.1,0.15,0.15)
$\hat{\pi}$	(0.0331,0.1391,0.0408,0.2989,0.2694,0.0395,0.1734,0.0057)
Steps	7679

Chapter 2

Introduction to Composite Likelihood

Maximum likelihood is a popular statistical method that plays a central role in the formal theory of statistical inference because it provides estimators with optimal statistical efficiency. However, many realistic statistical models are so complex in structure that it becomes computationally infeasible to find the MLE, especially in large data set. In such a situation, some convenient surrogates for ordinary likelihood are valuable. Bayesian strategies depend on the likelihood as well. Alternative approaches based on likelihood modifications have been developed by several authors, for example, the pseudo likelihood suggested by Besag (1974) and the partial likelihood advocated by Cox (1975). In this chapter, we will introduce composite likelihood method which is a generalization of pseudo and partial likelihood.

2.1 Definition

2.1.1 Notation and framework

Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_d)^T$ be a d-dimensional random vector with the probability model $f(\mathbf{y}; \theta)$, where θ is a parameter taking values in the parameter space $\boldsymbol{\Theta}$ that is a subset of a Euclidean space of dimension p. Then, the likelihood function is defined as $L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta)$, and the associated true score function is $U_{mle}(\theta; \mathbf{y}) = \nabla \log f(\mathbf{y}; \theta)$. If we observed n independent random samples, $\mathbf{y}_1, ..., \mathbf{y}_n$, from $f(\mathbf{y}; \theta)$, then the likelihood function and score function based on

samples are denoted as $L(\theta; \mathbf{y}_1, ..., \mathbf{y}_n) = f(\mathbf{y}_1, ..., \mathbf{y}_n; \theta)$ and $U_{mle}(\theta; \mathbf{y}_1, ..., \mathbf{y}_n) = \nabla \log f(\mathbf{y}_1, ..., \mathbf{y}_n; \theta)$.

For convenience, we also define the following notations, which will be used in the whole thesis. Let $f_i(y_i; \theta)$, $f_{ij}(y_i, y_j; \theta)$, and $f_{i|j}(y_i|y_j; \theta)$ denote marginal density for Y_i, pairwise density for $(Y_i, Y_j)^T$, and conditional density for Y_i given Y_j, respectively, where $i, j \in (1, 2, ..., d)$ and $i \neq j$. Associated likelihood functions are defined as $L_i = f_i(y_i; \theta)$, $L_{ij} = f_{ij}(y_i, y_j; \theta)$, and $L_{i|j} = f_{i|j}(y_i|y_j; \theta)$. Similarly, we use $U_i = \nabla \log L_i$, $U_{ij} = \nabla \log L_{ij}$, and $U_{i|j} = \nabla \log L_{i|j}$ to denote the corresponding score functions.

Moreover, three types of marginal likelihood functions and associated score functions, which play key roles in this proposal, are defined as following.

- Onewise marginal likelihood and score: $\prod_i L_i$ and $\sum_i U_i$
  Note that this likelihood corresponds to treat the data as independent.
- Pairwise marginal likelihood and score: $\prod_{i<j} L_{ij}$ and $\sum_{i<j} U_{ij}$
- All pairwise conditional likelihoods and score: $\prod_{i \neq j} L_{i|j}$ and $\sum_{i \neq j} U_{i|j}$

2.1.2 Definition for composite likelihood

Composite likelihood was first defined by Lindsay (1988) as the product of any set of conditional or marginal likelihoods. More precisely, Let s be a be an index representing a marginal or conditional event, let L_s be the likelihood corresponding to the variables indexed by s, and U_s be the resulting score function. Lindsay (1988) proposed the composite likelihood (CL), defined as

$$\prod_s L_s, \tag{2.1}$$

where the product can be evaluated over a an arbitrary collection of indices s. It is clear that the availability of the full likelihood ensures the formulation of (2.1).

However, knowledge of (2.1) may not necessarily recover the full distribution. It is important to note that we are not restricting $Y_1, ..., Y_d$ to be independent variables. In most important applications they are not.

In parallel with likelihood, Lindsay (1988) also defined the associated composite scores for the composite likelihood as

$$CS(\theta) = \sum_s U_s$$

where U_s is called as the component score.

A generalization of this definition would allow arbitrary (but fixed) non-negative weights w_s in the definition.

$$\prod_s L_s^{w_s}, \tag{2.2}$$

where $w_s \geq 0$. We will call this *weighted composite likelihood*.

The previously defined onewise, pairwise and conditional pairwise likelihoods are all example of composite likelihood with $w_s = 1$.

2.2 Previous work for composite likelihood

In this section, we will review several papers in the literature which have discussed construction methods and the corresponding properties for composite likelihood.

2.2.1 Lindsay (1988). Composite likelihood methods

Though the idea of pseudo likelihood was first proposed by Besag (1974), the more general concept of composite likelihood was first introduced in Lindsay (1988). Including the definition given above, Lindsay (1988) also discussed general properties of composite likelihood.

Lindsay pointed out that the composite likelihood provides a generally consistent method of estimation. Because the structure of composite likelihood, each component likelihood is a true likelihood. Hence, the Kullback Leibler information inequality holds for each component log likelihood, $l_s = \log L_s$

$$E_{\theta_0}(l_s(\theta)) \leq E_{\theta_0}(l_s(\theta_0))$$

where θ_0 is the true value for θ. Therefore, we have, for the *weighted composite log likelihood* (cl),

$$\sup_{\theta} E_{\theta_0}(cl(\theta)) = E_{\theta_0}(cl(\theta_0))$$

which means the Kullback Leibler information inequality also holds for weighted composite likelihood. Hence, maximizing the composite likelihood leads to a consistent method of estimation under a set of regularity conditions and under a suitable asymptotic framework.

In addition, the MCLE comes from an unbiased estimating function since each component likelihood is a true likelihood; for each component likelihood, the composite scores satisfy $E_\theta(U_s(\theta)) = 0$, which leads to $E_\theta(CS(\theta)) = \sum_s E_\theta(U_s(\theta)) = 0$.

Understanding how to use information calculations to compare efficiency is important for composite likelihood. If the estimating function $g(\theta; \mathbf{y}) = g(\theta) = (g_1(\theta), \cdots, g_p(\theta))'$ has finite second moment $E(gg')$ and satisfies $E(g)$=0, then, $g(\theta)$ is called an *unbiased statistical estimating function*. Its associated information, called *Godambe information* or *the sandwich information*, is defined as

$$\begin{aligned} I_g(\theta) &= E(\nabla g(\theta))(Var(g(\theta)))^{-1}E(\nabla g(\theta))' \\ &= E(U_{mle}g')V_g^{-1}E(gU'_{mle}) \end{aligned}$$

where U_{mle} is the true score function. The second identity holds because of $E(\nabla g) = -E(U_{mle} g')$.

The composite scores each belong to the class of unbiased estimating equations, and for each component score, we have $Var(U_s) = E(-\nabla U_s)$. Thus, the information for composite likelihood can be written as

$$I_{CL} = (\sum_s Var(U_s))(Var(\sum_s U_s))^{-1}(\sum_s Var(U_s))'$$

However, in above equation, there could be correlation between U_s's, so that $Var(\sum_s U_s) \neq \sum_s Var(U_s)$. Therefore the total information I_{CL} is not equal to the sum of the component information. For example, if we replace a component score with one having more information, it does not necessarily cause an increase in total information.

Furthermore, if the parameter θ is a scalar, then the information for an unbiased estimating function is

$$I_g(\theta) = corr^2(U_{mle}, g) \times I(\theta)$$

where $I(\theta) = Var(U_{mle})$ is the *Fisher information*. Hence, the attainment of full information is associated with a linear relationship between U_{mle} and g. Among other things this means the efficiency for composite scores can be enhanced by using a weighting scheme for the composite score.

Suppose we want to maximize the information over weights $\{w : w = \sum_s w_s(\theta) U_s(\theta)\}$. The maximizing information problem is equivalent to the least squares problem

$$\min_w E_\theta (U_{mle} - w^T S)^2$$

where $w = (w_1(\theta), ..., w_m(\theta))^T$ and $S = (U_1(\theta), ..., U_m(\theta))^T$.

By solving this LS problem, we obtain the optimal weights for the weighted composite score is $w_{opt} = (VarS)^{-1}E(U_{mle}S)$, and the corresponding optimal information is $I_{opt} = V'(VarS)^{-1}V$, where V is the vector with sth coordinate $E(U_s^2) = E(U_{mle}U_s)$. Particularly, if the optimal weights are equal weights, i.e. $w_{opt} = k \cdot \mathbf{1}$, the following equality must hold

$$k^{-1} \sum_t Cov(U_s, U_t)/Var(U_s) = 1,$$

for all s. Notice that the weighted score $\sum w_s(\theta)U_s(\theta)$ is not the derivative of $\sum w_s(\theta) \log L_s(\theta)$, so there does not exist a corresponding weighted likelihood. Moreover, as we shall see, the optimal weights need not be nonnegative.

When $dim(\theta) = p$, we can in theory construct the best weighted score using $p \times p$ weight matrices $w_s(\theta)$ in a generalization of the above, and the associated I_{opt} would dominate in the sense of positive definiteness every other information matrix. However, because of the high dimensional matrix inversion requires, this is not always computationally feasible. (This is true even for scalar parameter models.)

However, Lindsay pointed out that under some particular structural features of the model, it is possible to show that the equal weighting solution is strongly optimal within a linear class. For example, the author proved that equal weighting of component scores that are equal in marginal information necessarily yields an improvement in information over using one component. In addition, stronger results can also be obtained by adding more structure. Suppose R_θ is a any random variable with property that U_{mle} is a function of R_θ, then for any unbiased estimating function g, define

$$g^*(\theta) = E_\theta[g(\theta)|R_\theta],$$

then Lindsay showed that $I_{g^*} \geq_{p.d} I_g$.

2.2.2 Cox and Reid (2004). A note on pseudo likelihood constructed from marginal densities

In this paper, Cox and Reid studied how to construct a composite likelihood from an arbitrary combination of univariate and bivariate densities, and discussed the asymptotic properties of the maximum composite likelihood estimator (MCLE). These properties depend on the rate at which the Godambe information converges to infinity.

Let $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_d)^T \sim f(\mathbf{y}; \theta)$, where θ is an unknown parameter, scalar. Now, suppose the univariate and bivariate densities are

$$f_s(y_s; \theta), f_{st}(y_s, y_t; \theta), \text{ where } s \neq t, s, t = 1, ..., d,$$

Then the two composite likelihoods, as the authors defined them, can be constructed as

$$\begin{aligned} l_1(\theta; \mathbf{Y}) &= \sum_s \log f_s(y_s; \theta) \\ l_2(\theta; \mathbf{Y}) &= \sum_{s<t} \log f_{st}(y_s, y_t; \theta) - a \times d \times l_1(\theta; \mathbf{Y}) \end{aligned}$$

In the above construction, l_1 is the onewise marginal log likelihood, and if $a = 0$, then l_2 is the pairwise marginal log likelihood. In addition, if $a = 1/2$, then l_2 is all pairwise conditional log likelihood, which has similar structure to Besag's pseudo likelihood (1974). They stated that the best constant a is typically nonnegative, except in the rare case that most of the information is in l_1 and relative little in l_2.

If the observations $\mathbf{y}_1, ..., \mathbf{y}_n$ are iid, they define corresponding composite likelihoods

$$\begin{aligned} l_1(\theta; \mathbf{y}_1, ..., \mathbf{y}_n) &= \sum_i l_1(\theta; \mathbf{y}_i) \\ l_2(\theta; \mathbf{y}_1, ..., \mathbf{y}_n) &= \sum_i l_2(\theta; \mathbf{y}_i) \end{aligned}$$

Similar to Lindsay (1988), the associated composite score functions are defined to be

$$\begin{aligned} U_1(\theta;\mathbf{Y}) &= \partial l_1(\theta;\mathbf{Y})/\partial\theta = \sum_s U_{1s}(\theta) \\ U_2(\theta;\mathbf{Y}) &= \partial l_2(\theta;\mathbf{Y})/\partial\theta = \sum_{s<t} U_{2st}(\theta) - aq\sum_s U_{1s}(\theta) \\ U_1(\theta;\mathbf{y}_1,...,\mathbf{y}_n) &= \sum_i U_1(\theta;\mathbf{y}_i) \\ U_2(\theta;\mathbf{y}_1,...,\mathbf{y}_n) &= \sum_i U_2(\theta;\mathbf{y}_i). \end{aligned}$$

Hence the estimating equations they consider are

$$U_\nu(\widetilde{\theta};\mathbf{y}_1,...,\mathbf{y}_n) = 0, \text{ for } \nu = 1,2.$$

To discuss the asymptotic properties, the authors distinguish between two asymptotic scenarios depending on the data structure. The first one is of many independent replicates of the data, i.e. fixed d while increasing n. This kind of data structure includes longitudinal data. The second one is of few individuals but with large sequences, i.e. fixed n while increasing d. Time series and spatial data just belong to this class.

For the first setting, the consistency and asymptotic normality of the composite likelihood estimators are satisfied under some regularity conditions. That is, we have the asymptotic distribution of $\widetilde{\theta}$ is

$$\widetilde{\theta} \to N(\theta, \text{Var})$$

where $\text{Var} = [E(-U'_\nu(\theta))]^{-2}E(U^2_\nu(\theta))$, whose terms can be estimated by

$$\begin{aligned} \widehat{E}(U^2_\nu(\theta)) &= \frac{1}{n}\sum_i U^2_\nu(\widetilde{\theta};\mathbf{y}_i) \\ \widehat{E}(-U'_\nu(\theta)) &= -l''_\nu(\widetilde{\theta})/n \end{aligned}$$

However, for the second data setting, although the estimating equation is still unbiased, the unbiased structure no longer ensures satisfactory properties for

the resulting estimator. The author proved that by doing the first order expansion for $U_1(\widetilde{\theta}; Y) = 0$ around θ, we have

$$d^{-1}\sum U_{1s}(\theta) + d^{-1}(\widetilde{\theta} - \theta)\sum U'_{1s}(\theta) = 0.$$

Typically, the second term is $O_p(1)$, while the mean value for first term is 0, and the variance for first term of sum is

$$d^{-2}\{\sum_s Var(U_{1s}) + 2\sum_{s<t} Cov(U_{1s}, U_{1t})\} \approx O_p(d^{k-2})$$

where $1 \leq k \leq 2$. Thus, the first random sum is of order $O_p(d^{k/2-1})$. If $k = 2$, then the order is $O_p(1)$, which is the same as the second term. Hence, $\widetilde{\theta}$ will not be consistent as d increases. If $k < 2$, $d^{1-k/2}(\widetilde{\theta} - \theta) \to N(0, Var)$. However, if k is close to 2, then the convergence will be very slow.

More specifically, suppose the marginal density of $\mathbf{Y}$ belongs to any exponential family distribution with mean θ, and with some correlation structure. If the correlation structure is a short-range dependent stationary time series, then $k = 1$ and the convergence of the overall sample mean to θ is at usually rate $1/\sqrt{d}$. If the correlation is a long-range dependent process with Hurst coefficient $H > 1/2$, then $k = 2H$ and the convergence is slow. If all pairs are equally correlated, then $k = 2$ and $\widetilde{\theta}$ is not consistent.

At the end of this paper, the authors also mentioned that if the parameter of interest is the correlation between two elements of the vectors, then no information is available about it in the univariate margins. If the parameter of interest appears in both bivariate and univariate margins, then it is possible by choosing a to maximize the information provided.

In our example, it is possible to envision both n = number of individuals and d = length of sequence being very large, and do relevant in any asymptotic approximation.

2.2.3 A hybrid pairwise likelihood method

In Kuk (2007), the author introduced a modification to the pairwise likelihood method, which aims to improve the estimation of marginal (onewise) distribution parameters. The idea was to estimate marginal distribution parameters by the optimal linear combination of the onewise marginal scores while using the pairwise scores to estimate the nuisance parameters needed for the optimal weight. He called this the *hybrid pairwise likelihood* although technically it is a composite score method, as there is not a corresponding likelihood function that is being maximized. The constructed hybrid pairwise likelihood is robust to misspecification of the bivariate distributions as long as the univariate marginal distributions are correctly specified.

The author proposed this method for clustered data of the following type. Let $\mathbf{Y}_i = (Y_{i1}, ..., Y_{id_i})^T$, where $i = 1, ..., n$, be independent vectors of observations. The bivariate density of Y_{ij} and Y_{ik}, for $j \neq k$, is $f(y_{ij}, y_{ik}; \theta)$. The unknown parameters θ can be partitioned as $\theta = (\psi, \lambda)$, where ψ consists of the components of θ that are involved in the univariate marginal densities $f(y_{ij}; \psi)$, and λ are the remaining parameters.

The *pairwise log likelihood* is defined to be

$$l_p(\theta) = l_p(\psi, \lambda) = \sum_{i=1}^{n} \sum_{j>k} l_{ijk}$$

where $l_{ijk} = \log f(y_{ij}, y_{ik}; \psi, \lambda)$ is the bivariate log density for the observation pair $(y_{ij}, y_{ik})^T$ in cluster i. By solving the estimating equations

$$0 = \sum_{i=1}^{n} \sum_{j>k} l'_{ijk} = \begin{pmatrix} \sum_{i=1}^{n} \sum_{j>k} \partial l_{ijk}/\partial\psi \\ \\ \sum_{i=1}^{n} \sum_{j>k} \partial l_{ijk}/\partial\lambda \end{pmatrix} \tag{2.3}$$

One could maximize the pairwise likelihood.

Instead of using pairwise marginal, let

$$G_{ij}(\psi) = \frac{\partial l_{ij}}{\partial \psi}$$

where $l_{ij} = \log f(y_{ij}, \psi)$ be the onewise score. Assuming the data are independent within clusters, we could estimate ψ by solving

$$\sum_{i=1}^{n}\sum_{j=1}^{d_i}\frac{\partial l_{ij}}{\partial \psi} = \sum_{i=1}^{n}\sum_{j=1}^{d_i} G_{ij} = \sum_{i=1}^{n} C_i G_i = 0$$

where $G_i = vec(G_{i1}, ..., G_{id_i})$, and $C_i = (I_q, ..., I_q)$ consists of d_i copies of the $q \times q$ identity matrix with $q = dim(\psi)$. By choosing a more general C_i, we can find a better linear combinations of the marginal scores. According to the theory of optimal estimating functions (McCullagh and Nelder, 1989, §9.4), the optimal weights are $C_i = D_i^T V_i^{-1}$, where $V_i = V_i(\psi, \lambda) = Cov(G_i)$, and $D_i = -\partial G_i/\partial \psi$. Hence, the optimal estimating equation for ψ based on linear combinations of the marginal scores is given by

$$0 = \sum_{i=1}^{n} D_i^T V_i^{-1} G_i \tag{2.4}$$

However, the λ in the weight matrix C_i are still unknown, so the author proposed to iterate between solving (2.4) for ψ with λ fixed and then solving

$$0 = \sum_{i=1}^{n}\sum_{j>k}\frac{\partial l_{ijk}}{\partial \lambda}$$

for λ with ψ fixed. Then after convergence, we have obtained the solutions $(\widehat{\psi}, \widehat{\lambda})$ from the simultaneous equations

$$\begin{pmatrix} \sum_{i=1}^{n} D_i^T V_i^{-1} G_i \\ \\ \sum_{i=1}^{n}\sum_{j>k}\frac{\partial l_{ijk}}{\partial \lambda} \end{pmatrix} = 0. \tag{2.5}$$

Comparing (2.5) with (2.3), we can see that (2.5) is a hybrid estimating equation that combines the derivative of the pairwise log likelihood with respect

to λ with the optimal linear combinations of the marginal score function of ψ. The author induced the following approximation by expanding (2.5) at the true parameter values and ignoring terms with zero expectations

$$\begin{pmatrix} \widehat{\psi} - \psi \\ \\ \widehat{\lambda} - \lambda \end{pmatrix} = - \begin{pmatrix} \sum_{i=1}^{n} D_i^T V_i^{-1} G_i & 0 \\ \\ \sum_{i=1}^{n} \sum_{j>k} \frac{\partial^2 l_{ijk}}{\partial \psi \partial \lambda} & \sum_{i=1}^{n} \sum_{j>k} \frac{\partial^2 l_{ijk}}{\partial \lambda^2} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} D_i^T V_i^{-1} G_i \\ \\ \sum_{i=1}^{n} \sum_{j>k} \frac{\partial l_{ijk}}{\partial \lambda} \end{pmatrix}$$

Then, the asymptotic covariance matrix of $\widehat{\psi}$ and λ, as the number of clusters $n \to \infty$, is given by $B^{-1}C(B^{-1})^T$, where B is the triangular matrix appearing in the above equation and C is the covariance matrix of

$$\begin{pmatrix} \sum_{i=1}^{n} D_i^T V_i^{-1} G_i \\ \\ \sum_{i=1}^{n} \sum_{j>k} \frac{\partial l_{ijk}}{\partial \lambda} \end{pmatrix}$$

Finally, the author also pointed out some possible extensions. We could go further to consider the optimal linear combinations of the pairwise scores $\partial l_{ijk}/\partial\lambda$ for parameters λ. The idea is to stack all the pairwise scores $\partial l_{ijk}/\partial\lambda$, $1 \leq j \leq k \leq d_i$, into a vector H_i for cluster i, and form the optimal estimating equation

$$0 = \sum_{i=1}^{n} (\partial H/\partial\lambda)^T (Cov(H_i))^{-1} H_i$$

for λ. However, the computation of the covariance matrix $Cov(H_i)$ will in general require knowledge about the fourth order joint distributions, which is not always available. Even if it is available, $Cov(H_i)$ is of order $O(d_i^2) \times O(d_i^2)$, and its inversion is computationally infeasible for large clusters.

2.2.4 Varin (2008). On composite marginal likelihood

In this paper, Varin (2008) gave a thorough review of the marginal composite likelihood and its applications. However he did exclude biological applications.

In addition, the author also discussed some topics on composite likelihood from a global point of view.

Firstly, Varin (2008) discussed the classification of composite likelihood by type, which was proposed in Varin and Vidoni (2005). The authors think that the composite likelihood can be mainly classified into two classes based on the construction method. The first class is omission methods. The idea is that the composite likelihood is obtained by removing some terms that make the evaluation of the full likelihood complicated. If the removed part is not very informative on the parameter of interest, then the loss of efficiency may be tolerated. For example, the pseudo likelihood introduced by Besag (1974, 1977), the m-order likelihood for stationary processes (Azzalini (1983)), and the approximate likelihood of Stein (2004). The second class is constructed from marginal likelihoods. For example, the methods reviewed in Cox and Reid (2004) and Kuk (2007).

Secondly, Varin (2008) also discussed justification for the composite likelihood. Basically, there are two methods to justify the composite likelihood.

The first method is from unbiased estimating equation point of view. The conclusion is that if the true parameter θ_0 belongs to the interior of a compact parametric space, then the *maximum composite likelihood estimator* (MCLE) is the solution, if unique, of the composite score equation

$$CS(\theta, \mathbf{y}) = \nabla cl(\theta; \mathbf{y}) = \sum_{s=1}^{m} w_s U_s(\theta; \mathbf{y}) = 0$$

It is obvious that the above composite score is unbiased score function, since each component U_s is unbiased. The asymptotic properties of MCLE depends on the rate at which the Godambe information $I_{CL}(\theta)$ converges to infinity as Fisher information $I(\theta)$ increases. If it converges fast enough, then MCLE is consistent and asymptotic normally distributed.

The second method is from Kullback-Leibler divergence point of view. Varin

and Vidoni (2005) define the composite Kullback-Leibler divergence as the linear combination of the Kullback-Leibler divergences associated with each component of the composite likelihood. In other terms, given a random variable $\mathbf{Y} = (Y_1, ..., Y_d)^T$, with density $g(\mathbf{y})$, the composite Kullback-Leibler divergence of a density $h(\mathbf{y})$ relative to $g(\mathbf{y})$ is

$$\triangle c(g,h) \;\; = \;\; E_g\left[\log \frac{CL(g,\mathbf{y})}{CL(h,\mathbf{y})}\right]$$

Then we have $\triangle c(g,h) \geq 0$ because each component is an ordinary Kullback-Leibler divergence, which is nonnegative.

Under natural regularity conditions and in the right dependency setting, the maximum composite likelihood estimator converges to the minimizer of the composite Kullback-Leibler divergence between the assumed model f and the true model g, which is unknown. In general, this minimizer is a pseudo true parameter value because the parametric model for the marginal density could be misspecified.

The author also proposed how to estimate the Godambe information matrix, $I_{CL}(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$, where $H(\theta) = E(-\nabla CS)$ and $J(\theta) = Var(CS)$. The estimated $H(\theta)$ is

$$\widehat{H}(\mathbf{y}) = -\frac{1}{W_m}\sum_{s=1}^{m} w_s \nabla U_s(\widehat{\theta}_{MCLE};\mathbf{y})$$

where $W_m = \sum_s w_s$. Since the second Bartlett identity is valid for each individual likelihood term, we can also use the estimator

$$\widehat{H}(\mathbf{y}) = \frac{1}{W_m}\sum_{s=1}^{m} w_i U_s(\widehat{\theta}_{MCLE};\mathbf{y})U_s(\widehat{\theta}_{MCLE};\mathbf{y})^T$$

The estimation for $J(\theta)$ is quite difficult. If we observe independent or pseudo independent replicated data, for example, the longitudinal or clustered data, then the empirical estimator can be used. For example, in longitudinal data,

suppose $\mathbf{y}_i = (y_{i1}, ..., y_{id_i})^T$ is the observation for the ith cluster, where $i = 1, ..., n$, then

$$\widehat{J}(Y) = \frac{1}{n} \sum_{i=1}^{n} \nabla cl(\widehat{\theta}_{MCLE}; \mathbf{y}_i) \nabla cl(\widehat{\theta}_{MCLE}; \mathbf{y}_i)^T$$

For non-clustered data, or data in pseudo independent subgroups, $J(\theta)$ can be estimated by using empirical and jackknife estimator. Finally, it is also possible to use the Monte Carlo integration method to estimate the Godambe information. However, we then need to use the assumptions about the full joint distribution of the data.

Chapter 3

Construction of Composite Scores

As discussed in Lindsay (1988), typically there is information loss for the maximum composite likelihood estimator (MCLE) when compared with the MLE, except for some very specific parameter values. The information loss results in an efficiency loss for the composite likelihood method. Hence, one interesting problem is how to construct a proper composite likelihood so that it maintains efficiency as much as possible and keeps the computations economical at the same time. In this chapter, we will discuss issues about the construction of composite likelihood and weighted composite scores from the point of view of optimal efficiency and discuss the properties of the resulting estimators.

3.1 Hoeffding scores and likelihood

In this section, we will discuss the construction of composite likelihood based on Hoeffding projection. The idea is to project the true score function of a random variable Y on some linear classes of estimating functions, so that we can find the best elements in these classes in the sense of Godambe information or asymptotic variance, to substitute the true score function.

3.1.1 Notation and framework

Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_d)^T$ be a d-dimensional random vector with density $f(\mathbf{y}; \theta)$, where θ is a p dimensional unknown parameter. Note that $Y_1, ..., Y_d$ are not

necessarily independent. Let $U_{mle}(\theta; \mathbf{y}) = \nabla_\theta \log f(\mathbf{y}; \theta)$ be the true score function for $\mathbf{Y}$. As we have seen, there could be situations where the full distribution $f(\mathbf{y}; \theta)$ is difficult to specify, but the specification of onewise marginal, pairwise or pairwise conditional model are possible. In our discussion, we will not consider this case, focussing rather on the case where a true score is well defined.

Let L_i and U_i denote onewise marginal likelihoods and scores, L_{ij} and U_{ij} be the pairwise marginal likelihoods and scores, and $L_{i|j}$ and $U_{i|j}$ be the pairwise conditional likelihoods and scores, respectively. According to the definition for composite likelihood in chapter 2, all above likelihoods are composite likelihoods. Simple calculations give the following lemma.

Lemma 3.1. *$E(U_{mle}|Y_i) = U_i$ for all i, where U_i is the marginal score for Y_i.*

Proof. For notation simplicity, we assume $d = 2$, so $f(y_1, y_2; \theta) = f_1(y_1; \theta) f(y_2|y_1; \theta)$, where $f_1(y_1; \theta)$ is the marginal density of Y_1 and $f(y_2|y_1; \theta)$ is the conditional density of Y_2 given Y_1. Hence

$$\begin{aligned} E[U_{mle}|Y_1] &= E[\frac{\partial}{\partial\theta} \log f_1(y_1) + \frac{\partial}{\partial\theta} f(y_2|y_1)|Y_1] \\ &= \frac{\partial}{\partial\theta} \log f_1(y_1) + E[\frac{\partial}{\partial\theta} f(y_2|y_1)|Y_1]] \\ &= U_1 + 0 \end{aligned}$$

where U_1 is the marginal score for Y_1, and the second term equals zero because score function has mean zero in its probability distribution. □

We next define independent parameter value by

Definition 3.1. *Let an independence parameter value θ_{ind} be any value of θ for which all Y_i's are independent.*

For example, suppose $\mathbf{Y}$ has a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{ij})_{d\times d}$, where $\sigma_{ii} = \sigma^2$ and $\sigma_{ij} = \rho\sigma^2 (i \neq j)$, and $\theta = (\rho, \sigma^2)^T$. Then $\theta_{ind} = (0, \sigma^2)^T$ is an indepence parameter value.

3.1.2 Inference function and estimating function

We next provide a short review of inference functions and estimating functions. Suppose we observe n independent random samples,$\mathbf{y}_1, ..., \mathbf{y}_n$, from $f(\mathbf{y}; \theta)$, a distribution with an unknown parameter θ. To estimate θ, we can find an inference function, say, $G(\theta; \mathbf{y})$, and then optimize (typically maximize) its information over samples with respect to θ. For example, if we take $G(\theta; \mathbf{y}) = f(\mathbf{y}; \theta)$ as the inference function and maximize the likelihood function $G(\theta; \mathbf{y})$, then the maximizer is the maximum likelihood estimator. Instead of using an inference function, another approach to estimate θ is based on a $p \times 1$ vector of estimating functions $g(\theta; \mathbf{y})$ by solving $g(\theta; \mathbf{y}) = \mathbf{0}$ for θ. For instance, we can solve equations $g(\theta; \mathbf{y}) = \sum_i \nabla \log f(\mathbf{y}; \theta) = \mathbf{0}$ to obtain the MLE in iid data.

In regular cases, optimizing the inference function $G(\theta; \mathbf{y})$ is equivalent to solving its score function, i.e., finding the solution to $\nabla_\theta G(\theta; \mathbf{y}) = \mathbf{0}$. (However, these two methods could be different in boundary cases.) If $G(\theta; \mathbf{y})$ is available, we could use $g(\theta; \mathbf{y}) = \nabla G(\theta; \mathbf{y})$ as the estimating function. In other cases, we may merely interested in estimating function $g(\theta; \mathbf{y})$ without referring back to the parent inference function $G(\theta; \mathbf{y})$. This is true when $G(\theta; \mathbf{y})$ is not easy to construct or there may not exist a meaningful parent function.

Inference functions, such as the likelihood, have the strength of having a clear conceptual framework. As we will see later, when we use composite likelihood, they can also lead to reliable algorithms for optimization, for example, the EM algorithm. In addition, in our later analysis, composite likelihood provides resolution for irregularity problems, such as multiple solutions and boundary problems. On the other hand, estimating functions provide more flexibility in construction of efficient methods, and a rich optimality theory has been developed for combining available information. Moreover, they can be more robust than inference functions because the estimating functions can be created using fewer distributional

assumptions.

3.1.3 Hoeffding scores

Keep the notations as above and suppose $\mathbf{Y} = (Y_1, ..., Y_d)^T \sim f(\mathbf{y}; \theta)$. As before $Y_1, ..., Y_d$ are not necessarily independent. Let $U_{mle}(\theta; \mathbf{y}) = \nabla_\theta \log(f(\theta; \mathbf{y})$ be the true score function for $\mathbf{Y}$. Then, we have the following efficiency theorem

Theorem 3.2. *Suppose that for each θ, $V^* = V_\theta^* \in \mathcal{L}$, where $\mathcal{L}$ is a linear space of estimating functions, satisfies*

$$inf_{V \in \mathcal{L}} E(U_{mle} - V)(U_{mle} - V)^T = E(U_{mle} - V^*)(U_{mle} - V^*)^T$$

for all $V \in \mathcal{L}$. Then, $V^* = V_\theta^*$ is the most efficient element of $\mathcal{L}$ in terms of Godambe information.

Proof. Without loss of generality, we assume that the elements in $\mathcal{L}$ are unbiased estimating functions. Otherwise, they can be mean-centered.

By the projection theorem in Hilbert space (Small and McLeish (1994)), the minimizer V^* of $E(U_{mle} - V)(U_{mle} - V)^T$ satisfies

$$E[(U_{mle} - V^*)V^T] = 0, \text{ for any } V \in \mathcal{L}.$$

Hence, we have

$$E(V^* V^T) = E(U_{mle} V^T) = E[-\nabla V^T]. \tag{3.1}$$

Since V^* also belongs to $\mathcal{L}$, we also have

$$E(V^* V^{*T}) = E(U_{mle} V^{*T}) = E[-\nabla V^{*T}]. \tag{3.2}$$

Let $\delta = V^* - E(V^*V^T)[E(VV^T)]^{-1}V$. Then

$$\begin{aligned} 0 &\leq E(\delta\delta^T) \\ &= E\{V^*V^{*T} - V^*V^T[E(VV^T)]^{-1}E(VV^{*T}) \\ &\quad -E(V^*V^T)[E(VV^T)]^{-1}VV^{*T} \\ &\quad +E(V^*V^T)[E(VV^T)]^{-1}VV^T[E(VV^T)]^{-1}E(VV^{*T})\} \\ &= E(V^*V^{*T}) - E(V^*V^T)[E(VV^T)]^{-1}E(VV^{*T}) \end{aligned}$$

Thus,

$$E(V^*V^{*T}) \geq E(V^*V^T)[E(VV^T)]^{-1}E(VV^{*T}). \tag{3.3}$$

Note, by (3.1)

$$\begin{aligned} & E(V^*V^T)[E(VV^T)]^{-1}E(VV^{*T}) \\ = \; & E[-\nabla V^T][E(VV^T)]^{-1}E[-\nabla V] \\ = \; & I_V \end{aligned}$$

Thus, the right hand side of (3.3) is just the Godambe information for V.

Similarly, by (3.2), the left hand side of (3.3) equals

$$\begin{aligned} E(V^*V^{*T}) &= E(V^*V^{*T})[E(V^*V^{*T})]^{-1}E(V^*V^{*T}) \\ &= E[-\nabla V^{*T}][E(V^*V^{*T})]^{-1}E[-\nabla V^*] \\ &= I_{V^*}, \end{aligned}$$

which is the Godambe information for V^*.

Therefore, we have proved $I_{V^*} \geq I_V$, and as a result $V^* = V_\theta^*$ is the most efficient element of $\mathcal{L}$ in terms of Godambe information. □

Now, given a fixed θ, define the linear classes of estimating functions as

$$\begin{aligned}
\mathcal{H}_1 &= \{\mu + \sum_{i=1}^{n} g_i(Y_i)\}, \\
\mathcal{H}_2 &= \{\mu + \sum_{i=1}^{n} g_i(Y_i) + \sum_{i<j} g_{ij}(Y_i, Y_j)\}, \\
\mathcal{H}_3 &= \{\mu + \sum_{i=1}^{n} g_i(Y_i) + \sum_{i<j} g_{ij}(Y_i, Y_j) + \sum_{i<j<k} g_{ijk}(Y_i, Y_j, Y_k)\}, \\
&\quad etc.
\end{aligned}$$

where μ is in $\mathbb{R}$ and g_i, g_{ij}, and g_{ijk}, are arbitrary finite variance functions with mean zero under θ. We call these *additive estimating functions*. Our purpose is to find, for each θ, an optimal element from $\mathcal{H}_k$, $k = 1, 2, 3, ...$ in terms of estimating efficiency for each θ. Notice that candidate elements of $\mathcal{H}_k$ includes all marginal or conditional scores of low order in the number of variables, so the optimal $\mathcal{H}_k$ scores must have more information than the best weighted composite scores of the same order.

If we want to use one element from $\mathcal{H}_k$, $k = 1, 2, 3, ...$, to approximate the true score function U_{mle}, then the best choice is to use the projection of U_{mle} on $\mathcal{H}_k$ under θ in the L_2 sense based on Theorem 3.2. That is, suppose $h_k(Y_1, ..., Y_d) \in \mathcal{H}_k$ and for any function $H(Y_1, ..., Y_d) \in \mathcal{H}_k$, we have h_k minimizes $E_\theta[(U_{mle} - H)(U_{mle} - H)^T]$, then h_k is the best approximation for U_{mle} in the class $\mathcal{H}_k$. A necessary and sufficient condition for projection to be h_k is that for any other function $H \in \mathcal{H}_k$, we have $E_\theta[(U_{mle} - h_k)H] = 0$.

Because of the dependency among Y_i's, it is not necessarily easy to find the projection h_k. However, under θ_{ind}, we have the following Hoeffding projection theorem, which depends on the theory of U-statistics.

Theorem 3.3. *Under θ_{ind}, the projections of true score function U_{mle} on $\mathcal{H}_k$,*

$k = 1, 2, 3, ...$ *are*

$$\begin{aligned} h_1(\theta, Y) &= \sum_{i=1} U_i \\ h_2(\theta, Y) &= h_1 + \sum_{i<j} (U_{ij} - U_i - U_j) \\ h_3(\theta, Y) &= h_2 + \sum_{i<j<k} (U_{ijk} - U_{ij} - U_{ik} - U_{jk} + U_i + U_j + U_k) \\ & \quad etc. \end{aligned}$$

where U_i, U_{ij}, and U_{ijk} are corresponding marginal, pairwise, and triplewise score functions.

Proof. For the simplicity, we just prove that h_1 is the projection of U_{mle} on $\mathcal{H}_1$. The proof for h_k, $k = 2, 3, ...$, is similar.

To prove h_1 is the projection, it is equivalent to prove that for any arbitrary function $g_j(y_j)$, $j = 1, ..., d$, the equality $E[(U_{mle} - h_1) \sum_j g_j(y_j)] = 0$ holds under θ_{ind}.

$$\begin{aligned} E[(U_{mle} - h_1) g_j(y_j)] &= E\left[\left(U_{mle} - \sum_i E(U_{mle}|Y_i) \right) g_j(y_j) \right] \\ &= E\left\{ E\left[\left(U_{mle} - \sum_i E(U_{mle}|Y_i) \right) g_j(y_j) \mid Y_j \right] \right\} \\ &= E\left\{ g_j(y_j) E\left[\left(U_{mle} - \sum_i E(U_{mle}|Y_i) \right) |Y_j \right] \right\} \\ &= E\left\{ g_j(y_j) \left[E(U_{mle}|Y_j) - \sum_i E[E(U_{mle}|Y_i)|Y_j] \right] \right\} \end{aligned}$$

Note that $E[E(U_{mle}|Y_j)|Y_j] = E(U_{mle}|Y_j)$ and $E[E(U_{mle}|Y_i)|Y_j] = E[E(U_{mle}|Y_i)] = E(U_i) = 0$ because of the independence between Y_i and Y_j, thus the above equality equals zero, which leads to the proof that $h_1 = \sum_i U_i$ is the desired projection. □

The above Theorem 3.3 gives us the projected scores on the classes of marginal or pairwise estimating functions under θ_{ind}. Thus, we can call h_k, $k = 1, 2, ...$, the *k-th order Hoeffding score*.

Let $U^*_{ij} = U_{ij} - U_i - U_j$. Then, U^*_{ij} can be interpreted as *corrected pairwise score*, that is, it removes overused marginal information from the pairwise score U_{ij}. Hence, h_2 is the sum of all marginal and corrected pairwise scores. By further algebra, we can write h_2 as

$$h_2(\theta, Y) = \sum_{i<j} U_{ij} - (d-2) \sum_i U_i.$$

Moreover, if we go back from Hoeffding scores to inference functions, the above projections give us new kinds of likelihood ratios, which we call the *Hoeffding likelihoods*

Definition 3.2 (Hoeffding likelihood)**.**

$$\begin{aligned} L_{h_1} &= \prod_i f_i(Y_i) \\ L_{h_2} &= L_{h_1} \prod_{i<j} \frac{f_{ij}(Y_i, Y_j)}{f_i(Y_i) f_j(Y_j)} \\ L_{h_3} &= L_{h_2} \prod_{i<j<k} \frac{f_{ijk}(Y_i, Y_j, Y_j) f_i(Y_i) f_j(Y_j) f_k(Y_k)}{f_{ij}(Y_i, Y_j) f_{ik}(Y_i, Y_k) f_{kj}(Y_k, Y_j)} \\ & \quad \textit{etc.} \end{aligned}$$

where $f_i(y_i)$, $f_{ij}(y_i, y_j)$, and $f_{ijk}(y_i, y_j, y_k)$ are corresponding marginal, pairwise, and triplewise densities.

Note that the above Hoeffding likelihoods are invariant under permutation of the indices because for example $f_{ij}(y_i, y_j) = f_{ji}(y_j, y_i)$.

Again, similar as for Hoeffding scores, basic algebra shows that L_{h_2} can be written as

$$L_{h_2} = \frac{\prod_{i<j} f_{ij}(y_i, y_j)}{[\prod_i f_i(y_i)]^{d-2}}$$

It is easily seen that the Hoeffding likelihoods in Definition 3.2 are ratios between true likelihoods, but they are not always 'true' composite likelihoods because of the possibility of negative weights according to (2.2). For example, L_{h_1}

is a true composite likelihood, but L_{h_2} is not a true composite likelihood unless $d = 2$, where L_{h_2} is equivalent to pairwise likelihood, or $d = 3$, where L_{h_2} is equivalent to all pairwise conditional likelihoods. That is, unlike the true composite scores having all non-negative weights, the Hoeffding scores $h_k(\theta; \mathbf{y}) = \nabla L_{h_k}$ are the linear combinations $\sum_s w_s U_s$ of composite scores with some negative weights imposed on U_s. Hence, one big question arising from Hoeffding scores is, under general θ, whether it makes sense to use $\log L_{h_k}$ as an inference function. In other words, do the negative weights in $h_k(\theta; \mathbf{y})$ mean negative information?

3.2 Discussion for Hoeffding likelihoods

Theorem 3.3 tells us the pointwise optimality at θ_{ind} of L_{h_k}. This gives us the idea that we might still use these Hoeffding scores, or more specifically, the corresponding Hoeffding likelihood to substitute for the true likelihood and estimate the unknown parameters. This replacement will be much more computationally economical if there exists complex relationships between random variable components. However, as discussed at the end of last section, it is not clear how well L_{h_k} performs at $\theta \neq \theta_{ind}$.

As discussed in Lindsay (1988), an optimal estimating function should be linearly correlated with the true score function U_{mle}. However, it is not necessary that Hoeffding scores have linear relationship with U_{mle} except at θ_{ind}. Hence, it may not be optimal elsewhere. However, it is still possible that $\log(L_{h_k})$ can be used for reasonable inference. To discuss this problem, let's first see why log true composite likelihood, log(TCL), denoted as $G(\theta; \mathbf{y})$ works as as an inference function.

Theorem 3.4 (Global maximization property)**.** *$E_{\theta_\tau}[G(\theta; \mathbf{Y})]$ is maximized at $\theta = \theta_\tau$, where θ_τ is the true value of θ.*

Proof : Since $G(\theta; \mathbf{y})$ denotes the log true composite likelihood, we have $G(\theta; \mathbf{y}) =$

$\sum_s \log L_s = \sum_s \log f(\mathbf{Y}_s;\theta)$, where s is an index representing a marginal or conditional event $\mathbf{Y}_s$.

$$E_{\theta_\tau}[G(\theta;\mathbf{Y})] = E_{\theta_\tau}\left[\sum_s \log f(\mathbf{Y}_s;\theta)\right] = \sum_s \int \log\{f(\mathbf{Y}_s;\theta)\} f(\mathbf{Y}_s;\theta_\tau) d\mathbf{Y}_s$$

$$\begin{aligned} E_{\theta_\tau}[G(\theta;\mathbf{Y})] - E_{\theta_\tau}[G(\theta_\tau;\mathbf{Y})] &= \sum_s \int \log\left\{\frac{f(\mathbf{Y}_s;\theta)}{f(\mathbf{Y}_s;\theta_\tau)}\right\} f(\mathbf{Y}_s;\theta_\tau) d\mathbf{Y}_s \\ &\leq \sum_s \log\left\{\int \frac{f(\mathbf{Y}_s;\theta)}{f(\mathbf{Y}_s;\theta_\tau)} f(\mathbf{Y}_s;\theta_\tau) d\mathbf{Y}_s\right\} \\ &= \sum_s \log\left\{\int f(\mathbf{Y}_s;\theta) d\mathbf{Y}_s\right\} \\ &= \sum_s \log 1 = 0, \end{aligned}$$

where the inequality comes from Jensen's inequality. □

Let us call an inference function G that satisfies this property a *max-consistent* inference function.

If $-\nabla^2 E_{\theta_\tau}[G(\theta_\tau;\mathbf{Y})]$ is positive definite, we will say it has the local maximization property or is locally max-consistent.

The global max-consistency property often implies the local one, but not vice versa.

Now the question changes to be whether Hoeffding likelihoods have the global, or at least local, consistency properties. That is, does it make sense to maximize L_{h_k} to obtain estimators. In fact, the examples provided later in this section provide strong evidence that Hoeffding likelihood fails when θ is not near θ_{ind}. As a result, we seek to repair the Hoeffding scores using ideas from estimating functions. To do this, we start with *information identity*.

Definition 3.3 (Information identity)**.** *An inference function G satisfies the information identity if*

$$E\{-\nabla^2 G\} = E\{\nabla G \cdot \nabla^T G\}.$$

An estimating function g satisfies the information identity if

$$E\{-\nabla g\} = E\{gg^T\}.$$

If G is a true likelihood, then the information identity is the same as the Bartlett identity; if g is taken as a score function from G, then g inherits the information identity property of G. The satisfaction of the information identity by an inference function will induce the local max-consistency property. For a true composite likelihood, because each composite score satisfies the information identity, linear combinations of the composite scores with nonnegative weights automatically have the local max-consistency property. Thus, true composite likelihood are sensible inference functions. However, because of the negative weights, this is not the case for the Hoeffding likelihoods.

3.3 Modified Hoeffding likelihood

As discussed in previous section, Hoeffding scores do not satisfy the information identity. Hence, we cannot conclude the local maximization property. We therefore proposed to modify the Hoeffding scores by using *semi-optimal* weights.

Now, let us consider a class of estimating functions, g, which have no parent inference functions, G. We can set up a weaker version of the maximization properties called *positive likelihood association*, to help ensure the estimator from estimating function, $\widehat{\theta}_{EF}$, would have reasonable consistency properties.

Definition 3.4 (Positive likelihood association)**.** *A estimating function g is said to have positive likelihood association if*

$$E(gU_{mle}^T) = E[-\nabla g]$$

is nonnegative definite for all θ.

Notice that U_{mle}, as the gradient of the likelihood, always points to 'uphill' on the likelihood. If g has positive likelihood association, then g and the true score function U_{mle} both generally point in the same direction. Hence, the estimating function g should make the estimator $\widehat{\theta}_{EF}$ similar to the likelihood maximum.

Moreover, we also have the conclusion that if an estimating function g satisfies the information identity, then g also has positive likelihood association (but not vice versa). Moreover, a sum of estimating functions with positive likelihood association also has positive likelihood association. Thus all the estimating functions tend to point uphill on the likelihood.

Now, back to the Hoeffding score $U_{ij}^* = U_{ij} - U_i - U_j$. It is not necessary that U_{ij}^* has positive likelihood association. For example, suppose $\mathbf{Y} = (Y_1, Y_2, ..., Y_d)^T \sim N(\theta\mathbf{1}, \mathbf{\Sigma})$, where $\mathbf{1}$ is a $d \times 1$ vector of elements 1, and Σ is a $d \times d$ matrix with diagonal elements 1 and off-diagonal elements ρ. Here θ is an unknown parameter, but ρ is assumed known. For $i \neq j$, we have the pairwise and onewise scores

$$U_{ij} = \frac{1}{1+\rho}(y_i + y_j - 2\theta), \quad U_i = y_i - \theta, \quad U_j = y_j - \theta,$$

which yield

$$E[\nabla U_{ij}] = -\frac{2}{1+\rho}, \quad E[\nabla U_i] = E[\nabla U_j] = -1.$$

Thus,

$$\begin{aligned} E[U_{ij}^* U_{mle}] &= E[U_{ij}U_{mle}] - E[U_i U_{mle}] - E[U_j U_{mle}] \\ &= E[-\nabla U_{ij}] - E[-\nabla U_i] - E[-\nabla U_j] \\ &= \frac{-2\rho}{1+\rho}, \end{aligned}$$

Therefore, we can see that when $\rho \leq 0$, U_{ij}^* is not positively associated with U_{mle}.

Recall that $U_{ij}^* = U_{ij} - U_i - U_j$ is orthogonal to the class of all onewise marginal score functions under θ_{ind}. To modify the Hoeffding scores, we can replace

U_{ij}^* with the adjusted Hoeffding term

$$U_{ij}^{**} = U_{ij} - \beta_i(\theta)U_i - \beta_j(\theta)U_j,$$

where β's are $p \times p$ weight matrices which remove the marginal U_i and U_j effect. That is, U_{ij}^{**} is orthogonal to U_i and U_j under general θ. Then U_{ij}^{**} satisfies information identity as we show below.

To find $\beta_i(\theta)$ and $\beta_j(\theta)$, we need to minimize

$$\min_{\beta_i,\beta_j} E(U_{ij} - \beta_i U_i - \beta_j U_j)(U_{ij} - \beta_i U_i - \beta_j U_j)^T$$

Taking its partial derivative on β_i and β_j respectively, and then setting the derivatives equal zero, we have the following equations

$$\begin{cases} E[(U_{ij} - \beta_i U_i - \beta_j U_j)U_i^T] &= 0 \\ E[(U_{ij} - \beta_i U_i - \beta_j U_j)U_j^T] &= 0 \end{cases}$$

Further calculation shows that the above equations are equivalent to

$$\begin{cases} I_i - \beta_i I_i - \beta_j C_{ij} &= 0 \\ I_j - \beta_j C_{ij} - \beta_j I_j &= 0 \end{cases}$$

where $I_i = E(U_i U_i^T)$, $I_j = E(U_j U_j^T)$, and $C_{ij} = E(U_i U_j)$. Hence, the solutions are

$$\begin{cases} \beta_i &= (I_i - C_{ij})(I_i - C_{ij} I_j^{-1} C_{ij})^{-1} \\ \beta_j &= (I_j - C_{ij})(I_j - C_{ij} I_i^{-1} C_{ij})^{-1} \end{cases}$$

To illustrate that U_{ij}^{**} satisfies information identity, we consider the case in which θ is a scalar. In this situation, according to the discussion above, we have U_{ij}^{**} is orthogonal to U_i and U_j. Hence,

$$\begin{aligned} E(U_{ij}^{**2}) &= E[U_{ij}^{**}(U_{ij} - \beta_i U_i - \beta_j U_j)] \\ &= E[U_{ij}(U_{ij} - \beta_i U_i - \beta_j U_j)] \end{aligned}$$

Notice that

$$\begin{aligned} E[U_{mle}(U_{ij} - \beta_i U_i - \beta_j U_j)] &= E\{E[U_{mle}(U_{ij} - \beta_i U_i - \beta_j U_j)|Y_i, Y_j]\} \\ &= E\{(U_{ij} - \beta_i U_i - \beta_j U_j)E(U_{mle}|Y_i, Y_j)\} \\ &= E\{(U_{ij} - \beta_i U_i - \beta_j U_j)U_{ij}\} \end{aligned}$$

Thus, we have

$$\begin{aligned} E[U_{ij}(U_{ij} - \beta_i U_i - \beta_j U_j)] &= E[U_{mle}(U_{ij} - \beta_i U_i - \beta_j U_j)] \\ &= E(-\nabla U_{ij}^{**}), \end{aligned}$$

where the identity $E(U_{mle}g) + E(\nabla g) = 0$ (g is an unbiased estimating function) is used in last step.

Therefore, we prove $E(U_{ij}^{**2}) = E(-\nabla U_{ij}^{**})$, which induces that U_{ij}^{**} satisfies information identity.

In addition, because the adjusted Hoeffding term, U_{ij}^{**}, requires only pairwise distributions for the computation, we have kept the computational difficulty small meanwhile. Thus, we can define Modified Hoeffding scores as

Definition 3.5 (Modified Hoeffding scores)**.** *We define the Modified Hoeffding second order score as*

$$\sum_i U_i + \sum_{i<j} U_{ij}^{**}.$$

Because each term in Modified Hoeffding scores satisfies the information identity, they each have positive likelihood association. Thus, we have that Modified Hoeffding scores have positive likelihood association.

One simple example is that θ is a scalar. Then, the adjusted Hoeffding term is equivalent to regress U_{ij} on U_i and U_j, which gives β's as

$$\begin{pmatrix} \beta_i \\ \beta_j \end{pmatrix} = \begin{pmatrix} E(U_i^2) & E(U_iU_j) \\ E(U_iU_j) & E(U_j^2) \end{pmatrix}^{-1} \begin{pmatrix} E(U_i^2) \\ E(U_j^2) \end{pmatrix}$$

3.4 Multivariate normal example

In this section, we will consider examples of Hoeffding scores and Modified Hoeffding scores under the multivariate normal distribution. In addition, we will also compare the estimators obtained from Modified/Hoeffding scores with other estimators from true likelihood, pairwise likelihood, and all pairwise conditional likelihood.

3.4.1 Constant correlation model

Suppose $Y = (Y_1, ..., Y_d)^T \sim N_d(0, \Sigma)$, where $\Sigma = \sigma^2[(1-\rho)I_{d\times d} + \rho \mathbf{1}\mathbf{1}^T]$. Here, σ^2 is an unknown parameter and ρ is assumed to be a known constant. We mainly consider the following five types of scores:

- True score;
- Pairwise score;
- All conditional pairwise score;
- Hoeffding score;
- Modified Hoeffding score;

Firstly, we notice that all above scores can be represented as linear combination of onewise marginal and pairwise scores, $\alpha_1 \sum_i U_i + \alpha_2 \sum_{i<j} U_{ij}$, with different weights α_1 and α_2. We define the relative weight between onewise marginal and pairwise scores to be $-\alpha_1/\alpha_2$. The MLE, coming from true score, is the best estimator in the sense of asymptotic variance. Therefore, the constructed composite score will be optimal if the corresponding relative weight is close to the relative weight for the true score. The detailed comparison is found in the following table. Note that, true to the theory, the MLE and Hoeffding weights at $\rho = 0$ are identically $(d-2)$.

Table 3.1: Relative weights comparison for constant correlation model

Score	Rel. Wt.	Same as MLE
True	$\frac{d-2}{1+\rho}$	all ρ
Pairwise	0	never
All cond'ls	$\frac{d-1}{2}$	$\rho = 1 - \frac{2}{d-1}$
Hoeffding	$d-2$	$\rho = 0$
Mod. Hoeffding	$-\frac{\rho^2}{1+\rho^2} + \frac{d-2}{1+\rho^2}$	$\rho = 0, \rho \approx 1 - \frac{1}{d-1}$

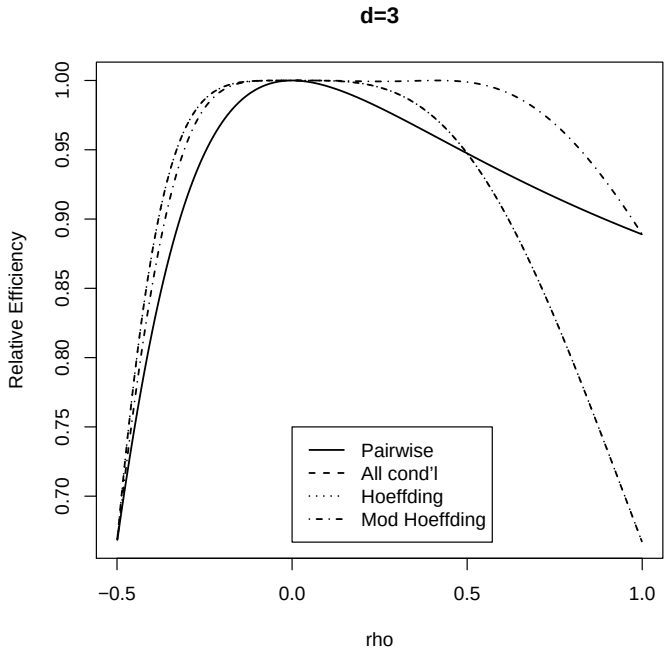
Secondly, under the normal distribution, all above five types of scores can also be written as $Y^T AY - E(Y^T AY)$ with different matrices A. In addition, the matrix A can be expressed as linear combination of A_{onewise} for onewise marginal score and A_{pairwise} for pairwise score according to the same weight α_1 and α_2 mentioned above. If the solution for the composite score equation is unique, we can use this solution as the composite score estimator. Therefore, we can compare the relative efficiency, r, between composite score estimator and MLE as following.

$$\begin{aligned} r &= \frac{Var(\hat{\theta}_{\text{mle}})}{Var(\hat{\theta}_{\text{cs}})} \\ &= \frac{(VarU)^{-1}}{VarU_{\text{cs}}/(EUU_{\text{cs}})^2} \\ &= \frac{(Cov(U, U_{\text{cs}}))^2}{VarU \cdot VarU_{\text{cs}}} \\ &= \frac{[tr(A_{\text{mle}}\Sigma A_{\text{cs}}\Sigma)]^2}{tr(A_{\text{mle}}\Sigma A_{\text{mle}}\Sigma) \cdot tr(A_{\text{cs}}\Sigma A_{\text{cs}}\Sigma)} \end{aligned}$$

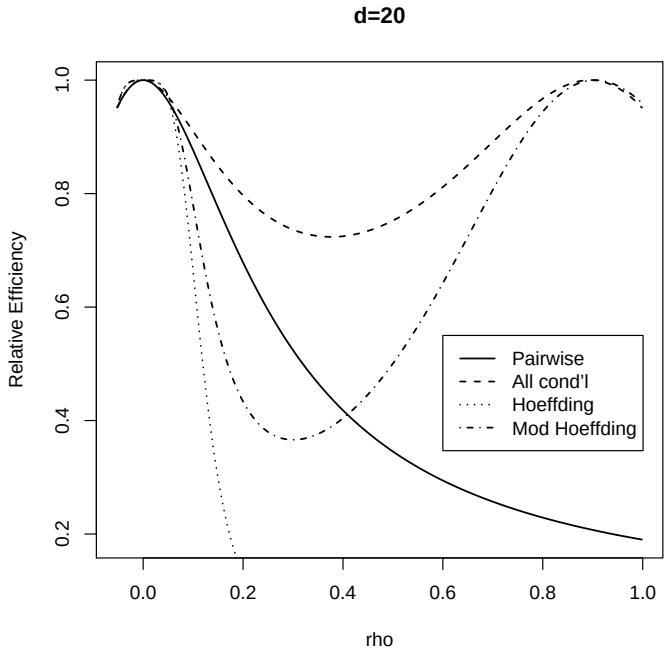
In the following, we compare the relative efficiencies for pairwise, all condition, Hoeffding, modified Hoeffding w.r.t. MLE when $d = 3, 20, 50$, under different values for $\rho \in (-1/(d-1), 1)$.

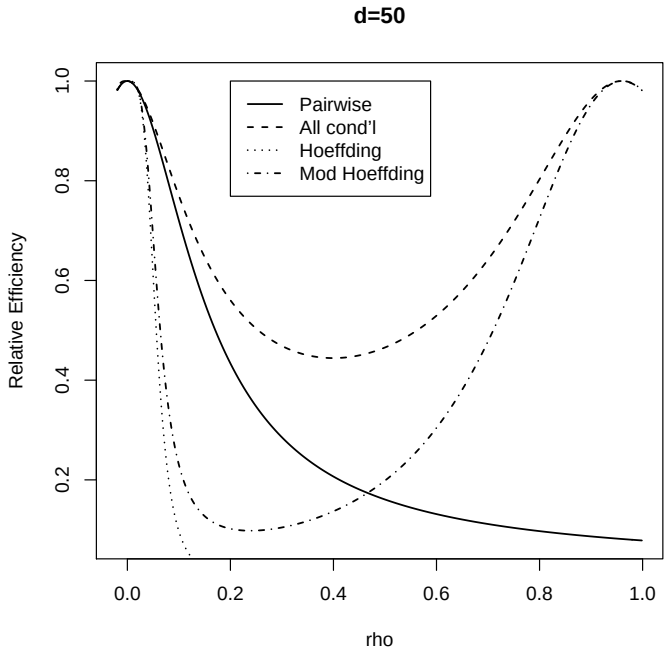
The first conclusion drawn from plots is that no matter what the values for d and ρ are, the maximum relative efficiency r is always 1. This means, we can always

find a best MCLE, which performs as well as MLE at $\rho = 0$. Another conclusion is that as d increase, the possible regions for ρ, where MCLE performs better, turn to be close to $\rho = 0$ or $\rho = 1$. As d increases, Modified Hoeffding performs better than Hoeffding which becomes noninformative for large ρ. However, it seems that the conditional pairwise likelihoods perform better overall for large d. Finally, notice that for $d = 3$, there is a local effect the makes the Hoeffding methods better near $\rho = 0$. However, this local effect diminishes as d increases.



When $d = 3$, all-conditionals is the same as Hoeffding


d=20
Relative Efficiency
rho
Pairwise
All cond'l
Hoeffding
Mod Hoeffding
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


d=50
Relative Efficiency
rho
Pairwise
All cond'l
Hoeffding
Mod Hoeffding
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

3.4.2 Constant partial correlation model

We consider a second simple example. Suppose $Y = (Y_1, ..., Y_d)^T \sim N_d(0, \Sigma)$, where $\Sigma^{-1} = \sigma^{-2}[(1-\beta)I_{d\times d} + \beta \mathbf{1}\mathbf{1}^T]$. Here, σ^2 is an unknown parameter and β is assumed to be a known constant.

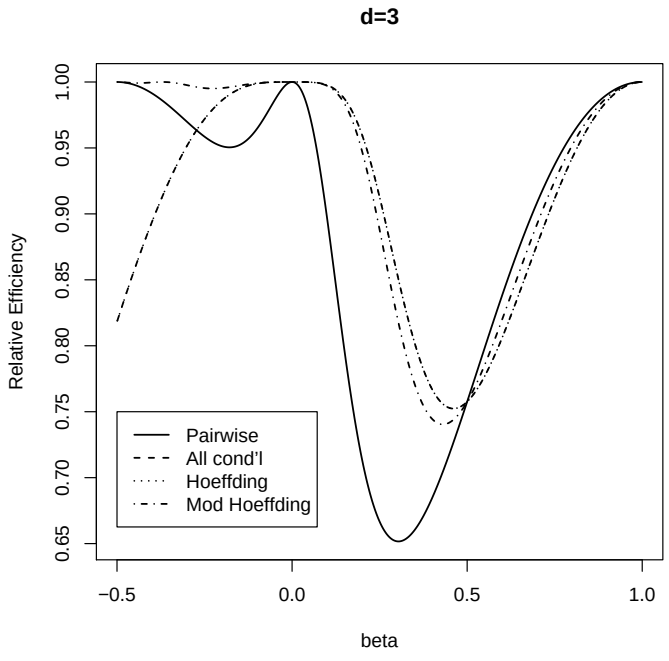
According to the same discussion as for constant correlation model, we have the following results for the relative weights (Table 3.2).

Table 3.2: Relative weights comparison for constant partial correlation model

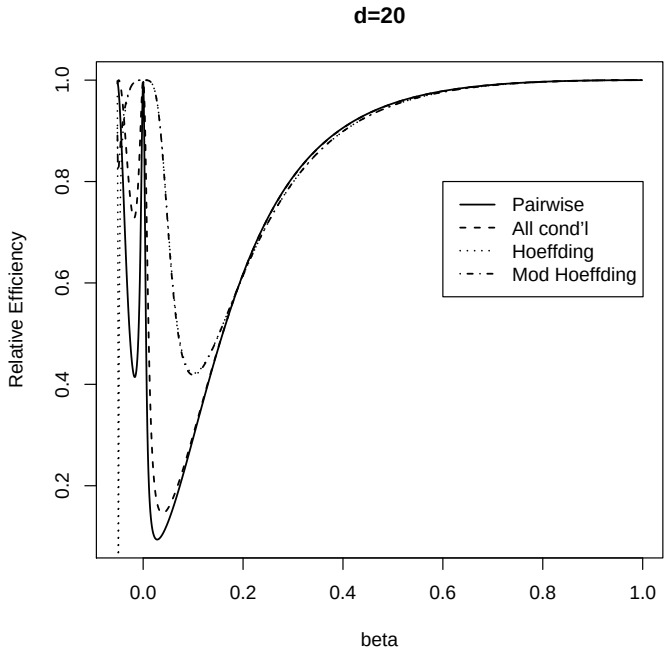
Score	Rel. Wt.	Same as MLE
True	$A(\beta)$	all β
Pairwise	0	never
All cond'ls	$\frac{d-1}{2}$	unknown?
Hoeffding	$B(\beta)$	$\beta = 0$
Mod. Hoeffding	$d-2$	$\beta = 0$,some unknown values?

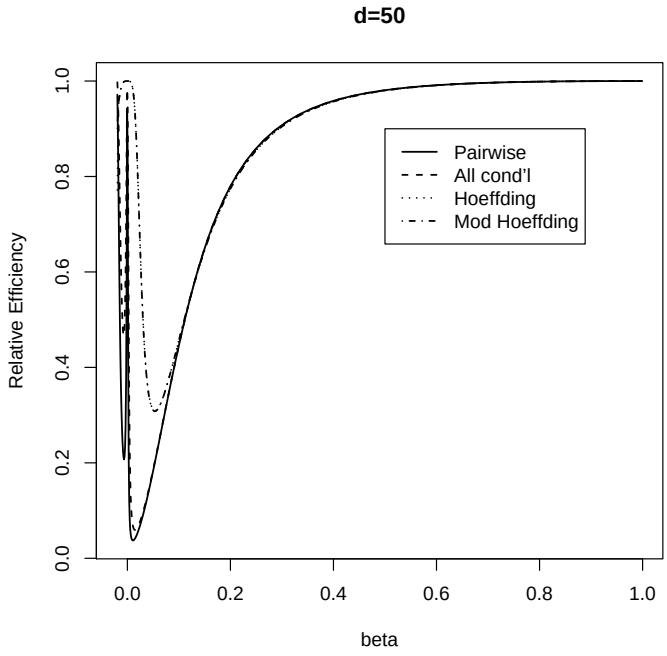
where $A(\beta)$ and $B(\beta)$ are quite complicated.

Again, we also compare the relative efficiencies for pairwise, all pairwise condition, Hoeffding, modified Hoeffding with respect to MLE when $d = 3, 20, 50$, under different values for $\beta \in (-1/(d-1), 1)$.



When $d = 3$, all-conditionals is the same as Hoeffding





When $d = 3$, both all-conditionals and modified Hoeffding performed similarly for $\beta \geq 0$, but modified Hoeffding was a clear winner for $\beta < 0$. For large

d, modified Hoeffding performed the best, although all methods had worrisome performance for ρ in the rage 0.05 to 0.2, but excellent performance near $\rho = 1$. Among the true composite likelihoods, all conditionals was best.

3.5 Hoeffding likelihood ratio

We now consider another potential extension of the Hoeffding projection method. It will not be developed further in this thesis; it is included as possible future work. In the previous three sections, we discussed how to construct the composite score by projecting the true score U_{mle} on some linear classes of estimating functions $\mathcal{H}_k$. The same idea can be applied to the log likelihood $\log L(\theta, y)$ itself. However, the corresponding projection is difficult to calculate because of the complexity of the term $E(\log L(\theta, y)|y_i)$. Instead we consider projecting the likelihood ratio, under θ_{ind}. Under independence, the projections have a simple form:

$$\begin{aligned} E_{\theta_{ind}}\left[\frac{L(y,\theta)}{L(y,\theta_{ind})}|y_i\right] &= \int \frac{f(y_i,\theta)f(y|y_i,\theta)}{f(y_i,\theta_{ind})f(y|y_i,\theta_{ind})} f(y|y_i,\theta_{ind})dX \\ &= \frac{f(y_i,\theta)}{f(y_i,\theta_{ind})}. \end{aligned}$$

Thus, we can project the likelihood ratio, $\frac{L(y,\theta)}{L(y,\theta_{ind})}$, on $\mathcal{H}_k$, $k = 1, 2, ...$ and obtain linear combinations of marginal likelihoods. The corresponding first and second projections are as follows:

$$\begin{aligned} R_1(\theta,\theta_{ind}) &= 1 + \sum_i \left(\frac{f(y_i,\theta)}{f(y_i,\theta_{ind})} - 1\right), \\ R_2(\theta,\theta_{ind}) &= R_1(\theta,\theta_{ind}) \\ &\quad + \sum_{i<j}\left(\frac{f(y_i,y_j,\theta)}{f(y_i,y_j,\theta_{ind})} - \frac{f(y_i,\theta)}{f(y_i,\theta_{ind})} - \frac{f(y_j,\theta)}{f(y_j,\theta_{ind})} + 1\right) \end{aligned}$$

Suppose we use the second projection, $R_2(\theta, \theta_{ind})$, to approximate the likelihood ratio. Then we have

$$\frac{L(\theta,y)}{L(\theta_{ind},y)} \approx R_2(\theta,\theta_{ind}).$$

If we consider θ_{ind} be be fixed, then we have the estimator $\widehat{\theta}$, which can maximize the LHS, must be the MLE. So, the estimator $\widehat{\theta}^*$, which can maximize the RHS might give a 'useful' approximation to the MLE.

Furthermore, we can choose an arbitrary independence model $L(y, \tau)$ with the same support as the dependence model $L(y, \theta)$. Suppose $\tau = \tau(\theta)$ is the choice of independence parameter so that $L(y, \tau(\theta))$ is closed to $L(y, \theta)$, or more strictly, the ratio of them is closed to 1. Then we suppose that $\frac{L(y,\theta)}{L(y,\tau(\theta))} \approx R_2(\theta, \tau(\theta))$. To recover a likelihood approximation, we let

$$L_2(y, \theta) \triangleq L(y, \tau(\theta)) R_2(\theta, \tau(\theta)) \approx L(y, \theta)$$

That suggests that we might use $L_2(y, \theta)$, a product of an independence model and a dependence correction term, as a new inference function to do the inference, which is computationally cheaper than $L(y, \theta)$.

Chapter 4

Applying Composite Likelihood to the Recombination Model

In chapter 1, we described a motivating problem for this thesis. It was to use a model for binary sequences with both mutation and recombination to estimate the ancestral distribution from a data set of current observed binary sequences. We also pointed out that the classical likelihood method is computationally infeasible for this problem. We then proposed to use composite likelihood method to solve the computation complexity. In the following two chapters, we discuss the properties of composite likelihood as related to this construction method. In this chapter, we will discuss how one can apply the composite likelihood method to the motivating problem in detail.

4.1 Model with only recombination

To explore the motivating problem of chapter 1, we proposed a model with both mutation and recombination to estimate the ancestral distribution. The strategy was to view the observed binary sequence, Y, as a mutated version of sequence X, where X is generated from ancestors with potential recombination. In Chen and Lindsay (2006), the authors have discussed how to estimate the ancestral distribution from the model with mutation only. In order to simplify our analysis, we will focus here on the inference on model with only recombination. That is, our goal will be to estimate the ancestral distribution from the observed sequences

X by considering only the recombination factor. This problem in itself has a high computational complexity.

4.1.1 Recombination model for sequence with length 2

Suppose the sequence length L is 2, then there are four possible types of ancestors sequences: (0,0), (0,1), (1,0), and (1,1). Assume the population frequencies for these four types of ancestors are $\pi_{1,2}(0,0)$, $\pi_{1,2}(0,1)$, $\pi_{1,2}(1,0)$, and $\pi_{1,2}(1,1)$, where $\pi_{1,2}(i,j) \geq 0$, $\sum_{i,j} \pi_{1,2}(i,j) = 1$, and $\pi_{1,2}(i,j)$ denote the frequency of the ancestor with value i on site 1 and value j on site 2, $i,j \in \{0,1\}$. Let Z represent the random vector (Z_1, Z_2) with density π, so that $P(Z_1 = i, Z_2 = j) = \pi_{1,2}(i,j)$. Let $X = (i,j)$ denote the observed sequence, and let q be the recombination rate $\in [0,1]$. In the model with recombination only, we have

$$P(X = (i,j)) = (1-q)\pi_{1,2}(i,j) + q\pi_1(i)\pi_2(j),$$

where $\pi_1(i) = \sum_j \pi_{1,2}(i,j)$ and $\pi_2(j) = \sum_i \pi_{1,2}(i,j)$. That is with probability $1-q$ the ancestor sequence came from a single ancestor, while with probability q a recombination occurred, and two independent draws were made.

4.1.2 Recombination model for sequence with length L

The above model can be easily generalized for sequence with length $L = 3$. When $L = 3$, there are two locations where recombination could occur. Thus, the recombination model is

$$\begin{aligned} P(X = (i,j,k)) \;=\; & (1-q)^2\pi_{1,2,3}(i,j,k) \\ & +q(1-q)[\pi_1(i)\pi_{2,3}(j,k) + \pi_{1,2}(i,j)\pi_3(k)] \\ & +q^2\pi_1(i)\pi_2(j)\pi_3(k). \end{aligned}$$

More generally, when the sequence length is L, there are $L-1$ locations

where recombination could occur. Thus the generalized recombination model is

$$
\begin{aligned}
& P(X = (x_1, x_2, \ldots, x_L)) \qquad (4.1)\\
= \;& (1-q)^{L-1}\pi_{1,2,\ldots,L}(x_1, x_2, \ldots, x_L)\\
+ \;& q(1-q)^{L-2}\sum_{s=1}^{L-1}\pi_{1,\ldots,s}(x_1, \ldots, x_s)\pi_{s+1,\ldots,L}(x_{s+1}, \ldots, x_L)\\
+ \;& q^2(1-q)^{L-3}\sum_{s_1=1}^{L-2}\sum_{s_2=s_1+1}^{L-1}[\pi_{1,\ldots,s_1}(x_1, \ldots, x_{s_1}) \times \pi_{s_1+1,\ldots,s_2}(x_{s_1+1}, \ldots, x_{s_2})\\
& \qquad \times \pi_{s_2+1,\ldots,L}(x_{s_2+1}, \ldots, x_L)]\\
+ \;& \cdots\\
+ \;& q^{L-1}\pi_1(x_1)\pi_2(x_2)\cdots\pi_L(x_L).
\end{aligned}
$$

Again, we will let $Z_1, Z_2, \ldots, Z_L$ be the random vector with the ancestor density $P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_L = z_L) = \pi_{1,2,\ldots,L}(z_1, z_2, \ldots, z_L)$.

4.2 Markov Chain Composite likelihood

As mentioned in chapter 1, though we can write out the above probability explicitly, there are significant computation challenges when applying the classical likelihood method to do the inference due to enormous number of recombination possibilities when L is large. Therefore, we propose to use composite likelihood as a substitute of full likelihood to reduce the computation complexity. In particularly, we will use a Markov chain composite likelihood (MCCL) to solve the computation problem for our desired statistical model.

Markov chains, especially continuous time Markov chains, are widely used in genetic evolution problems (Hjort and Varin (2007)) because they provide rich models, such as Jukes and Cantor (1969) and Kimura (1980), to describe the evolution of DNA sequences. Here, we will also build the Markov chain idea into our recombination model and hence to construct the MCCL as a substitute for the full likelihood of the recombination model.

There is a certain natural logic to use the MCCL in this problem. First, it greatly reduces the computational needs for finding estimates. Secondly, the action of recombination is to break the sequences into independent pieces. Thus, if there is any recombination at all between two sites a, b, the values of X_a and X_b are independent, and so the pair contain little information about the dependence in ancestral sequences at those two sites. Thus, we expect most of the recombination information to come from shorter sequences, and these will be built into our MCCL.

Suppose a sequence of random variables $X_1, X_2, \ldots, X_L$ is a Markov chain of order m, then we have

$$\begin{aligned} P(X_1, ..., X_L) &= P(X_1, ..., X_{L-1})P(X_L|X_{L-1}, ..., X_1) \\ &= P(X_1, ..., X_{L-1})P(X_L|X_{L-1}, ..., X_{L-m}) \\ &= P(X_1, ..., X_{L-2})P(X_{L-}|X_{L-2}, ..., X_{L-m-1})P(X_L|X_{L-1}, ..., X_{L-m}) \\ &= \cdots \\ &= P(X_1, ..., X_m)P(X_{m+1}|X_1, ..., X_m) \times \cdots \times P(X_L|X_{L-1}, ..., X_{L-m}). \end{aligned}$$

For example, if $m = 1$

$$P(X_1, ..., X_L) = P(X_1)P(X_2|X_1)P(X_3|X_2) \cdots P(X_L|X_{L-1});$$

and if $m = 2$, we have

$$P(X_1, ..., X_L) = P(X_1, X_2)P(X_3|X_2, X_1)P(X_4|X_3, X_2) \cdots P(X_L|X_{L-1}, X_{L-2}).$$

If we assume that there exists an order-m Markov chain structure on the observed n binary sequences, the the log likelihood function can be written as

$$\begin{aligned} l(\pi) &= \sum_{x_1,\ldots,x_L} n(x_1, \ldots, x_L) \log P(X_1 = x_1, X_2 = x_2, \ldots, X_L = x_L) \\ &= \sum_{x_1,\ldots,x_L} n(x_1, \ldots, x_L) \log[P(X_1 = x_1, ..., X_m = x_m)P(X_{m+1} = x_{m+1}|X_1 = x_1, ..., X_m = x_m) \\ &\qquad \times \cdots \times P(X_L = x_L|X_{L-1} = x_{L-1}, ..., X_{L-m} = x_{L-m})] \\ &= \sum_{x_1,\ldots,x_L} n(x_1, \ldots, x_L) \log \left[\frac{P(X_1 = x_1, ..., X_{m+1} = x_{m+1}) \cdots P(X_{L-m} = x_{L-m}, ..., X_L = x_L)}{P(X_2 = x_2, ..., X_{m+1} = x_{m+1}) \cdots P(X_{L-m} = x_{L-m}, ..., X_{L-1} = x_{L-1})} \right], \end{aligned}$$

where $n(x_1, \ldots, x_L)$ denote the number of observations which have sequence as $X = (x_1, \ldots, x_L)$, $x_s \in \{0, 1\}$ for $s = 1, ..., L$; and $P(X_s = x_s, X_{s+1} = x_{s+1}, ..., X_t = x_t)$ can be calculated by recombination model (4.1) in the above section.

Thus, we can use the following log likelihood function as a surrogate for the full log likelihood, and call this surrogate log likelihood as the log Markov chain composite likelihood (log MCCL). It is a good surrogate if the observed sequences have the Markov property. In addition, the MCCL converges to full likelihood when the order, m, of Markov chain order increases.

$$l(\pi) = \sum_{x_1,...,x_L} n(x_1, \ldots, x_L) \log \left[\frac{f(x_1, ..., x_{m+1}) f(x_2, ..., x_{m+2}) \cdots f(x_{L-m}, ..., x_L)}{f(x_2, ..., x_{m+1}) \cdots f(x_{L-m}, ..., x_{L-1})} \right], \quad (4.2)$$

where

$$f(x_s, x_{s+1}..., x_t) = P(X_s = x_s, X_{s+1} = x_{s+1}, \ldots, X_t = x_t)$$

can be calculated by recombination probability in equation (4.1).

From the log MCCL (4.2), we can see that if we use the MCCL in stead of the full likelihood to do the inference, in fact, we can estimate only the $(L - m)$ different sequential margins of ancestral distribution, i.e. $\pi_{s,...,s+m}(x_s, ..., x_{s+m})$ for $s = 1, ..., L - m$, not the full joint ancestral distribution $\pi_{1,...,L}(x_1, \ldots, x_L)$, $x_s \in \{0, 1\}$ for $s = 1, \ldots, L$. In order to estimate this full joint ancestral distribution, we could make an additional assumption that the joint density also has an order-m Markov property. If so, we can estimate the joint after obtaining the $(m+1)$-wise marginal ancestral distribution estimations, by

$$\begin{aligned} & \hat{\pi}_{1,...,L}(x_1, \ldots, x_L) \\ = \; & \hat{\pi}_{1,...,m}(x_1, ..., x_m) \hat{\pi}_{m+1|m,...,1}(x_{m+1}|x_m, ..., x_1) \cdots \hat{\pi}_{L|L-1,...,L-m}(x_L|x_{L-1}, ..., x_{L-m}) \\ = \; & \frac{\hat{\pi}_{1,...,m+1}(x_1, ..., x_{m+1}) \hat{\pi}_{2,...,m+2}(x_2, ..., x_{m+2}) \cdots \hat{\pi}_{L-m,...,L}(x_{L-m}, ..., x_L)}{\hat{\pi}_{2,...,m+1}(x_2, ..., x_{m+1}) \cdots \hat{\pi}_{L-m,...,L-1}(x_{L-m}, ..., x_{L-1})} \end{aligned} \quad (4.3)$$

Of course, in practice we will view (4.3) as an approximation to the joint density, one that improves in quality as m increases.

Remark: In this thesis, we have paired the MCCL objective function together with a Markov Chain assumptions on π that can be used if one wishes reconstruct the full joint distribution of π from its margins. An alternative approach to this problem is to start by assuming that π is a Markov Chain. In this case, the data sequence $X_1, \ldots, X_L$ is a hidden Markov Chain. This would make the computation feasible for the full likelihood. As part of our future work, we plan to investigate this approach and compare it with MCCL.

4.3 Challenges

Though we can reduce the computation complexity by using MCCL instead of the full likelihood, there are a number of challenges to find the estimator that maximizes the MCCL.

In the MCCL, both $(m+1)$-wise margins, $\pi_{s,s+1,\ldots,s+m}(x_s, i_{s+1}, \ldots, x_{s+m})$, and m-wise margins, $\pi_{s+1,\ldots,s+m}(x_{s+1}, \ldots, x_{s+m})$ for $s = 1, 2, \ldots, L-m$, are unknown. But, since the m-wise margins can be calculated from the $(m+1)$-wise margins directly, the unknown parameters in the MCCL are the $L-m$ different $(m+1)$-wise margins of ancestral distribution,

$$\pi_{s,s+1,\ldots,s+m}(x_s, x_{s+1}, \ldots, x_{s+m}), \text{ for } s = 1, 2, \ldots, L-m.$$

If we view these as our parameters, there are severe restrictions on the parameter space. The easy one is the *between zero and one* constraint, which means that

$$0 \leq \pi_{s,s+1,\ldots,s+m}(x_s, x_{s+1}, \ldots, x_{s+m}) \leq 1, \text{ for } s = 1, 2, \ldots, L-m.$$

Besides this constraint, the parameters have to satisfy *lower order margin consistency* property. For example, the estimated $(m+1)$-wise margins must satisfy

$$\begin{aligned} & \pi_{s+1,\ldots,s+m}(x_{s+1},\ldots,x_{s+m}) \\ = & \sum_{x_s=0}^{1} \pi_{s,s+1,\ldots,s+m}(x_s,x_{s+1},\ldots,x_{s+m}) \\ = & \sum_{x_{s+m+1}=0}^{1} \pi_{s+1,\ldots,s+m+1}(x_{s+1},\ldots,x_{s+m+1}), \end{aligned}$$

for $s = 1, 2, \ldots, L-m-1$.

These complicated constraints make it difficult to build an estimator that will maximize the MCCL within the parameter space. Moreover, the optimization problem is very high dimensional as a function of L. When $s = 1$, to estimate the $(m+1)$-wise margin $\pi_{1,\ldots,m+1}(x_1,\ldots,x_{m+1})$, we have $2^{m+1}-1$ free parameters because of the 'sum to one' constraint, i.e.

$$\sum_{x_1=0}^{1} \cdots \sum_{x_{m+1}=0}^{1} \pi_{1,\ldots,m+1}(x_1,\ldots,x_{m+1}) = 1.$$

When $s = 2$, there are in total 2^{m+1} unknown parameters, $\pi_{2,\ldots,m+2}(x_2,\ldots,x_{m+2})$. Except for the 'sum to one' constraint, there also exist $2^m - 1$ lower margin consistency constraints, that is

$$\sum_{x_1=0}^{1} \pi_{1,\ldots,m+1}(x_1,\ldots,x_{m+1}) = \sum_{x_{m+2}=0}^{1} \pi_{2,\ldots,m+2}(x_2,\ldots,x_{m+2}),$$

for all the combinations of $x_2,\ldots,x_{m+1}$. Thus, there are $2^{m+1}-1-(2^m-1) = 2^m$ free parameters to estimate. Similarly, we have 2^m free parameters to estimate the $(m+1)$-wise margins $\pi_{s,\ldots,s+m}(x_s,\ldots,x_{s+m})$ when $s = 2,\ldots,L-m$. Therefore, in total we have $(2^{m+1}-1)+2^m(L-m-1)$ free parameters. For this reason we have focused on this problem first from the perspective of finding quick (and possibly inefficient) methods to come near to maximizing the MCCL for long sequences L.

4.4 A left to right estimator

In this section, we will propose a fast estimate method, *left to right estimator*, which is used to find the estimator for $(m+1)$-wise margins, $\pi_{s,\ldots,s+m}(x_s,\ldots,x_{s+m})$ where $s=1,2,\ldots,L-m$, sequentially.

4.4.1 A reparametrization

In order to demonstrate our proposed estimator clearly, we will first discuss the reparametrized system that is used in our left to right estimator.

Suppose we want find the estimates of $L-m$ different $(m+1)$-wise marginal ancestral distribution, $\pi_{s,\ldots,s+m}(x_s,\ldots,x_{s+m})$ where $s=1,2,\ldots,L-m$ and $x_s\in\{0,1\}$, from order-m log MCCL in equation (4.2). We define new parameters

$$\theta_{s+1}(\underline{x}) \;=\; P(Z_{s:s+m}=(\underline{x},0)|Z_{s:s+m-1}=(\underline{x})),$$

for $s=1,2,\ldots,L-m$, where $\underline{x}=(x_s,\ldots,x_{s+m-1})$ is a $1\times m$ vector, $Z_{s:s+m-1}(\underline{x})=(Z_s=x_s,\ldots,Z_{s+m-1}=x_{s+m-1})$ is also a $1\times m$ random vector with ancestral density

$$P(Z_{s:s+m-1}(\underline{x}))=\pi_{s,\ldots,s+m-1}(x_s,\ldots,x_{s+m-1})$$

and $Z_{s:s+m}=(\underline{x},0)$ is a $1\times(m+1)$ random vector with ancestral density

$$P(Z_{s:s+m}(\underline{x},0))=\pi_{s,\ldots,s+m-1,s+m}(x_s,\ldots,x_{s+m-1},0).$$

Thus, we reparameterize the parameters $\pi_{s,\ldots,s+m}(x_s,\ldots,x_{s+m})$ into $\theta_{s+1}(\underline{x})$, where $\theta_{s+1}(\underline{x})\in[0,1]$ and $\underline{x}=(x_s,\ldots,x_{s+m-1})$. In addition, the above reparametrization is 1-1 mapping. For example, consider order-1 log MCCL,

$$l(\pi) \;=\; \sum_{x_1=0}^{1}\cdots\sum_{x_L=0}^{1} n(x_1,\ldots,x_L)\log\left\{\frac{f(x_1,x_2)f(x_2,x_3)\cdots f(x_{L-1},x_L)}{f(x_2)f(x_3)\cdots f(x_L)}\right\} \quad (4.4)$$

$$\;=\; \sum_{x_1=0}^{1}\cdots\sum_{x_L=0}^{1} n(x_1,\ldots,x_L)\log[f(x_1)f(x_2|x_1)f(x_3|x_2)\cdots f(x_L|x_{l-1})] \quad (4.5)$$

where

$$\begin{aligned}
f(x_s, x_{s+1}) &= P(X_s = x_s, X_{s-1} = x_{s-1}) \\
&= (1-q)\pi_{s,s+1}(x_s, x_{s+1}) + q\pi_s(x_s)\pi_{s+1}(x_{s+1}) \\
f(x_s) &= P(X_s = x_s) = \pi_s(x_s) \\
f(x_{s+1}|x_s) &= f(x_s, x_{s+1})/f(x_s) \\
&= P(X_s = x_s, X_{s-1} = x_{s-1})/P(X_s = x_s) \\
&= P(X_s = x_s, X_{s-1} = x_{s-1}|X_s = x_s) \\
&= (1-q)\pi_{s+1|s}(x_{s+1}|x_s) + q\pi_{s+1}(x_{s+1}).
\end{aligned}$$

By reparametrization, our parameters of interest change from $\pi_{1,2}(x_1, x_2), \ldots,$ $\pi_{L-1,L}(x_{L-1}, x_L)$ to $\pi_1(x_1), \theta_2(x_1), \ldots, \theta_L(x_{L-1})$. All $\pi_{s,s+1}(x_s, x_{s+1})$'s for $s = 1, \ldots, L-1$ can be written as

$$\begin{aligned}
\pi_{s,s+1}(x_s, x_{s+1}) &= \pi_{s+1|s}(x_{s+1}|x_s)\pi_s(x_s) \\
&= \pi_s(x_s)\left[\theta_{s+1}(x_s)\right]^{1-x_{s+1}}\left[1-\theta_{s+1}(x_s)\right]^{x_{s+1}}.
\end{aligned}$$

Hence, the new parameters are clearly a 1-1 reparametization. However, there is still a computational problem with this parametrization because the margins, $\pi_s(x_s)$, needed in the log MCCL, are an increasingly complicated function of $\pi_1(x_1)$ and the proceeding $\theta_{s+1}(x_s)$ values as we proceed from left to right. For example,

$$\begin{aligned}
\pi_2(0) &= \pi_1(0)\pi_{2|1}(0|0) + \pi_1(1)\pi_{2|1}(0|1) \\
&= \pi_1(0)\theta_2(0) + \pi_1(1)\theta_2(1) \\
\pi_3(0) &= \pi_2(0)\pi_{3|2}(0|0) + \pi_2(1)\pi_{3|2}(0|1) \\
&= \left[\pi_1(0)\theta_2(0) + \pi_1(1)\theta_2(1)\right]\theta_3(0) + \left\{1 - \left[\pi_1(0)\theta_2(0) + \pi_1(1)\theta_2(1)\right]\right\}\theta_3(1) \\
\pi_4(0) &= \pi_3(0)\pi_{4|3}(0|0) + \pi_3(1)\pi_{4|3}(0|1) \\
&= \cdots
\end{aligned}$$

This gives the terms in the MCCL a highly nonseparable nature. The goal of creating a fast method for this problem led us to the following sequential maximization.

4.4.2 The estimator

The estimator proposed here does not provide a complete maximization of log MCCL (4.2). Instead it does a sequential maximization in which we update the parameters from left to right using a conditional maximization.

We will illustrate this process as follows: we first obtain the estimation of the leftmost margin $\hat{\pi}_{1,\ldots,m}(x_1,\ldots,x_m)$. Using the leftmost factor of the MCCL, we then estimate $\pi_{1,\ldots,m,m+1}(x_1,\ldots,x_m,x_{m+1})$ given $\hat{\pi}_{1,\ldots,m}(x_1,\ldots,x_m)$ fixed. This optimization only depends on $\theta_2(\underline{x})$, where $\underline{x}=(x_1,\ldots,x_m)$. We then calculate $\hat{\pi}_{2,\ldots,m+1}(i_2,\ldots,i_{m+1})$ by summing over the first index in $\hat{\pi}_{1,\ldots,m+1}(x_1,\ldots,x_{m+1})$. Holding this fixed, we estimate $\pi_{2,\ldots,m+2}(x_2,\ldots,x_{m+2})$. This now depends only on $\theta_3(\underline{x})$, where $\underline{x}=(x_2,\ldots,x_{m+1})$. Repeat this sequential procedure from the left side to the right side of the sequence, until finally we have all estimated the unknown parameters.

Pairwise margin

To illustrate this process in greater detail in the simplest situation, let us discuss the situation where order-1 log MCCL is applied to recombination model. Under this situation, the target log likelihood is as (4.5).

The left-right estimator begins from estimating the onewise marginal ancestral distribution at the first site on the left of the ancestor sequences, i.e. $\pi_1(x_1)$ for $x_1\in\{0,1\}$. Since for onewise margin, we do not need to take into account of the recombination factor between sites, it is natural to use sample proportion as the estimation for $\pi_1(x_1)$. Thus, we have $\hat{\pi}_1(x_1)=n_1(x_1)/n$, where $n_1(x_1)$ denotes the number of observations that have value x_1 on the first site. This corresponds

to maximizing $\sum_{x_1} n_1(x_1) \log f(x_1)$ to find $\hat{\pi}_1(x_1)$.

After obtaining $\hat{\pi}_1(x_1)$, we estimate $\pi_{1,2}(x_1, x_2)$, where $x_1, x_2 \in \{0, 1\}$, by fixing $\hat{\pi}_1(x_1)$ and maximizing $\sum_{x_1,x_2} n_{1,2}(x_1, x_2) \log f(x_2|x_1)$ which is the next term in the log MCCL (4.5). Suppose that $\pi_{1,2}(0, 0) = \theta_2(0)\pi_1(0)$, for $\theta_2(0) \in [0, 1]$. It follows that $\pi_{1,2}(0, 1) = [1 - \theta_2(0)]\, \pi_1(0)$. Similarly, we can define $\pi_{1,2}(1, 0) = \theta_2(1)\pi_1(1)$ and $\pi_{1,2}(1, 1) = [1 - \theta_2(1)]\, \pi_1(1)$ for $\theta_2(1) \in [0, 1]$. Hence, we can rewrite the unknown parameters $\pi_{1,2}(x_1, x_2)$ in terms of $\theta_2 = [\theta_2(0), \theta_2(1)]^T$ as below:

$$\pi_{1,2}(0, 0) = \theta_2(0)\hat{\pi}_1(0), \;\; \pi_{1,2}(0, 1) = (1 - \theta_2(0))\hat{\pi}_1(0)$$
$$\pi_{1,2}(1, 0) = \theta_2(1)\hat{\pi}_1(1), \;\; \pi_{1,2}(1, 1) = (1 - \theta_2(1))\hat{\pi}_1(1).$$

When we have the estimates of $\theta_2 = [\theta_2(0), \theta_2(1)]^T$, we can solve for the next margins as

$$\begin{aligned}
\hat{\pi}_2(0) &= \hat{\pi}_{1,2}(0, 0) + \hat{\pi}_{1,2}(1, 0) = \hat{\theta}_2(0)\hat{\pi}_1(0) + \hat{\theta}_2(1)\hat{\pi}_1(1) \\
\hat{\pi}_2(1) &= \hat{\pi}_{1,2}(0, 1) + \hat{\pi}_{1,2}(1, 1) = [1 - \hat{\theta}_2(0)]\hat{\pi}_1(0) + [1 - \hat{\theta}_2(1)]\hat{\pi}_1(1)
\end{aligned}$$

The next unknown parameters are $\theta_3 = [\theta_3(0), \theta_3(1)]^T$. Following the similar strategy as for θ_2, we estimate θ_3 by maximizing $\sum_{x_2,x_3} n_{2,3}(x_2, x_3) \log f(x_3|x_2)$, the third term in log MCCL (4.5), given $\hat{\pi}_2(x_2)$ fixed, and then calculate $\hat{\pi}_3(x_3)$. Repeat this process until we estimate all the unknown parameters θ_{s+1}, and consequently $\pi_{s,s+1}(x_s, x_{s+1})$ for $s = 1, \ldots, L$.

In such a way, as long as the estimate of $\pi_1(x_1)$ satisfies $0 \leq \hat{\pi}_1(x_1) \leq 1$, all the estimated pairwise margins satisfy $0 \leq \hat{\pi}_{s,s+1}(x_1, x_2) \leq 1$. Moreover, the above strategy also ensures that all the pairwise margins satisfy the lower margin consistency constraint. In addition, note that $\hat{\theta}_{s+1}(0) = \hat{\pi}_{s+1|s}(0|0)$ and $\hat{\theta}_{s+1}(1) = \hat{\pi}_{s+1|s}(0|1)$, they can be used directly when we reconstruct the ancestral distribution $\pi_{1,\ldots,L}$ based on equation (4.3) by using order-1 Markov chain property.

4.4.3 Details of the MCCL optimization

In this subsection, we present in detail how to estimate $\theta_{s+1} = [\theta_{s+1}(0), \theta_{s+1}(1)]^T$ from the log likelihood (4.4). To estimate θ_{s+1}, the corresponding log likelihood to be maximized is.

$$\begin{aligned} l(\theta_{s+1}) &= \sum_{x_s, x_{s+1}} n_{s,s+1}(x_s, x_{s+1}) \log f(x_{s+1}|x_s) \qquad (4.6) \\ &= n_{s,s+1}(0,0) \log\{(1-q)\theta_{s+1}(0)\hat{\pi}_s(0) + q\hat{\pi}_s(0)[\theta_{s+1}(0)\hat{\pi}_s(0) + \theta_{s+1}(1)\hat{\pi}_s(1)]\} \\ &\quad + n_{s,s+1}(0,1) \log\{(1-q)(1-\theta_{s+1}(0))\hat{\pi}_s(0) + q\hat{\pi}_s(0)[(1-\theta_{s+1}(0))\hat{\pi}_s(0) + (1-\theta_{s+1}(1))\hat{\pi}_s(1)]\} \\ &\quad + n_{s,s+1}(1,0) \log\{(1-q)\theta_{s+1}(1)\hat{\pi}_s(1) + q\hat{\pi}_s(1)[\theta_{s+1}(0)\hat{\pi}_s(0) + \theta_{s+1}(1)\hat{\pi}_s(1)]\} \\ &\quad + n_{s,s+1}(1,1) \log\{(1-q)(1-\theta_{s+1}(1))\hat{\pi}_s(1) + q\hat{\pi}_s(1)[(1-\theta_{s+1}(0))\hat{\pi}_s(0) + (1-\theta_{s+1}(1))\hat{\pi}_s(1)]\}, \end{aligned}$$

where $\theta_{s+1}(i) \in [0,1]$ for $i = 0, 1$.

Lemma 4.1. *The above log likelihood function* (4.6) *is a concave function respect to* $\theta_{s+1}(0)$ *and* $\theta_{s+1}(1)$.

Proof. Denote

$$\begin{aligned} f^1_{\theta_{s+1}} &\triangleq f_{\theta_{s+1}}(x_s = 0, x_{s+1} = 0) \\ &= (1-q)\theta_{s+1}(0)\hat{\pi}_s(0) + q\hat{\pi}_s(0)[\theta_{s+1}(0)\hat{\pi}_s(0) + \theta_{s+1}(1)\hat{\pi}_s(1)] \\ f^2_{\theta_{s+1}} &\triangleq f_{\theta_{s+1}}(x_s = 0, x_{s+1} = 1) \\ &= (1-q)(1-\theta_{s+1}(0))\hat{\pi}_s(0) + q\hat{\pi}_s(0)[(1-\theta_{s+1}(0))\hat{\pi}_s(0) + (1-\theta_{s+1}(1))\hat{\pi}_s(1)] \\ f^3_{\theta_{s+1}} &\triangleq f_{\theta_{s+1}}(x_s = 1, x_{s+1} = 0) \\ &= (1-q)\theta_{s+1}(1)\hat{\pi}_s(1) + q\hat{\pi}_s(1)[\theta_{s+1}(0)\hat{\pi}_s(0) + \theta_{s+1}(1)\hat{\pi}_s(1)] \\ f^4_{\theta_{s+1}} &\triangleq f_{\theta_{s+1}}(x_s = 1, x_{s+1} = 1) \\ &= (1-q)(1-\theta_{s+1}(1))\hat{\pi}_s(1) + q\hat{\pi}_s(1)[(1-\theta_{s+1}(0))\hat{\pi}_s(0) + (1-\theta_{s+1}(1))\hat{\pi}_s(1)]. \end{aligned}$$

The log MCCL (4.6) can be simplified as

$$l(\theta_{s+1}) = \sum_{i=1}^{4} n^i \log f^i_{\theta_{s+1}},$$

where $n^1 = n_{s,s+1}(0,0)$, $n^2 = n_{s,s+1}(0,1)$, $n^3 = n_{s,s+1}(1,0)$, $n^4 = n_{s,s+1}(1,1)$.

Since we have

$$\begin{array}{ll}
\frac{\partial}{\partial\theta_{s+1}(0)} f^1_{\theta_{s+1}} = (1-q)\hat{\pi}_s(0) + q\hat{\pi}_s^2(0), & \frac{\partial}{\partial\theta_{s+1}(1)} f^1_{\theta_{s+1}} = q\hat{\pi}_s(0)\hat{\pi}_s(1) \\
\frac{\partial}{\partial\theta_{s+1}(0)} f^2_{\theta_{s+1}} = -\frac{\partial}{\partial\theta_{s+1}(0)} f^1_{\theta_{s+1}}, & \frac{\partial}{\partial\theta_{s+1}(1)} f^2_{\theta_{s+1}} = -\frac{\partial}{\partial\theta_{s+1}(1)} f^1_{\theta_{s+1}} \\
\frac{\partial}{\partial\theta_{s+1}(0)} f^3_{\theta_{s+1}} = q\hat{\pi}_s(0)\hat{\pi}_s(1), & \frac{\partial}{\partial\theta_{s+1}(1)} f^3_{\theta_{s+1}} = (1-q)\hat{\pi}_s(1) + q\hat{\pi}_s^2(1) \\
\frac{\partial}{\partial\theta_{s+1}(0)} f^4_{\theta_{s+1}} = -\frac{\partial}{\partial\theta_{s+1}(0)} f^3_{\theta_{s+1}}, & \frac{\partial}{\partial\theta_{s+1}(1)} f^4_{\theta_{s+1}} = -\frac{\partial}{\partial\theta_{s+1}(1)} f^3_{\theta_{s+1}};
\end{array}$$

the gradient matrix of the above log likelihood is

$$\nabla l(\theta_{s+1}) = \begin{pmatrix} \frac{\partial}{\partial\theta_{s+1}(0)} l(\theta_{s+1}) \\ \frac{\partial}{\partial\theta_{s+1}(1)} l(\theta_{s+1}) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^4 n^i \frac{\frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \\ \sum_{i=1}^4 n^i \frac{\frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \end{pmatrix}$$

In addition, because

$$\frac{\partial^2}{\partial\theta_{s+1}(0)^2} f^i_{\theta_{s+1}} = 0, \ \frac{\partial^2}{\partial\theta_{s+1}(1)^2} f^i_{\theta_{s+1}} = 0, \text{ and } \frac{\partial^2}{\partial\theta_{s+1}(0)\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}} = 0,$$

for $i = 1, 2, 3, 4$. Thus, the Hessian matrix for the above log likelihood function is

$$\begin{aligned}
\nabla^2 l(\theta_{s+1}) &= \begin{pmatrix} \frac{\partial^2}{\partial\theta_{s+1}(0)^2} l(\theta_{s+1}) & \frac{\partial^2}{\partial\theta_{s+1}(0)\partial\theta_{s+1}(1)} l(\theta_{s+1}) \\ \frac{\partial^2}{\partial\theta_{s+1}(0)\partial\theta_{s+1}(1)} l(\theta_{s+1}) & \frac{\partial^2}{\partial\theta_{s+1}(1)^2} l(\theta_{s+1}) \end{pmatrix} \\
&= -\begin{pmatrix} \sum_{i=1}^4 n^i \left(\frac{\frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \right)^2 & \sum_{i=1}^4 n^i \frac{\frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}} \frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{(f^i_{\theta_{s+1}})^2} \\ \sum_{i=1}^4 n^i \frac{\frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}} \frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{(f^i_{\theta_{s+1}})^2} & \sum_{i=1}^4 n^i \left(\frac{\frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \right)^2 \end{pmatrix} \\
&\triangleq -H.
\end{aligned}$$

The determinant of matrix H is

$$\begin{aligned}
|H| &= \left[\sum_{i=1}^4 n^i \left(\frac{\frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \right)^2 \right] \left[\sum_{i=1}^4 n^i \left(\frac{\frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \right)^2 \right] - \left[\sum_{i=1}^4 n^i \frac{\frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}} \frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{(f^i_{\theta_{s+1}})^2} \right]^2 \\
&= \left[\sum_{i=1}^4 \left(\sqrt{n^i} \frac{\frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \right)^2 \right] \left[\sum_{i=1}^4 \left(\sqrt{n^i} \frac{\frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{f^i_{\theta_{s+1}}} \right)^2 \right] - \left[\sum_{i=1}^4 \frac{\sqrt{n^i} \frac{\partial}{\partial\theta_{s+1}(0)} f^i_{\theta_{s+1}} \sqrt{n^i} \frac{\partial}{\partial\theta_{s+1}(1)} f^i_{\theta_{s+1}}}{(f^i_{\theta_{s+1}})^2} \right]^2 \\
&\geq 0,
\end{aligned}$$

where the last inequality is because of Cauchy-Schwarz inequality

$$\left|\sum_{i=1}^{n} a_i b_i\right|^2 \leq \sum_{j=1}^{n} |a_j|^2 \sum_{k=1}^{n} |b_k|^2,$$

for series of number a_j and b_k.

Thus, the determinants of all the ordered principal submatrices of matrix H are non-negative, and hence the matrix H is semi-positive definite, which implies the Hessian matrix $\nabla^2 l(\theta_{s+1})$ is semi-negative definite.

Therefore, the log likelihood function (4.6) is a concave function with respect to θ_{s+1}. □

The above lemma proved that the Hessian matrix of log likelihood function (4.6) is semi-negative definite. Thus, there exists a local maximum for $l(\theta_{s+1})$ and we can estimate θ_{s+1} by solving score equation

$$\nabla l(\theta_{s+1}) = 0$$

It is not easy to write out the closed form solution for above score equation, but because we can write out the gradient matrix and Hessian matrix of the log likelihood function (4.6) explicitly, we can use numerical method such as Newton-Raphson method to find the maximizer.

Let θ_{s+1}^c denote the current θ_{s+1}. The updated θ_{s+1} by using Newton-Raphson method is

$$\theta_{s+1}^{new} = \theta_{s+1}^c - [\nabla^2 l(\theta_{s+1}^c)]^{-1} \nabla l(\theta_{s+1}^c),$$

and the iteration for updating θ_{s+1} will stop until the log likelihood function (4.6) doesn't increase.

There are several issues need to be considered when applying Newton-Raphson method. First, the initial value for θ_{s+1}. Since $\theta_{s+1} = [\theta_{s+1}(0), \theta_{s+1}(1)]^T$,

where $\theta_{s+1}(i) \in [0,1]$ for $i = 0, 1$, it is reasonable to choose the number within $[0,1]$ interval, for example $1/2$, as the initial value. This is what we did in the simulation study.

Second, because the parameter space for θ_{s+1} is restricted, the local maximum may not be achieved within the parameter space. Under this situation, we stopped the updating iteration at the closest bound of either $\theta_{s+1}^{new}(0)$ or $\theta_{s+1}^{new}(1)$, depending on which $\theta_{s+1}^{new}(i)$ was outside the parameter space. Then we only did the Newton update for the other $\theta_{s+1}(j)$, which was not outside the boundary. Finally, during the updating iteration, it is possible that one of the $f_{\theta_{s+1}}$'s becomes zero. In such situation, we will have no definitions for $\nabla l(\theta_{s+1})$ and $\nabla^2 l(\theta_{s+1})$, and the log likelihood will be negative infinity. To avoid this situation, in the simulation study we added a very small number, $\epsilon = 10^{-10}$, to the diagonal elements of the Hessian matrix. The Newton method then behaved smoothly without the interruption of zero denominator problems.

4.4.4 Estimator for $m \geq 3$

The above pairwise margins estimator can be generalized to estimate threewise margins directly. The only difference is that when the estimates for $\pi_{1,2}(x_1, x_2)$ is fixed as $\hat{\pi}_{1,2}(x_1, x_2)$, to estimate the threewise margin $\pi_{1,2,3}(x_1, x_2, x_3)$, the reparametrization will be $\theta_2 = [\theta_2(0,0), \theta_2(0,1), \theta_2(1,0), \theta_2(1,1)]^T$, where $\theta_2(x_1, x_2) \in [0,1]$, and

$$\pi_{1,2,3}(0,0,0) = \theta_2(0,0)\hat{\pi}_{1,2}(0,0), \;\; \pi_{1,2,3}(0,0,1) = (1-\theta_2(0,0))\hat{\pi}_{1,2}(0,0)$$
$$\pi_{1,2,3}(0,1,0) = \theta_2(0,1)\hat{\pi}_{1,2}(0,1), \;\; \pi_{1,2,3}(0,1,1) = (1-\theta_2(0,1))\hat{\pi}_{1,2}(0,1)$$
$$\pi_{1,2,3}(1,0,0) = \theta_2(1,0)\hat{\pi}_{1,2}(1,0), \;\; \pi_{1,2,3}(1,0,1) = (1-\theta_2(1,0))\hat{\pi}_{1,2}(1,0)$$
$$\pi_{1,2,3}(1,1,0) = \theta_2(1,1)\hat{\pi}_{1,2}(1,1), \;\; \pi_{1,2,3}(1,1,1) = (1-\theta_2(1,1))\hat{\pi}_{1,2}(1,1)$$

Similar to the pairwise margin case, the Hessian matrix for $\nabla^2 l(\theta_2) \leq 0$. Hence, we again use Newton-Raphson method to estimate θ_2. Once we have

$\hat{\pi}_{1,2,3}(x_1, x_2, x_3)$, we estimate $\pi_{2,3,4}(x_2, x_3, x_4)$ similarly by fixing $\hat{\pi}_{2,3}(x_2, x_3) = \sum_i \hat{\pi}_{1,2,3}(x_1, x_2, x_3)$. We repeat this process until, finally, we obtain all the threewise margins $\hat{\pi}_{s,s+1,s+2}(x_1, x_2, x_3)$ for $s = 1, \ldots, L-2$.

One question for the threewise marginal estimates is that how to get the leftmost parameter $\hat{\pi}_{1,2}(x_1, x_2)$. To obtain $\hat{\pi}_{1,2}(x_1, x_2)$, we use the strategy for $m = 2$ again. That is, we first estimate $\hat{\pi}_1(x_1) = n_1(x_1)/n$, then calculate $\pi_{1,2}(x_1, x_2)$ given $\hat{\pi}_1(x_1)$.

Therefore, the general left-right estimator for estimating $(m+1)$-wise margins, $\pi_{s,s+1,\ldots,s+m}(x_s, x_{s+1}, \ldots, x_{s+m})$ is

1. Begin from $\hat{\pi}_1(x_1) = n_1(x_1)/n$.

2. Use left-right estimator for pairwise, threewise,..., until m-wise margins to obtain $\hat{\pi}_{1,2}(x_1, x_2)$, $\hat{\pi}_{1,2,3}(x_1, x_2, x_3)$,..., and $\hat{\pi}_{1,\ldots,m}(x_s, \ldots, x_m)$ sequentially.

3. Estimate $\pi_{s,s+1,\ldots,s+m}(x_s, x_{s+1}, \ldots, x_{s+m})$ by fixing $\hat{\pi}_{s,\ldots,s+m-1}(x_s, \ldots, x_{s+m-1})$ and through reparameterized $\theta_{s+1}(\underline{x})$, where $\underline{x} = (x_1, \ldots, x_{s+m-1})$ and $\theta_{s+1}((\underline{x})) \in [0, 1]$, and

$$\pi_{s,\ldots,s+m}((\underline{x}), 0) = \theta_{s+1}(\underline{x})\hat{\pi}_{s,\ldots,s+m-1}(\underline{x}),$$
$$\pi_{s,\ldots,s+m}((\underline{x}), 1) = [1 - \theta_{s+1}(\underline{x})]\hat{\pi}_{s,\ldots,s+m-1}(\underline{x}).$$

   The estimate of $\theta_{s+1}((\underline{x}))$ is calculated numerically by the Newton-Raphson method.

4.5 Discussion

The idea of the above left-right estimator is to estimated the marginal ancestral distribution given the lower margins sequentially from the left side to the right side of the sequence. A natural set of questions regarding this estimator

would are: Can we can apply the estimator from the right to the left side of the sequence? If we can, would we obtain the same estimates as those obtained in the original direction? If the estimations obtained from two directions are different, what should we do?

For the first question, our the answer is yes. By symmetry, we can apply the same estimator to estimate the marginal ancestral distribution beginning from the right side, It is just a matter of index relabeling.

Are the estimates the same? From the perspective of the MCCL objective function, the estimates obtained either from left or from right should be the same. This is because in equation (4.2), we have

$$\begin{aligned} & \frac{f(x_1,...,x_{m+1})f(x_2,...,x_{m+2})\cdots f(x_{L-m},...,x_L)}{f(x_2,...,x_{m+1})\cdots f(x_{L-m},...,x_{L-1})} \\ = & \frac{f(x_1,...,x_m)}{f(x_1,...,x_m)} \cdot \frac{f(x_1,...,x_{m+1})f(x_2,...,x_{m+2})\cdots f(x_{L-m},...,x_L)}{f(x_2,...,x_{m+1})\cdots f(x_{L-m},...,x_{L-1})} \\ = & f(x_1,...,x_m)f(x_{m+1}|x_m,...,x_1)f(x_{m+2}|x_{m+1},...,x_2)\cdots f(x_L|x_{L-1},...,x_{L-m}). \end{aligned}$$

In addition, we also have

$$\begin{aligned} & \frac{f(x_1,...,x_{m+1})f(x_2,...,x_{m+2})\cdots f(x_{L-m},...,x_L)}{f(x_2,...,x_{m+1})\cdots f(x_{L-m},...,x_{L-1})} \\ = & \frac{f(x_1,...,x_{m+1})f(x_2,...,x_{m+2})\cdots f(x_{L-m},...,x_L)}{f(x_2,...,x_{m+1})\cdots f(x_{L-m},...,x_{L-1})} \cdot \frac{f(x_{L-m+1},...,x_L)}{f(x_{L-m+1},...,x_L)} \\ = & f(x_{L-m+1},...,x_L)f(x_{L-m}|x_{L-m+1},...,x_L)\cdots f(x_1|x_2,...,x_{m+1}). \end{aligned}$$

Thus

$$\begin{aligned} & f(x_1,...,x_m)f(x_{m+1}|x_m,...,x_1)f(x_{m+2}|x_{m+1},...,x_2)\cdots f(x_L|x_{L-1},...,x_{L-m}) \\ = & f(x_{L-m+1},...,x_L)f(x_{L-m}|x_{L-m+1},...,x_L)f(x_{L-m-1}|x_{L-m},...,x_{L-1})\cdots f(x_1|x_2,...,x_{m+1}), \end{aligned}$$

which means there is no left-right effect, because the objective function is the same.

However, in reality, we are doing sequential conditional maximization with a different sequence of conditioning steps. That is, in left-right estimator, the

estimate for the current m-wise margin depends on the left neighbour estimated m-wise margin. When the sequence length L is large, the errors pass to the next step and finally the cumulative errors at the end could make the estimates obtained from two directions rather different. We call this phenomenon the *left-right effect*.

To better understand this effect, we generated $n = 100$ binary sequences with $L = 10$ from a known ancestral distribution via recombination model, using the recombination rate $q = 0.1$. We then used the left-right estimator to estimate the pairwise margins $\pi_{s,s+1}(x_s, x_{s+1})$, beginning from both left and right ends of the sequence. By comparing the estimations from two directions (Figure 4.1), we can see that there is a small but obvious left-right effect for $\hat{\pi}_{s,s+1}(0,0)$ and $\hat{\pi}_{s,s+1}(1,1)$.

To eliminate the left-right effect, we propose to use the *best linear combination* method. Let π^l denote the (marginal) ancestral distribution estimated from left to right, and π^r denote the (marginal) ancestral distribution estimated from right to left. If we form a convex combination of π^l and π^r, we would get a family of π-value, say $\pi^\alpha = (1-\alpha)\pi^l + \alpha\pi^r$, where $\alpha \in [0,1]$. To reduce the left-right effect, we would select the value of α that π^α maximizes the log MCCL among all $\alpha \in [0,1]$.

For example, for order-1 MC model, the log likelihood function for π^α is

$$l(\alpha) = \sum_{x_1} \cdots \sum_{x_L} n(x_1 \cdots x_L) \log \frac{f^\alpha(x_1,x_2) f^\alpha(x_2,x_3) \cdots f^\alpha(x_{L-1},x_L)}{f^\alpha(x_2) f^\alpha(x_3) \cdots f^\alpha(x_{L-1})},$$

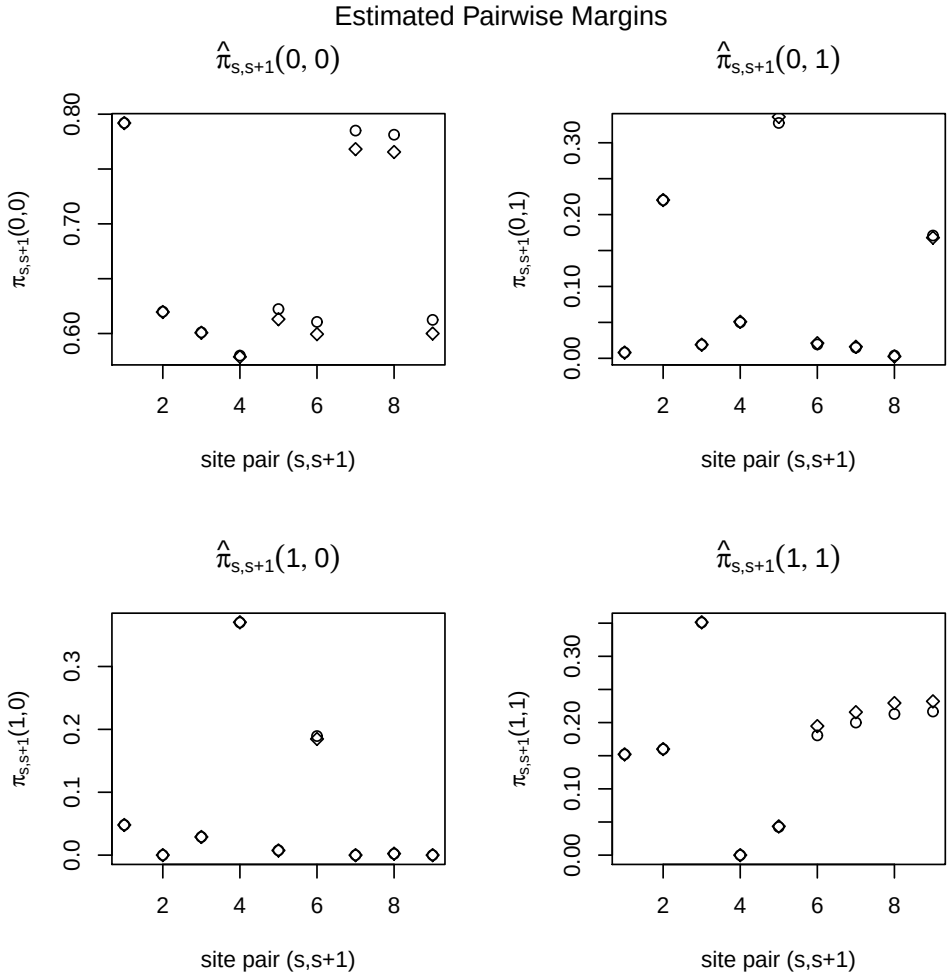


Figure 4.1: Left-Right Effect. Diamonds correspond to right-to-left estimates, circles are left-to-right. Here $n = 100$ and $L = 10$.

where $x = (x_1, x_2, \ldots, x_L)$ is the observed binary sequence with $x_i \in \{0, 1\}$, and

$$\begin{aligned} f^{\alpha}(x_s) &= \pi_s^{\alpha}(x_s) \\ &= (1-\alpha)\pi_s^{l}(x_s) + \alpha\pi_s^{r}(x_s) \\ f^{\alpha}(x_s, x_{s+1}) &= (1-q)\pi_{s,s+1}^{\alpha}(x_s, x_{s+1}) + q\pi_s^{\alpha}(x_s)\pi_{s+1}^{\alpha}(x_{s+1}) \\ &= (1-q)\left[(1-\alpha)\pi_{s,s+1}^{l}(x_s, x_{s+1}) + \alpha\pi_{s,s+1}^{r}(x_s, x_{s+1})\right] \\ &\quad +q\left[(1-\alpha)\pi_s^{l}(x_s) + \alpha\pi_s^{r}(x_s)\right]\left[(1-\alpha)\pi_{s+1}^{l}(x_{s+1}) + \alpha\pi_{s+1}^{r}(x_{s+1})\right] \end{aligned}$$

To get α that maximizes the above log likelihood function, $l(\alpha)$, we can use Newton-Raphson or a bisection algorithm. For example, using the above example, the α value which maximizes $l(\alpha)$ is 0.21754 if we use Newton-Raphson method, equivalent to the 0.21755 found using the bisection method.

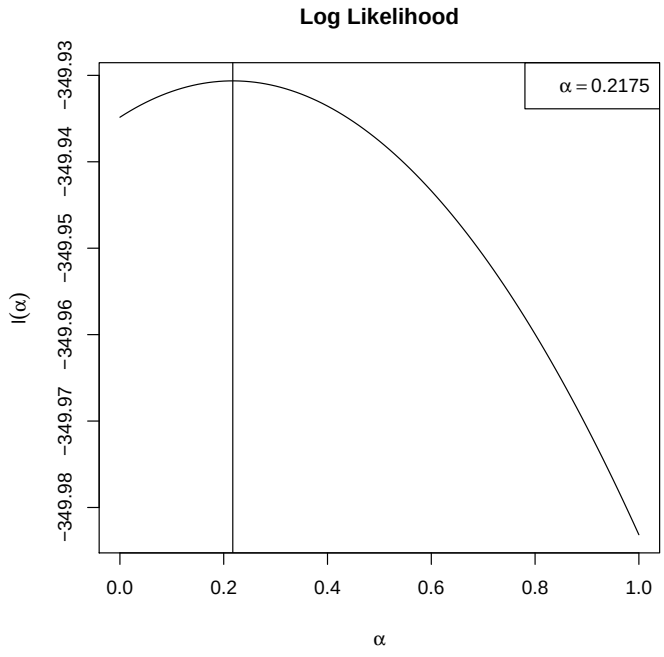


Figure 4.2: Best Linear Combination

In Figure 4.2, we can see the log likelihood for π^l is larger than that for π^r and the best linear combination does put more weight on π^l. However, the

difference between $l(\pi^l)$ and $l(\pi^r)$ is very small, less than 0.1, numerically.

4.6 A simulation study

In this section, we will examine the performance of the left-right estimator by some simulations.

4.6.1 Simulation design

Generate true ancestral distribution

To generate the simulated data, we first generated the ancestral distribution from a real data set. The real data set we used is a haplotype sequence data, called 'HapMap3 r2', released by *International HapMap Project* in Feb., 2009. (http://hapmap.ncbi.nlm.nih.gov/) Particularly, we used the haplotypes for the population of Yoruba in Ibadan, Nigeria (YRI), one of the 11 populations in HapMap phase 3. In the released YRI data, there are 167 individuals and 230 haplotypes for each of 22 chromosomes. To simplify the simulation, we just used a portion of the YRI data as the ancestral sequences. That is, we used 60 unrelated individuals and 120 haplotypes for 50 sites on one chromosome (Mao (2010)).

We generated ancestral distribution with three different sequence length, $L = 10, 15, 20$. To calculate the ancestral distribution with sequence length as L, we counted the number of distinct sequences among 120 haplotypes in the first L out of 50 sites, and then calculated the corresponding relative frequency. We used these distinct sequences as the ancestor sequences and the corresponding relative frequency as the probability mass for ancestor sequences.

Generate simulated data from a known ancestral distribution

We generated the simulated data from each of the three ancestral distributions obtained above by considering recombination through our recombination

model. The recombination rates we used are $q = 0.01, 0.05, 0.1$. Thus, we have in total nine simulated data sets, one for each L and q combination.

We generate $n*N$ binary sequences for each simulated data set as the observations, where $n = 100$ is the total number of observations within each replication and $N = 100$ is the number of replications.

Simulation results to report

In each simulation study, we reported the estimated m-wise marginal ancestral distribution, where $m = 2, 3, 4$, by applying the left to right estimator beginning from both left and right ends of the sequence. We also calculated the best linear combination π^{α} based on π^{l} and π^{r}, and then compared our estimations with the true m-wise margins. In addition, the reconstructed joint ancestral distributions could be calculated from marginal π^{l} and π^{r} by assuming an order-$(m-1)$ Markov property. Together with joint π^{α}, we compared these three estimated joint ancestral distribution with the true one. Finally, the computing time were also included in the report to check if our estimator is computationally feasible.

Note that, when we reconstruct the joint ancestral distribution through the estimated margins by using Markov property (4.3), theoretically, we need to calculate $\pi_{x_1,\ldots,x_L}$ for all the possible combinations of $x_1, \ldots, x_L$. That is, we need to find all 2^L probabilities, though in practice most of them would be zeroes. This computation is very time consuming when L is large. Thus, we need to figure out a feasible method for the reconstruction.

In the following simulation studies, instead of calculating the probabilities for all potential 2^L ancestors, we only calculate the probabilities for a set of *simulated ancestors*. To reconstruct the joint ancestral distribution given the estimated m and $(m+1)$-wise margins, we first simulated $B = 1000$ binary sequences with length L, by using order-m property, as a pool for the ancestors. For example,

for $m = 1$, we generate the simulated ancestral sequence as follows: generate $x_1 \sim Bernoulli(\hat{\pi}_1(1))$; given x_1 fixed, we generate $x_2 \sim Bernoulli(\hat{\pi}_{2|1}(1|x_1))$; fixing x_2, we generate $x_3 \sim Bernoulli(\hat{\pi}_{2|3}(1|x_2))$; $\cdots$; following the process, until given x_{L-1} fixed, we generate $x_L \sim Bernoulli(\hat{\pi}_{L|L-1}(1|x_{L-1}))$. The sequence $x = (x_1, x_2, \ldots, x_L)$ is just one simulated ancestor. We then reconstruct the joint distribution only for the distinct sequences in the simulated ancestors pool.

In reality, since the probability masses for the most of the 2^L possible ancestral sequences are zeroes, a lot of our estimated margins should be zeroes. Thus, those ancestral sequences with zero probability mass are very unlikely to be included in the simulated ancestors pool. Hence, the number of distinct sequences in the simulated ancestors pool is limited and much less then 2^L when L is large.

It is possible that we miss some ancestors with very small probabilities by using the simulated ancestors pool method. However, this method can identify those ancestors with larger population proportions. In addition, this method reduces the computation complexity greatly and makes it feasible to reconstruct the joint ancestral distribution for long sequence.

4.6.2 Simulation results

Notations

In order to illustrate the simulation results clearly, we first clarify some notations that will be used in the simulation reports. Suppose $\underline{x} = (x_1, x_2, \ldots, x_L)$ is a potential ancestor sequence, we denote the true ancestral distribution as $\tau(\underline{x})$. The estimated ancestral distribution, which is reconstructed by applying Markov chain property, will be denoted as $\hat{\pi}(\underline{x})$. The empirical distribution for simulated ancestors, as discussed in section 4.6.1, will be denoted as $\hat{\pi}_{sim}$. In addition, we let $\tau^m_{MC}(\underline{x})$ denote an order-m Markov chain approximation to the true ancestral

distribution. That is,

$$\tau^m_{MC}(\underline{x}) = \tau(x_1, \ldots, x_m)\tau(x_{m+1}|x_m, \ldots, x_1) \cdots \tau(x_L|x_{L-1}, \ldots, x_{L-m}).$$

Estimated marginal distributions

First, we examine the performance of the left to right estimator to estimate the pairwise, threewise, and four wise marginal ancestral distributions. From the simulation results, we find that the left-right effects are not obvious for most situation. For example, from Figure 4.3, the estimated pairwise margins, either from left ends, right ends, or the best linear combination $\hat{\pi}^\alpha = (1-\alpha)\hat{\pi}^l + \alpha\pi^l$, are quite close. Hence, in the following we only report the simulation results for the best linear combination estimates.

Figure 4.4 is the comparison between estimated pairwise and true margins. Here $L = 10$ and $q = 0.01$. In the plot, squares correspond to true margins and circles are the average estimated margins. Stars above and below the circles represent the average estimated margins plus or minus two times of standard error. That is, the interval between two stars at same x-axis value is about 95% confidence interval for the estimates.

From the figure we can see that our average estimated margins are very closed to the true margins with small standard errors. In addition, the standard errors for the estimates with small true values are smaller than those for the estimates larger true values. Similar results can be found for estimated threewise and fourwise margins for this data set, and for other data sets. See Figures B.1 to B.6 in the Appendix B for details.

We also considered how the recombination factor would affect our inferences. Figure 4.5 and Figure 4.6 show the estimated pairwise margin when $q = 0.01, 0.05, 0.1$ under the same sequence length $L = 10$. From the plots we can see that the recombination rate did not have a large effect. A similar conclusion can

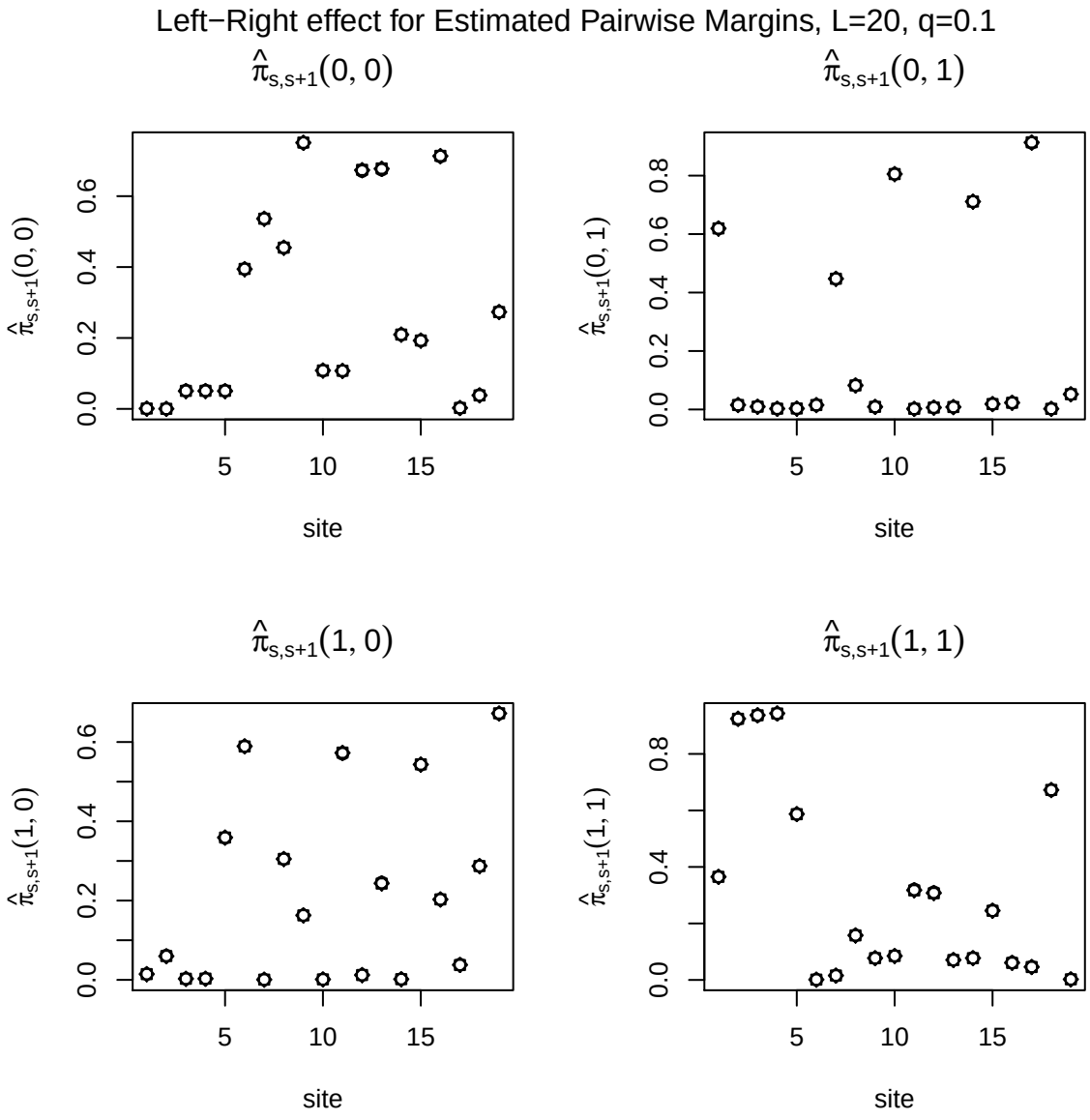


Figure 4.3: Estimated Pairwise Margins. Here $L = 20$, $q = 0.1$. Circles represent $\hat{\pi}^l$, squares correspond to $\hat{\pi}^r$, and diamonds are $\hat{\pi}^\alpha$. Note that circles, squares and diamonds are almost overlapped.

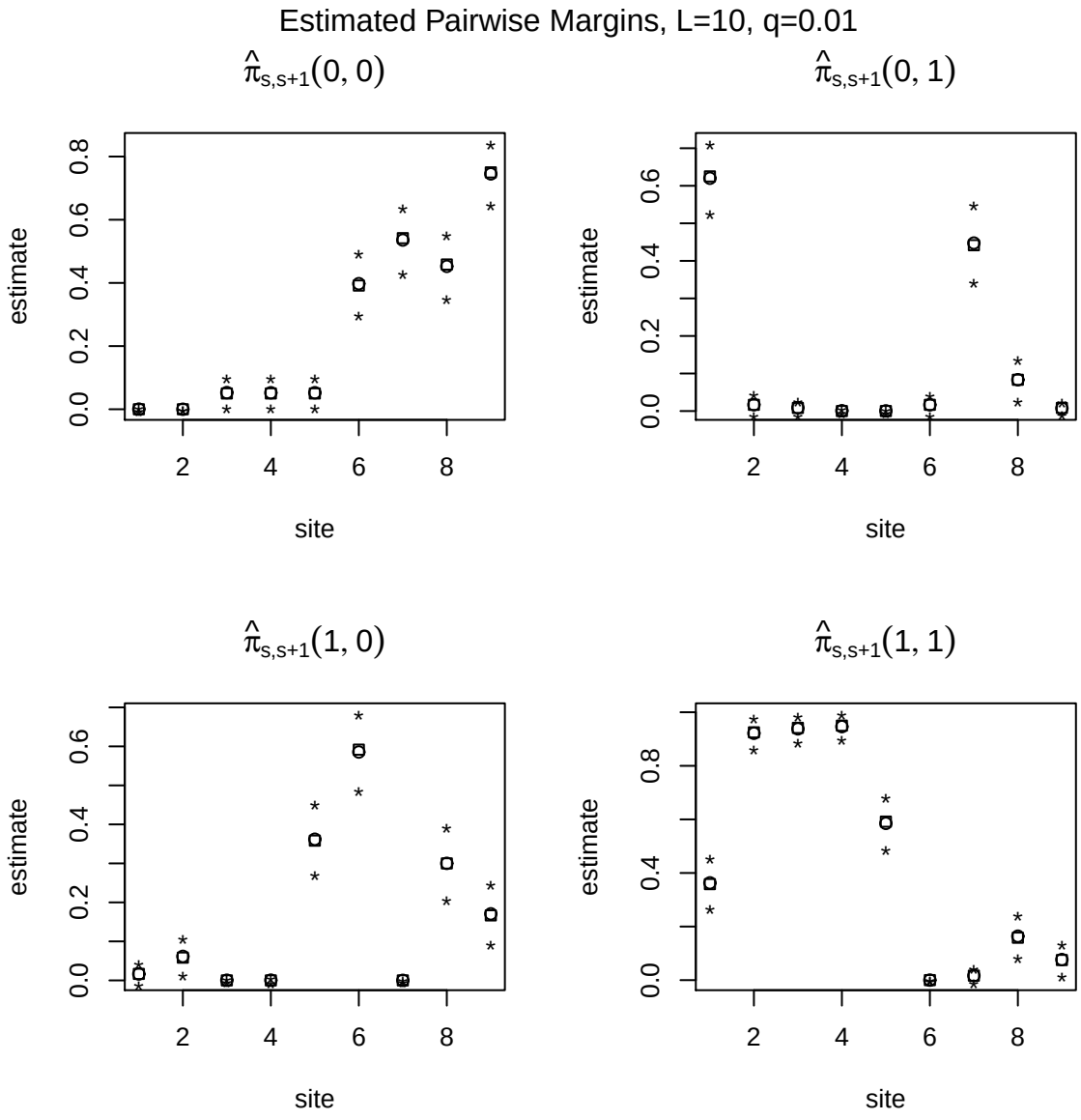


Figure 4.4: Estimated Pairwise Margins. Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

be drawn for higher order margins and other data sets, see Figures B.7 to B.14 in the Appendix B for details.

Estimated joint distributions

After checking the performance of left to right estimator for estimated marginal ancestral distribution, we compared the MC reconstruction of the joint ancestral distribution based on $\hat{\pi}^{\alpha}(\underline{x})$, to the true joint ancestral distribution. Please note that the full joint distribution $\tau(\underline{x})$ is not identifiable from the order m margins, so that a considerable portion of the errors in this methodology are due to the Markov chain approximation being used.

We first examine if the estimated joint ancestral distribution can identify all the true ancestors, i.e. ancestors with positive population proportion. In data set 1, where $L = 10$ and $q = 0.01$, there are 13 true ancestors with positive population proportion, denoted by $\mu^i, i = 1, \ldots, 13$. Table 4.1 shows the comparison between true and the average estimated probability for all 13 true ancestors when applying the order-1, 2, and 3 MC to reconstruct the joint distribution.

From the simulation results, we can see that our reconstruction method has considerable bias for the probabilities of the ancestral distribution. For example, the sum of estimated probabilities when applying order-1 MC reconstruction at the true ancestors is 0.4715. So around half of probabilities are assigned to non-ancestral sequences. In addition, we can see that the higher order MC reconstruction is better than the lower one for estimating the true ancestral distribution. Particularly, in this case the sum of estimated probabilities at the true ancestors are 0.4715 for order-1 MC reconstruction, 0.7844 for order-2 MC reconstruction, and 0.8254 for order-3 MC reconstruction.

Figure 4.7 shows the above comparison between true and estimated joint ancestral distribution. The solid line is the bar plot for true probability at the

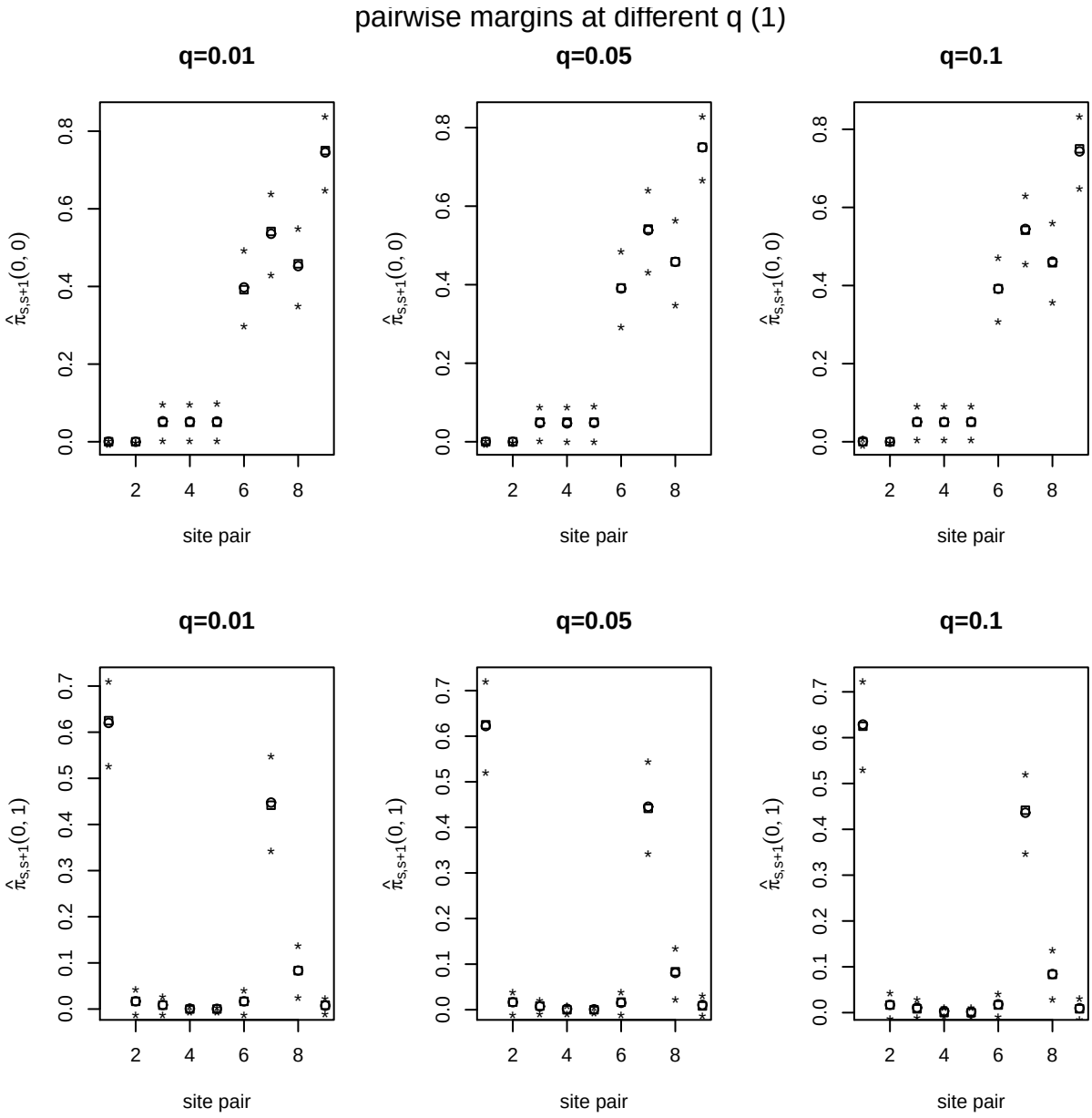


Figure 4.5: Estimated Pairwise Margins at Different q. Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

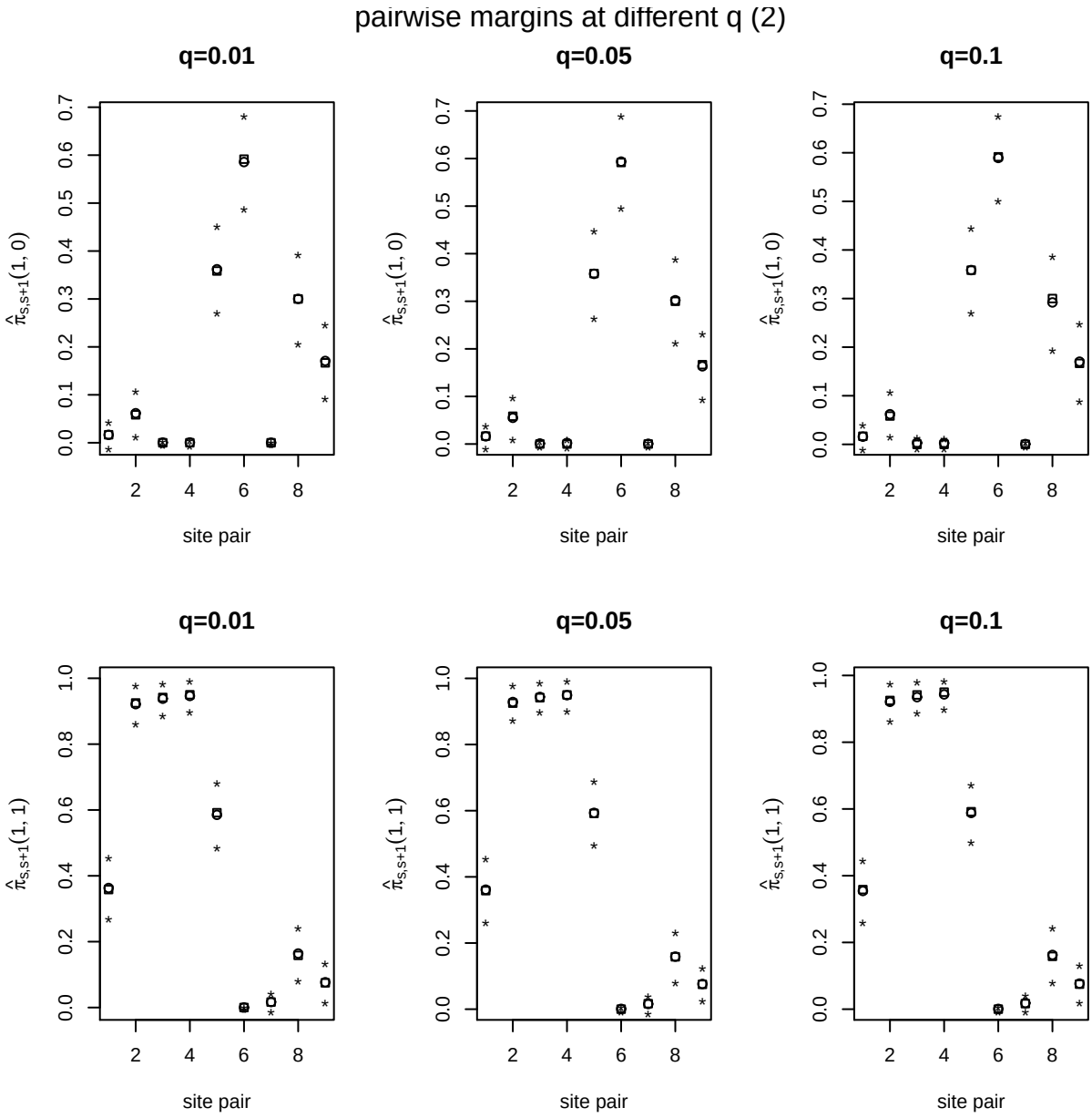


Figure 4.6: Estimated Pairwise Margins at Different q. Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

Table 4.1: Comparison between true, $\tau(\mu)$, and estimated, $\hat{\pi}(\mu)$, probability at true ancestor μ when $L = 10$ and $q = 0.01$

ancestors	μ^1	μ^2	μ^3	μ^4	μ^5	μ^6	μ^7	μ^8	μ^9	μ^{10}	μ^{11}	μ^{12}	μ^{13}	
	true ancestral distribution													total
$\tau(\mu)$	0.4250	0.2917	0.0750	0.0500	0.0417	0.0250	0.0250	0.0167	0.0167	0.0083	0.0083	0.0083	0.0083	1
	order-1 MC reconstruction													
$\hat{\pi}(\mu)$	0.1658	0.0382	0.0218	0.0017	0.0137	0.0395	0.0181	0.0000	0.0950	0.0015	0.0622	0.0016	0.0124	0.4715
sd	0.0378	0.0160	0.0100	0.0017	0.0041	0.0106	0.0081	0.0001	0.0138	0.0018	0.0143	0.0022	0.0051	
	order-2 MC reconstruction													
$\hat{\pi}(\mu)$	0.2624	0.0904	0.0493	0.0081	0.0226	0.0047	0.0043	0.0002	0.1684	0.0036	0.1350	0.0041	0.0314	0.7844
sd	0.0444	0.0289	0.0177	0.0072	0.0074	0.0037	0.0035	0.0004	0.0205	0.0045	0.0218	0.0052	0.0102	
	order-3 MC reconstruction													
$\hat{\pi}(\mu)$	0.2602	0.1018	0.0492	0.0073	0.0154	0.0144	0.0153	0.0002	0.1669	0.0066	0.1526	0.0043	0.0313	0.8254
sd	0.0446	0.0302	0.0177	0.0061	0.0064	0.0075	0.0081	0.0005	0.0203	0.0071	0.0223	0.0051	0.0101	

true ancestors, given in a decreasing magnitude left to right. The circles show our average estimation, $\hat{\pi}$, at the true ancestors. Stars above and below the circles correspond to $\hat{\pi} \pm 2s$, where s is the standard errors. We can draw the same conclusion from this figure as from Table 4.1. Therefore, in the following we will mainly compare the true ancestral distribution and our estimates reconstructed by using order-3 MC property.

As we noted, one reason we expected our estimators to have considerable bias is that, in practice, the true ancestral distribution is not a Markov chain. To reduce this, we also compared our estimates with the Markov chain reconstruction of the true ancestral distribution. That is we compare $\hat{\pi}(\mu_{MC})$ with $\tau^m_{MC}(\mu_{MC})$ for $m = 1, 2, 3$, where μ_{MC} is the binary sequence such that $\tau^m_{MC}(\mu_{MC}) > 0$.

Figure 4.8 shows this comparison based on simulation results in data set 1. The solid line represents the true probability, the dashed line is $\hat{\pi}$, and the interval between two dotted line is $\hat{\pi} \pm 2s$. We can see that our estimates are very good when compared with the MC version of the true ancestral distribution, τ^m_{MC}. Tables 4.2 is the sum of estimated probabilities at μ_{MC} when $m = 1, 2, 3$, which also confirms the above conclusion.

Table 4.2: Sum of estimated probabilities, $\hat{\pi}(\mu_{MC})$, at the binary sequence, μ_{MC}. Here $L = 10$ and $q = 0.01$.

	$m = 1$	$m = 2$	$m = 3$
$\sum_{\mu_{MC}} \hat{\pi}(\mu_{MC})$	0.9853	0.9845	0.9874

We next consider the role of L and q in the estimators behavior. Figure 4.9 compares the estimated ancestral distribution, $\hat{\pi}$ by using order-3 MC reconstruction, with the true ancestral distribution, τ, for all combinations of $L = 10, 20$ and $q = 0.01, 0.1$. From the figure we can see as the sequence length increases, the bias increases. In addition, Table 4.3, which lists the sum of the estimated probabilities

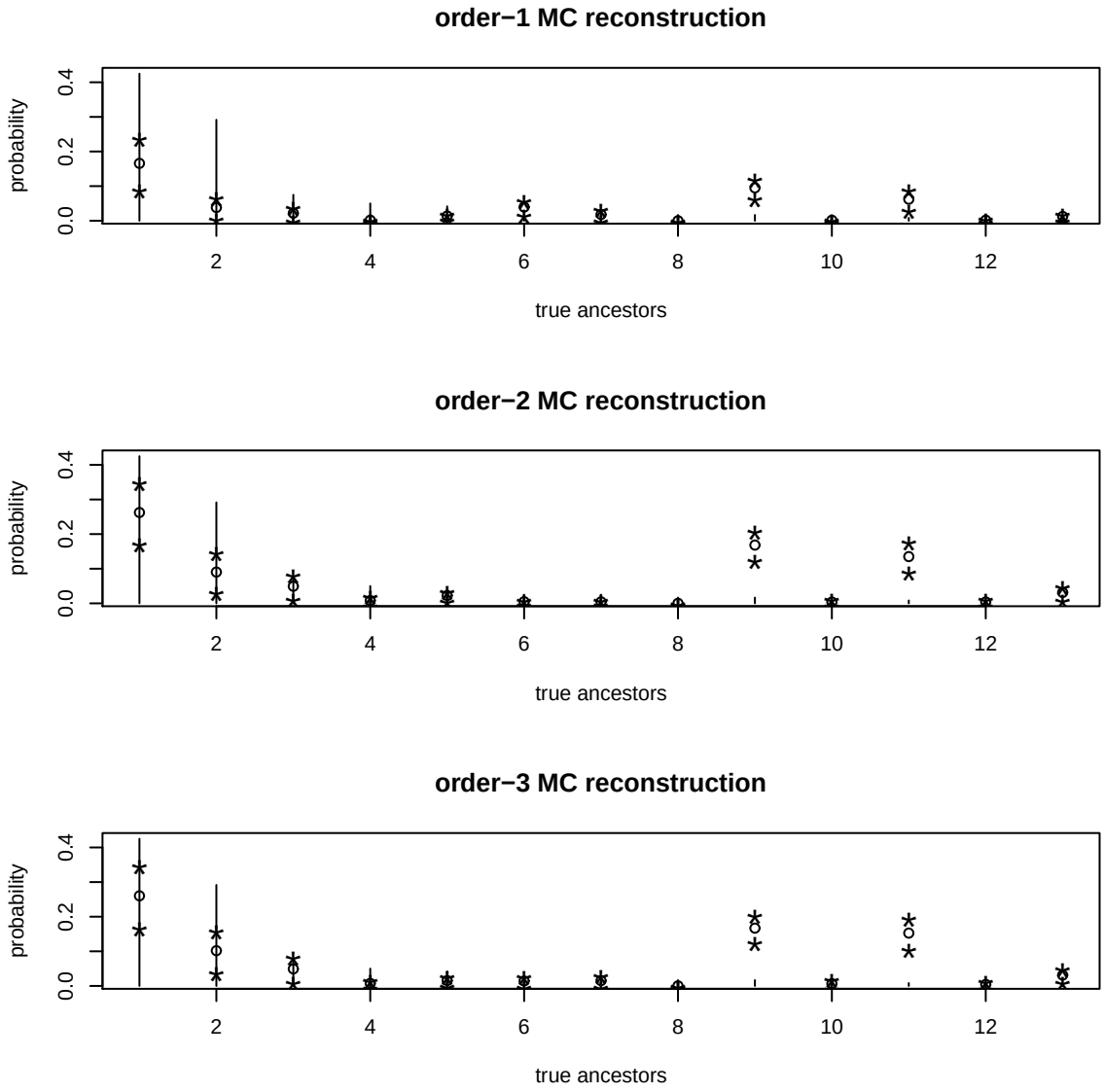


Figure 4.7: Estimated Ancestral Distribution. Here $L = 10$, $q = 0.01$ and order-1,2,3 MC are applied. Solid vertical line is the true probability at the true ancestors, drawn by a decreasing order. Circles are average $\hat{\pi}$ at the true ancestors. Stars above and below the circles correspond to $\hat{\pi} \pm 2s$.

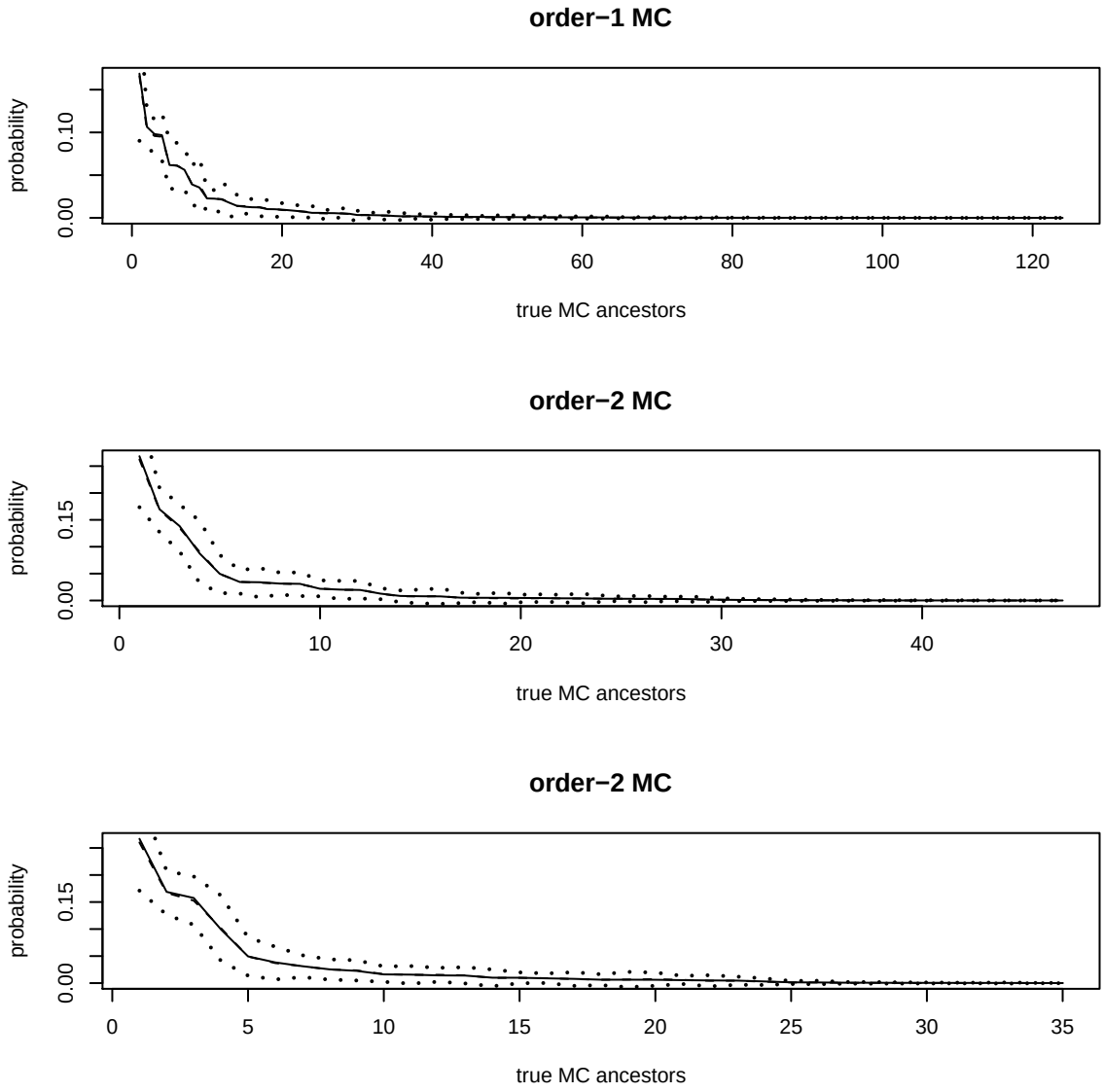


Figure 4.8: Estimated Ancestral Distribution. Here $L = 10$, $q = 0.01$ and order-1,2,3 MC are applied. Solid line is the true probability at the true ancestors, drawn by a decreasing order. Dashed line is $\hat{\pi}$ at the true ancestors. Dotted lines above and below the dashed line correspond to $\hat{\pi} \pm 2s$.

at the true ancestors, μ, also demonstrates the same conclusion. Moreover, from the table we can also find when q increases, the sum of probability decreases a little bit, which also implies that the bias in estimating the zero probability cases increases. Note that when L is different, the true ancestral distribution τ and the number of true ancestors are also different.

Table 4.3: Sum of estimated probabilities $\hat{\pi}(\mu)$ at the true ancestors μ.

	$q = 0.01$	$q = 0.05$	$q = 0.1$
$L = 10$	0.8258	0.8187	0.7966
$L = 15$	0.5540	0.5381	0.5267
$L = 20$	0.3455	0.3398	0.3293

One reason that more probabilities are assigned to non-ancestral sequences as L increases could be the diffusion of probabilities to the neighbors of true ancestors. Thus, we also summarized the sum of probabilities at the true ancestors and their neighbors. Here, the neighbor sequences are selected by Hamming distance, d_h, which is defined as the number of positions where two strings a and b have different elements. For example, all the sequences having Hamming distance 1 with the true ancestor μ are the sequences that only have one different element compared with μ, or, by switching only one symbol of μ at each time. In Table 4.4, we summarize the results when considering neighbor sequences with $d_h = 1$.

Computation time

In the following table, Table 4.5, we report the computation time (in seconds) for the marginal estimation by using R language on computational cluster Lion-XO in $N = 100$ replications. The time reported in the table is the average of computation time for π^l and π^r, because from the estimator point of view, there is no difference between starting from left or right. The best linear combination

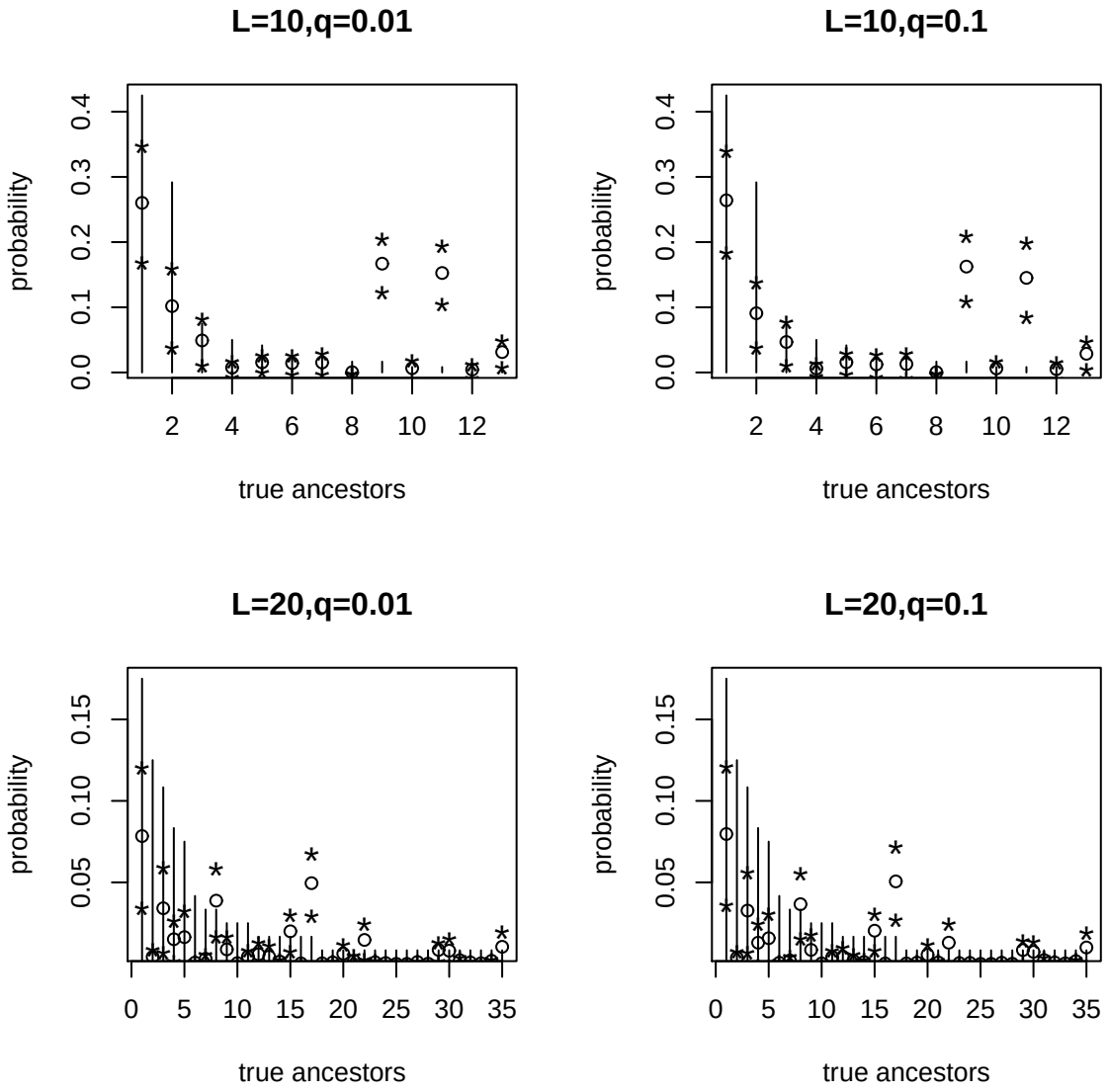


Figure 4.9: Estimated Ancestral Distribution for $L = 10, 20$ $q = 0.01, 0.1$, and order-3 MC are applied. Solid vertical line is the true probability at the true ancestors, drawn by a decreasing order. Circles are average $\hat{\pi}$ at the true ancestors. Stars above and below the circles correspond to $\hat{\pi} \pm 2s$.

Table 4.4: Sum of estimated probabilities $\hat{\pi}$ at the true ancestors μ and their neighbor sequences with Hamming distance as 1.

	$q = 0.01$	$q = 0.05$	$q = 0.1$
$L = 10$	0.9485	0.9499	0.9438
$L = 15$	0.8501	0.8409	0.8344
$L = 20$	0.6835	0.6726	0.6563

estimates are constructed from estimated margins from both left and right ends of the sequence, so we didn't report the computation time for π^{α}. From the table we can see that the computation is fast. As the order of margins increased, the computation time increased. Note that to estimate the m-wise marginal ancestral distribution, we have 2^{m-1} parameters to be estimated at each step in the left to right estimator.

Table 4.5: Computation time (in seconds) for estimating margins in 100 samples using left to right estimate method.

	time			log_2 time		
	pairwise	threewise	fourwise	pairwise	threewise	fourwise
$L = 10, q = 0.01$	1.76	6.25	20.64	0.82	2.64	4.37
$q = 0.05$	1.90	7.22	24.36	0.93	2.85	4.61
$q = 0.1$	2.05	7.06	20.75	1.04	2.82	4.38
$L = 15, q = 0.01$	2.84	9.71	26.83	1.51	3.28	4.75
$q = 0.05$	3.10	10.25	33.26	1.63	3.36	5.06
$q = 0.1$	2.92	11.49	33.94	1.55	3.52	5.08
$L = 20, q = 0.01$	4.32	13.26	72.92	2.11	3.73	6.19
$q = 0.05$	4.08	14.75	87.42	2.03	3.88	6.45
$q = 0.1$	4.59	14.99	94.77	2.20	3.91	6.57

Chapter 5

Hierarchical Estimator

In chapter 4, we discussed applying MCCL to the recombination model, and proposed a left to right estimator to estimate the $(m+1)$-wise marginal ancestral distribution, $\pi_{s,\ldots,s+m}(x_s,\ldots,x_{s+m})$, given the left neighbor, $\pi_{s,\ldots,s+m}(x_s,\ldots,x_{s+m-1})$, is fixed. In order to estimate the $(m+1)$-wise margins, we changed the parameters $\pi_{s,\ldots,s+m-1}$'s into θ_{s+1}'s. However, this made the margins an increasingly complicated function of the previous $\theta_2,\ldots,\theta_s$. Hence, the errors coming from previous step would be passed into current step, which could cause a left-right effect at the end of the sequences as discussed in section 4.4.1 and section 4.5. Moreover, these θ parameters are nonseparable in the MCCL objective function, causing complexity in finding the explicit estimators, so we have to use a multivariate optimization method to do the inference. Thus computational expense will rise with the order m. In this chapter, we present a new parametrization and a new estimator which has the potential to improve upon the left to right estimator.

5.1 A reparametrization

The new estimator we are going to discuss in this chapter requires one to change the unknown parameters $\pi_{s,\ldots,s+m}(x_s,\ldots,x_{s+m-1})$ into a new parameter system. We will illustrate how to do the reparametrization in this section.

5.1.1 Reparametrization for pairwise margins

To demonstrate the reparametrization clearly, let's first consider the simple situation: applying the order-1 MCCL to the recombination model.

We start by estimating $\hat{\pi}_s(0)$ for all s using the order-0 Markov chain. This means $\hat{\pi}_s(0) = n_s(0)/n$. To estimate the next higher margins $\pi_{s,s+1}(x_s, x_{s+1})$, we use the order-1 MCCL holding the $\hat{\pi}_s(0)$ parameters fixed. When applying order-1 MCCL, the log likelihood function equals equation (4.4). To estimate the pairwise marginal distributions $\pi_{s,s+1}(x_s, x_{s+1})$, for $s = 1, \ldots, L-1$ and $x_s, x_{s+1} \in \{0, 1\}$, we again apply conditional maximization fixing all the onewise margins as $\hat{\pi}_s(x_s)$ for $s = 1, ..., L$. Given all the onewise margins fixed, the unknown parameter $\pi_{s,s+1}(x_s, x_{s+1})$ only appears in the MCCL in the term $f(x_s, x_{s+1})$. It does not depend on other neighbor pairs $f(x_t, x_{t+1})$ or $f(x_s)$ in the log MCCL (4.4). Thus, to estimate $\pi_{s,s+1}(x_s, x_{s+1})$, it can maximize the following reduced log likelihood function

$$l(\pi_{s,s+1}(x_s, x_{s+1})) = \sum_{x_s, x_{s+1}} n_{s,s+1}(x_s, x_{s+1}) \log f(x_s, x_{s+1}),$$

where

$$\begin{aligned} f(x_s, x_{s+1}) &= P(X_s = x_s, X_{s+1} = x_{s+1}) \\ &= (1-q)\pi_{s,s+1}(x_s, x_{s+1}) + q\pi_s(x_s)\pi_{s+1}(x_{s+1}). \end{aligned}$$

In addition, since $\pi_{s,s+1}(x_s, x_{s+1})$ must satisfy the lower margin consistency property, we have the constraints

$$\begin{aligned} \pi_s(x_s) &= \sum_{x_{s+1}} \pi_{s,s+1}(x_s, x_{s+1}) \\ \pi_{s+1}(x_{s+1}) &= \sum_{x_s} \pi_{s,s+1}(x_s, x_{s+1}) \\ 1 &= \sum_{x_{s+1}} \sum_{x_s} \pi_{s,s+1}(x_s, x_{s+1}), \end{aligned}$$

and so only one free parameter remains, say $\pi_{s,s+1}(0,0)$. If we let $\pi_{s,s+1}(0,0) = \phi_s$, then the other $\pi_{s,s+1}(x_s, x_{s+1})$'s can be expressed as

$$\begin{aligned}
\pi_{s,s+1}(0,1) &= \hat{\pi}_s(0) - \phi_s \\
\pi_{s,s+1}(1,0) &= \hat{\pi}_{s+1}(0) - \phi_s \\
\pi_{s,s+1}(1,1) &= 1 - \hat{\pi}_s(0) - \hat{\pi}_{s+1}(0) + \phi_s.
\end{aligned}$$

In addition, since all $\pi_{s,s+1}(x_s, x_{s+1})$ should satisfy the constraint: $0 \leq \pi_{s,s+1}(x_s, x_{s+1}) \leq 1$, the parameter space for the free parameter ϕ_s is

$$\max\{0, \hat{\pi}_s(0) + \hat{\pi}_{s+1}(0) - 1\} \leq \phi_s \leq \min\{\hat{\pi}_s(0), \hat{\pi}_{s+1}(0)\}.$$

In summary, given all the onewise margins fixed as $\hat{\pi}_s(x_s)$, to estimate the pairwise margins, the unknown parameters are ϕ_s for $s = 1, \ldots, L-1$. They can be estimated by maximizing the individual log likelihood

$$\begin{aligned}
l(\phi_s) &= \sum_{x_s} \sum_{x_{s+1}} n_{s,s+1} \log f(x_s, x_{s+1}) \qquad (5.1) \\
&= n_{s,s+1}(0,0) \log\{(1-q)\phi_s + q\hat{\pi}_s(0)\hat{\pi}_{s+1}(0)\} \\
&+ n_{s,s+1}(0,1) \log\{(1-q)[\hat{\pi}_s(0) - \phi_s] + q\hat{\pi}_s(0)\hat{\pi}_{s+1}(1)\} \\
&+ n_{s,s+1}(1,0) \log\{(1-q)[\hat{\pi}_{s+1}(0) - \phi_s] + q\hat{\pi}_s(1)\hat{\pi}_{s+1}(0)\} \\
&+ n_{s,s+1}(1,1) \log\{(1-q)[1 - \hat{\pi}_s(0) - \hat{\pi}_{s+1}(0) + \phi_s] + q\hat{\pi}_s(1)\hat{\pi}_{s+1}(1)\},
\end{aligned}$$

subject to $\max\{0, \hat{\pi}_s(0) + \hat{\pi}_{s+1}(0) - 1\} \leq \phi_s \leq \min\{\hat{\pi}_s(0), \hat{\pi}_{s+1}(0)\}$. The resulting estimates $\phi_1, \ldots, \phi_{L-1}$ conditionally maximize the order-1 MCCL conditioned on the order-0 MCCL estimates.

5.1.2 Reparametrization for threewise margins

Similar to pairwise margins, we can estimate the threewise marginal ancestral distribution $\pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2})$, for $s = 1, ..., L-2$ and $x_s, x_{s+1}, x_{s+2} \in \{0, 1\}$, by fixing the onewise margins as $\hat{\pi}_s(x_s)$ and the adjacent pairwise margins

as $\hat{\pi}_{s,s+1}(x_s, x_{s+1})$ respectively. Because the $\pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2})$ also need to satisfy the six lower margin consistency constraints,

$$\begin{aligned}
\pi_{s,s+1}(x_s, x_{s+1}) &= \sum_{x_{s+2}} \pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2}) \\
\pi_{s+1,s+2}(x_{s+1}, x_{s+2}) &= \sum_{x_s} \pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2}) \\
\pi_s(x_s) &= \sum_{x_{s+1}} \sum_{x_{s+2}} \pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2}) \\
\pi_{s+1}(x_{s+1}) &= \sum_{x_s} \sum_{x_{s+2}} \pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2}) \\
\pi_{s+2}(x_{s+2}) &= \sum_{x_s} \sum_{x_{s+1}} \pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2}) \\
1 &= \sum_{x_s} \sum_{x_{s+1}} \sum_{x_{s+2}} \pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2}),
\end{aligned}$$

only two of the eight $\pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2})$'s are free. Notice also that we did not constrain the pairwise margin $\pi_{s,s+2}(x_s, x_{s+2})$ because it was not estimated in the order-1 MCCL. Thus we have two free parameters remaining. All other six unknown parameters can be reparameterized as a function of these two free parameters.

To select these two free parameters, we first partition all eight unknown parameters into two groups with four parameters in each, according to whether the value for site $s+1$ is 0 or 1. That is, $\pi_{s,s+1,s+2}(0,0,0)$, $\pi_{s,s+1,s+2}(0,0,1)$, $\pi_{s,s+1,s+2}(1,0,0)$, and $\pi_{s,s+1,s+2}(1,0,1)$ are in one group, and $\pi_{s,s+1,s+2}(0,1,0)$, $\pi_{s,s+1,s+2}(0,1,1)$, $\pi_{s,s+1,s+2}(1,1,0)$, and $\pi_{s,s+1,s+2}(1,1,1)$ are in another group. Then, arbitrarily select one parameter from each group. The selected two parameters are the free parameters we will use. For easier generalization, we will use $\pi_{s,s+1,s+2}(0,0,0)$ and $\pi_{s,s+1,s+2}(0,1,0)$ as our free parameters, and denote them as

$$\pi_{s,s+1,s+2}(0,0,0) \triangleq \phi_s(0), \pi_{s,s+1,s+2}(0,1,0) \triangleq \phi_s(1).$$

The other six parameters can be reconstructed from fixed terms and the

ϕ's as

$$\begin{cases} \pi_{s,s+1,s+2}(0,0,1) &= \hat{\pi}_{1,2}(0,0) - \phi_s(0) \\ \pi_{s,s+1,s+2}(1,0,0) &= \hat{\pi}_{2,3}(0,0) - \phi_s(0) \\ \pi_{s,s+1,s+2}(1,0,1) &= \hat{\pi}_2(0) - \hat{\pi}_{1,2}(0,0) - \hat{\pi}_{2,3}(0,0) + \phi_s(0) \end{cases}$$

$$\begin{cases} \pi_{s,s+1,s+2}(0,1,1) &= \hat{\pi}_{1,2}(0,1) - \phi_s(1) \\ \pi_{s,s+1,s+2}(1,1,0) &= \hat{\pi}_{2,3}(1,0) - \phi_s(1) \\ \pi_{s,s+1,s+2}(1,1,1) &= \hat{\pi}_2(1) - \hat{\pi}_{1,2}(0,1) - \hat{\pi}_{2,3}(1,0) + \phi_s(1) \end{cases}$$

In order to satisfy the constraint, $0 \leq \pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2}) \leq 1$ for $x_s, x_{s+1}, x_{s+2} \in \{0,1\}$, the ranges for the above two free parameters are

$$\begin{aligned} \phi_s(0) &\in [\max\{0, \hat{\pi}_{s,s+1}(0,0) + \hat{\pi}_{s+1,s+2}(0,0) - \hat{\pi}_{s+1}(0)\}, \min\{\hat{\pi}_{s,s+1}(0,0), \hat{\pi}_{s+1,s+2}(0,0)\}] \\ \phi_s(1) &\in [\max\{0, \hat{\pi}_{s,s+1}(0,1) + \hat{\pi}_{s+1,s+2}(1,0) - \hat{\pi}_{s+1}(1)\}, \min\{\hat{\pi}_{s,s+1}(0,1), \hat{\pi}_{s+1,s+2}(1,0)\}]. \end{aligned}$$

From above reparametrization, we can see that unlike the θ parameters in the left to right estimator, the ϕ parameters used here are separable in the log likelihood function, which greatly simplifies algorithms that maximize the likelihood.

To estimate two free parameters $\phi_s(0)$ and $\phi_s(1)$ from the order-2 MCCL, we need to maximize the log likelihood

$$l(\phi_s(0), \phi_s(1)) = \sum_{x_s} \sum_{x_{s+1}} \sum_{x_{s+2}} n_{s,s+1,s+2} \log f(x_s, x_{s+1}, x_{s+2}) \qquad (5.2)$$

within the parameter space, where $f(x_s, x_{s+1}, x_{s+2}) = P(X_s = x_s, X_{s+1} = x_{s+1}, X_{s+2} = x_{s+2})$ can be calculated by (4.1).

5.1.3 Reparametrization for $(m+1)$-wise margins

More generally, in order to estimate $(m+1)$-wise margins,

$$\pi_{s,s+1,\ldots,s+m}(x_s, x_{s+1}, \cdots, x_{s+m}),$$

for $s = 1, \ldots, L-m$ from the order-m MCCL, we first fix the corresponding lower margins by estimating them from lower order MCCL's. Again, because $\pi_{s,s+1,\ldots,s+m}(x_s, x_{s+1}, \cdots, x_{s+m})$ must satisfy lower margin consistency constraint, not all of them are free parameters. In this case, we actually have 2^{m-1} free parameters.

To select these free parameters, we partition all 2^{m+1} unknown parameters into 2^{m-1} groups with four parameters in each group, such that all the parameters that have same values on middle sites $s+1, \ldots, s+m-1$ should be assigned into one group. The free parameter can be chosen by randomly selecting one parameter from each of 2^{m-1} groups. We have chosen $\pi_{s,s+1,\ldots s+m-1,s+m}(0, x_{s+1}, \cdots, x_{s+m-1}, 0)$ for $s = 1, \ldots, L-m$ as the free parameters.

Let

$$\pi_{s,s+1,\ldots s+m-1,s+m}(0, x_{s+1}, \cdots, x_{s+m-1}, 0) \triangleq \phi_s(\underline{x}),$$

where $\underline{x} = (x_{s+1}, \cdots, x_{s+m-1})$. The other three parameters in the same group can be rewritten as

$$\begin{aligned}
\pi_{s,\ldots s+m}(0, \underline{x}, 1) &= \hat{\pi}_{s,\ldots s+m-1}(0, \underline{x}) - \phi_s(\underline{x}) \\
\pi_{s,\ldots s+m}(1, \underline{x}, 0) &= \hat{\pi}_{s+1,\ldots s+m}(\underline{x}, 0) - \phi_s(\underline{x}) \\
\pi_{s,\ldots s+m}(1, \underline{x}, 1) &= \hat{\pi}_{s+1,\ldots s+m-1}(\underline{x}) - \hat{\pi}_{s,\ldots s+m-1}(0, \underline{x}) - \hat{\pi}_{s+1,\ldots s+m}(\underline{x}, 0) + \phi_s(\underline{x}),
\end{aligned}$$

because all of the $\pi_{s,s+1,\ldots,s+m}(x_s, x_{s+1}, \cdots, x_{s+m})$'s must satisfy the lower margin consistency constraint.

In addition, because all of the 2^{m+1} parameters also need to satisfy the between zero and one restriction, the parameter space for the free parameter $\phi_s(\underline{x})$ is

$$L(\underline{x}) \leq \phi_s(\underline{x}) \leq U(\underline{x}), \tag{5.3}$$

where

$$\begin{aligned} L(\underline{x}) &= \max\{0, \hat{\pi}_{s,\ldots s+m-1}(0,\underline{x}) + \hat{\pi}_{s+1,\ldots s+m}(\underline{x},0) - \hat{\pi}_{s+1,\ldots s+m-1}(\underline{x})\} \\ U(\underline{x}) &= \min\{\hat{\pi}_{s,\ldots s+m-1}(0,\underline{x}), \hat{\pi}_{s+1,\ldots s+m}(\underline{x})\}. \end{aligned}$$

The corresponding order-m log MCCL can be separated into terms for each parameter $\phi_s(\underline{x})$, and the log likelihood part used to estimate $\phi_s(\underline{x})$ is simply a sum of terms:

$$l(\phi_s(\underline{x})) = \sum_{x_s} \sum_{x_{s+m}} n_{s,\underline{x},s+m}(x_s, \underline{x}, x_{s+m}) \log f(x_s, \underline{x}, x_{s+m}), \tag{5.4}$$

where $f(x_s, \underline{x}, x_{s+m}) = P(X_s = x_s, \underline{X} = \underline{x}, X_{s+m} = x_{s+m})$ can be calculated by equation (4.1).

5.2 A hierarchical estimator

In the previous section, we have introduced our idea for the new estimator briefly: we first estimate all the onewise marginal ancestral distribution, then estimate the pairwise margins given all the onewise fixed. The threewise margins are estimated by fixing all the pairwise ones. Following this process hierarchically until we have the estimates for the desired m-wise marginal distribution. We call this new estimator the *hierarchical estimator* and will discuss the detailed estimator in this section.

5.2.1 Estimator for pairwise margins

For clear illustration, we first discuss the hierarchical estimator for estimating the pairwise marginal ancestral distribution, $\pi_{s,s+1}(x_s, x_{s+1})$, for $s = 1, \ldots, L-1$ and $x_s, x_{s+1} \in \{0, 1\}$, when all the onewise margins are fixed as $\hat{\pi}_s(x_s)$.

The first step for the estimator is to estimate onewise margins $\pi_s(x_s)$, $s = 1, ..., L$. As discussed in section 4.4.2, we could use sample proportions as the estimates for $\pi_s(x_s)$. That is $\hat{\pi}_s(x_s) = n_s(x_s)/n$.

Then, we estimate $\pi_{s,s+1}(x_s, x_{s+1})$ through new parameter $\phi_s = \pi_{s,s+1}(0,0)$ given $\hat{\pi}_s(x_s)$ fixed. To estimate ϕ_s, we maximize the log likelihood function (5.1) by solving the score equation

$$\frac{d}{d\phi_s} l(\phi_s) = 0, \tag{5.5}$$

where

$$\begin{aligned}
\frac{d}{d\phi_s} l(\phi_s) &= \frac{n_{s,s+1}(0,0)(1-q)}{(1-q)\phi_s + q\hat{\pi}_s(0)\hat{\pi}_{s+1}(0)} \\
&+ \frac{n_{s,s+1}(0,1)[-(1-q)]}{(1-q)[\hat{\pi}_s(0) - \phi_s] + q\hat{\pi}_s(0)\hat{\pi}_{s+1}(1)} \\
&+ \frac{n_{s,s+1}(1,0)[-(1-q)]}{(1-q)[\hat{\pi}_{s+1}(0) - \phi_s] + q\hat{\pi}_s(1)\hat{\pi}_{s+1}(0)} \\
&+ \frac{n_{s,s+1}(1,1)(1-q)}{(1-q)[1 - \hat{\pi}_s(0) - \hat{\pi}_{s+1}(0) + \hat{\pi}_{s,s+1}(0,0)] + q\hat{\pi}_s(1)\hat{\pi}_{s+1}(1)}.
\end{aligned} \tag{5.6}$$

In addition, the second derivative of $l(\phi_s)$ with respect to ϕ_s is

$$\begin{aligned}
\frac{d^2}{d\phi_s^2} l(\phi_s) &= (-1) \cdot \frac{n_{s,s+1}(0,0)(1-q)^2}{[(1-q)\phi_s + q\hat{\pi}_s(0)\hat{\pi}_{s+1}(0)]^2} \\
&+ (-1) \cdot \frac{n_{s,s+1}(0,1)[-(1-q)]^2}{[(1-q)[\hat{\pi}_s(0) - \phi_s] + q\hat{\pi}_s(0)\hat{\pi}_{s+1}(1)]^2} \\
&+ (-1) \cdot \frac{n_{s,s+1}(1,0)[-(1-q)]^2}{[(1-q)[\hat{\pi}_{s+1}(0) - \phi_s] + q\hat{\pi}_s(1)\hat{\pi}_{s+1}(0)]^2} \\
&+ (-1) \cdot \frac{n_{s,s+1}(1,1)(1-q)^2}{[(1-q)[1 - \hat{\pi}_s(0) - \hat{\pi}_{s+1}(0) + \hat{\pi}_{s,s+1}(0,0)] + q\hat{\pi}_s(1)\hat{\pi}_{s+1}(1)]^2} \\
&\leq 0.
\end{aligned}$$

Thus, $l(\phi_s)$ is a concave function with respect to ϕ_s, which means there typically exists a unique local maximum for $l(\phi_s)$.

It is not easy to solve the above score equation (5.5) directly. However, the following lemma tells us there is a closed form solution for this score equation, and the solution is

$$\hat{\phi}_s = \frac{1}{1-q}\left[\frac{n_{s,s+1}(0,0)}{n} - q\hat{\pi}_s(0)\hat{\pi}_{s+1}(0)\right] \tag{5.7}$$

Lemma 5.1. *Function* (5.6) *achieves zero when*

$$\phi_s = \frac{1}{1-q}\left[\frac{n_{s,s+1}(0,0)}{n} - q\hat{\pi}_s(0)\hat{\pi}_{s+1}(0)\right],$$

given $\hat{\pi}_s(x_s) = n_s(x_s)/n$, *where* $s = 1, \ldots, L$ *and* $x_s \in \{0, 1\}$.

This lemma can be proved easily by plugging in

$$\phi_s = \frac{1}{1-q}\left[\frac{n_{s,s+1}(0,0)}{n} - q\hat{\pi}_s(0)\hat{\pi}_{s+1}(0)\right] \text{ and } \hat{\pi}_s(x_s) = \frac{n_s(x_s)}{n}$$

into the function (5.6).

Notice that the above solution (5.7) is the solution of score equation (5.5) when there is no constraint on the parameter space, but actually, the parameter space for ϕ_s is restricted to be

$$\max\{0, \hat{\pi}_s(0) + \hat{\pi}_{s+1}(0) - 1\} \leq \phi_s \leq \min\{\hat{\pi}_s(0), \hat{\pi}_{s+1}(0)\}. \tag{5.8}$$

If $\hat{\phi}_s$ is in the above interval, then it is the desired constrained MLE with respect to log likelihood function (5.1). If $\hat{\phi}_s$ isn't in parameter space, then the desired constrained MLE must locate on one of the two boundaries in the above interval because of the concaveness of $l(\phi_s)$. In particular, the constrained MLE must be the bound that is closest to $\hat{\phi}_s$, in the interval (5.8) .

Thus, the hierarchical estimator for pairwise margins could simply be:

1. Fix all onewise margins as $\hat{\pi}_s(x_s) = n_s(x_s)/n$ for $s = 1, \ldots, L$.
2. Calculate $\hat{\phi}_s$ by using equation (5.7).
3. Check if $\hat{\phi}_s$ is in the interval (5.8). If yes, $\hat{\phi}_s$ is the desired estimate. If no, the desired estimate should be the bound that is closest to $\hat{\phi}_s$ in the interval (5.8).
4. Calculate the estimate for $\pi_{s,s+1}(0,1)$, $\pi_{s,s+1}(1,0)$, and $\pi_{s,s+1}(1,1)$ through reparametrization in section 5.1.1.

5.2.2 Estimator for threewise margins

The above estimator can be generalized to estimate the threewise marginal distribution. To estimate $\pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2})$, for $s = 1, \ldots, L-2$, we first fix onewise margins $\hat{\pi}_s(x_s)$ and pairwise margins $\hat{\pi}_{s,s+1}(x_s, x_{s+1})$. The unknown parameters, $\pi_{s,s+1,s+2}(x_s, x_{s+1}, x_{s+2})$, can be reparametrized as $\phi_s(0) = \pi_{s,s+1,s+2}(0,0,0)$ and $\phi_s(1) = \pi_{s,s+1,s+2}(0,1,0)$ based on the discussions in section 5.1.2. These two new parameters can be estimated by optimizing the log likelihood function (5.2).

Moreover, since these two parameters $\phi_s(0)$ and $\phi_s(1)$ are separable in log likelihood function (5.2). Thus, we have

$$l(\phi_s(0), \phi_s(1)) = l(\phi_s(0)) + l(\phi_s(1)),$$

where

$$l(\phi_s(i)) = \sum_{x_s=0}^{1} \sum_{x_{s+2}=0}^{1} n_{s,s+1,s+2}(x_s, i, x_{s+2}) \log f(x_s, x_{s+1} = i, x_{s+2}), \quad (5.9)$$

for $i = 0$ or 1, and $f(x_s, x_{s+1=i}, x_{s+2}) = P(X_s = x_s, X_{s+1} = i, X_{s+2} = x_{s+2})$ can be calculated by equation (4.1).

To estimate the parameter $\phi_s(i)$, for $i = \{0, 1\}$, one can solve the score equation

$$\frac{d}{d\phi_s(i)} l(\phi_s(i)) = 0, \quad (5.10)$$

subject to

$$\begin{aligned} \phi_s(i) &\geq \max\{0, \hat{\pi}_{s,s+1}(0,i) + \hat{\pi}_{s+1,s+2}(i,0) - \hat{\pi}_{s+1}(i)\} \quad (5.11) \\ \phi_s(i) &\leq \min\{\hat{\pi}_{s,s+1}(0,i), \hat{\pi}_{s+1,s+2}(i,0)\} \end{aligned}$$

Similar to the pairwise margins case, we can prove that

$$\frac{d^2}{d\phi_s(i)^2} l(\phi_s(i)) \leq 0, \quad (5.12)$$

which imply that $l(\phi_s(i))$ is concave function respect to $\phi_s(i)$. Hence, there exists local maximum for $l(\phi_s(i))$. The general solution for score equation (5.10) will be derived soon, but the following lemma provides us a simple solution in special cases.

Lemma 5.2. *Function* $\frac{d}{d\phi_s(i)} l(\phi_s(i))$ *in equation* (5.10) *achieves zero when*

$$\begin{aligned}\phi_s(i) &= \frac{1}{(1-q)^2}\{\frac{n_{s,s+1,s+2}(0,i,0)}{n} - q(1-q)[\hat{\pi}_s(0)\hat{\pi}_{s+1,s+2}(i,0) + \hat{\pi}_{s,s+1}(0,i)\hat{\pi}_{s+2}(0)] \\ &\quad -q^2\hat{\pi}_s(0)\hat{\pi}_{s+1}(i)\hat{\pi}_{s+3}(0)\}, \end{aligned} \tag{5.13}$$

provided that $\hat{\pi}_s(x_s) = n_s(x_s)/n$ *and* $\hat{\pi}_{s,s+1}(0,0) = \hat{\phi}_s$ *is as in equation* (5.7) *(not a solution forced by constraint).*

It would seem, following the similar estimator for pairwise margins, that one could use solution (5.13) as the candidate estimate, say $\hat{\phi}_s(i)$, and then compare it with the parameter space (5.11) to decide whether we should use $\hat{\phi}_s(i)$ or one of the bounds in interval (5.11) as the constrained MLE for $\phi_s(i)$. However, as stated in Lemma 5.2, the solution (5.13) can make $\frac{d}{d\phi_s(i)} l(\phi_s(i))$ achieve zero only when $\hat{\pi}_{s,s+1}(0,0)$ satisfies (5.7), the non boundary case. That is, if the constrained MLE for $\hat{\pi}_{s,s+1}(0,0)$ is not $\hat{\phi}_s$ in equation (5.7), but rather one of two bounds in (5.8), it cannot be proved that (5.13) is the solution for score equation (5.10). Hence, the pairwise hierarchical estimator cannot be directly generalized to the threewise margin case.

By further exploring, we find that $f(x_s, x_{s+1} = i, x_{s+2})$ in log likelihood (5.9) are all in the linear form of

$$f(x_s, x_{s+1} = i, x_{s+2}) = a^s_{x_s,x_{s+2}}\phi_s(i) + b^s_{x_s,x_{s+2}},$$

for $x_s \in \{0,1\}$ and $x_{s+1} \in \{0,1\}$, where

$$
\begin{aligned}
a^s_{00} &= a^s_{11} = (1-q)^2, a^s_{01} = a^s_{10} = -(1-q)^2, \\
b^s_{00} &= q(1-q)[\hat{\pi}_s(0)\hat{\pi}_{s+1,s+2}(i,0) + \hat{\pi}_{s,s+1}(0,i)\hat{\pi}_{s+2}(0)] + q^2\hat{\pi}_s(0)\hat{\pi}_{s+1}(i)\hat{\pi}_{s+2}(0), \\
b^s_{01} &= (1-q)^2\hat{\pi}_{s,s+1}(0,i) + q(1-q)[\hat{\pi}_s(0)\hat{\pi}_{s+1,s+2}(i,1) + \hat{\pi}_{s,s+1}(0,i)\hat{\pi}_{s+2}(1)] \\
&\quad + q^2\hat{\pi}_s(0)\hat{\pi}_{s+1}(i)\hat{\pi}_{s+2}(1), \\
b^s_{10} &= (1-q)^2\hat{\pi}_{s+1,s+2}(i,0) + q(1-q)[\hat{\pi}_s(1)\hat{\pi}_{s+1,s+2}(i,0) + \hat{\pi}_{s,s+1}(1,i)\hat{\pi}_{s+2}(0)] \\
&\quad + q^2\hat{\pi}_s(1)\hat{\pi}_{s+1}(i)\hat{\pi}_{s+2}(0), \\
b^s_{11} &= (1-q)^2[\hat{\pi}_{s+1}(i) - \hat{\pi}_{s,s+1}(0,i) - \hat{\pi}_{s+1,s+2}(i,0)] \\
&\quad + q(1-q)[\hat{\pi}_s(1)\hat{\pi}_{s+1,s+2}(i,1) + \hat{\pi}_{s,s+1}(1,i)\hat{\pi}_{s+2}(1)] + q^2\hat{\pi}_s(1)\hat{\pi}_{s+1}(i)\hat{\pi}_{s+2}(1).
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
& \frac{d}{d\phi_s(i)} l(\phi_s(i)) \qquad (5.14) \\
= & \sum_{x_s=0}^{1}\sum_{x_{s+2}=0}^{1} n_{s,s+1,s+2}(x_s, i, x_{s+2}) \log f(x_s, x_{s+1}=i, x_{s+2}) \\
= & \sum_{x_s=0}^{1}\sum_{x_{s+2}=0}^{1} n_{s,s+1,s+2}(x_s, i, x_{s+2}) \frac{a^s_{x_s,x_{s+2}}}{f(x_s, x_{s+1}=i, x_{s+2})} \\
= & \frac{A_1\phi_s(i)^3 + A_2\phi_s(i)^2 + A_3\phi_s(i)^3 + A_4}{\prod_{x_s=0}^{1}\prod_{x_{s+2}=0}^{1} f(x_s, x_{s+1}=i, x_{s+2})},
\end{aligned}
$$

where

$$
\begin{aligned}
A_1^3 &= \left(\sum_{x_s,x_{s+2}} n^s(x_s,i,x_{s+2})\right)\left(\prod_{x_s,x_{s+2}} a^s_{x_s,x_{s+2}}\right),\\
A_2^3 &= (n^s(0,i,0)+n^s(0,i,1)+n^s(1,i,0))a^s_{00}a^s_{01}a^s_{10}b^s_{11}\\
&\quad +(n^s(0,i,0)+n^s(0,i,1)+n^s(1,i,1))a^s_{00}a^s_{01}b^s_{10}a^s_{11}\\
&\quad +(n^s(0,i,0)+n^s(1,i,0)+n^s(1,i,1))a^s_{00}b^s_{01}a^s_{10}a^s_{11}\\
&\quad +(n^s(0,i,1)+n^s(1,i,0)+n^s(1,i,1))b^s_{00}a^s_{01}a^s_{10}a^s_{11},\\
A_3^3 &= (n^s(0,i,0)+n^s(0,i,1))a^s_{00}a^s_{01}b^s_{10}b^s_{11}+(n^s(0,i,0)+n^s(1,i,0))a^s_{00}b^s_{01}a^s_{10}b^s_{11}\\
&\quad +(n^s(0,i,0)+n^s(1,i,1))a^s_{00}b^s_{01}b^s_{10}a^s_{11}+(n^s(0,i,1)+n^s(1,i,0))b^s_{00}a^s_{01}a^s_{10}b^s_{11}\\
&\quad +(n^s(0,i,1)+n^s(1,i,1))b^s_{00}a^s_{01}b^s_{10}a^s_{11}+(n^s(1,i,0)+n^s(1,i,1))b^s_{00}b^s_{01}a^s_{10}a^s_{11},\\
A_4^3 &= n^s(0,i,0)a^s_{00}b^s_{01}b^s_{10}b^s_{11}+n^s(0,i,1)b^s_{00}a^s_{01}b^s_{10}b^s_{11}\\
&\quad +n^s(1,i,0)b^s_{00}b^s_{01}a^s_{10}b^s_{11}+n^s(1,i,1)b^s_{00}b^s_{01}b^s_{10}a^s_{11},\\
n^s(x_s,i,x_{s+1}) &= n_{s,s+1,s+2}(x_s,i,x_{s+2}) \text{ where } x_s,x_{s+2}\in\{0,1\}.
\end{aligned}
$$

Solving the score equation (5.10), is equivalent to solving the cubic equation,

$$A_1^3\phi_s(i)^3+A_2^3\phi_s(i)^2+A_3^3\phi_s(i)+A_4^3=0, \tag{5.15}$$

provided that the denominator $f(x_s,x_{s+1}=i,x_{s+2})$ is not zero.

Because there are explicit solutions for any cubic equation, we can find the candidate estimate, $\hat{\phi}_s(i)$, by solving equation (5.15) and then comparing $\hat{\phi}_s(i)$ with the parameter space (5.11) to decide whether we should use $\hat{\phi}_s(i)$ or one of the boundaries of the parameter space (5.11) as the desired estimate.

Multiple real roots

For any cubic equations with real coefficients, there exist either one real root and two conjugate complex roots, or three real roots for this equation. Thus, there could exist three real roots for the above cubic equation (5.15). In such

situation, which real root should be chosen as the candidate estimate is a critical problem. Thus, we need to build up a rule for choosing a candidate estimate $\hat{\phi}_s(i)$ when there exist multiple real roots.

Before building the selection rule, we will first illustrate several facts. (Note: The proofs for the following lemma and theorem are all listed in the appendix.)

Lemma 5.3. *Functions* $\frac{d}{d\phi_s(i)}l(\phi_s(i))$ *and* $\frac{d^2}{d\phi_s(i)^2}l(\phi_s(i))$ *have no definitions when*

$$\phi_s(i) = -\frac{b^s_{00}}{a^s_{00}}, -\frac{b^s_{01}}{a^s_{01}}, -\frac{b^s_{10}}{a^s_{10}}, \text{ or } -\frac{b^s_{11}}{a^s_{11}}.$$

In addition, when $q < 1$, we have $a^s_{00} > 0$, $a^s_{11} > 0$, $a^s_{01} < 0$, $a^s_{10} < 0$, and all $b^s_{x_s x_{s+1}} \geq 0$. Denote

$$\phi^{(1)}_s(i) \triangleq -\frac{b^s_{00}}{a^s_{00}}, \ \ \phi^{(2)}_s(i) \triangleq -\frac{b^s_{01}}{a^s_{01}}, \ \ \phi^{(3)}_s(i) \triangleq -\frac{b^s_{10}}{a^s_{10}}, \ \ \phi^{(4)}_s(i) \triangleq -\frac{b^s_{11}}{a^s_{11}}.$$

Then, the relationship among $\phi^{(k)}_s(i)$, $k = 1, 2, 3, 4$, is

$$\min\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\} \leq \max\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\} \leq \min\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\} \leq \max\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}.$$

Lemma 5.4. *The cubic equation* (5.15) *has no real root in the intervals* $\left(-\infty, \min\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\right)$ *and* $\left(\max\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}, \infty\right)$.

Lemma 5.5. *There is at most one real solution for the cubic equation* (5.15) *when* $\phi_s(i)$ *belongs to the open intervals* $\left(\min\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\}, \max\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\}\right)$, $\left(\max\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\}, \min\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}\right)$, *and* $\left(\min\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}, \max\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}\right)$.

Lemma 5.6. *The parameter space for* $\phi_s(i)$, *i.e. interval* (5.11), *is a sub set of* $\left[\max\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\}, \min\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}\right]$.

With all above lemmas, we have the following theorem

Theorem 5.7. *If there exist three real roots for cubic function* (5.15), *there is only one real root is the eligible candidate estimate for* $\phi_s(i)$, *such that* $f(x_s, x_{s+1} = i, x_{s+2}) \geq 0$ *for* $x_s \in \{0, 1\}$ *and* $x_{s+1} \in \{0, 1\}$. *In addition, this root is the middle one among all three real roots.*

Based on Theorem 5.7, we build up the root selection rule for our hierarchical threewise estimator under the multiple roots situation: select the middle real root as the candidate estimate for $\phi_s(i)$ and denote it as $\hat{\phi}_s(i)$.

Since by Lemma 5.6, the parameter space of $\phi_s(i)$ is a subset of interval $\left[\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\right]$. Thus, if $\hat{\phi}_s(i)$ is also in the parameter space, it must be the conditional constrained MLE we want. If $\hat{\phi}_s(i)$ isn't in the parameter space, we should use the boundary that is nearest to $\hat{\phi}_s(i)$ in the interval (5.11) as the desired estimate.

Therefore, the hierarchical estimator for estimating threewise marginal ancestral distribution is

1. Fix all onewise margins as $\hat{\pi}_s(x_s) = n_s(x_s)/n$ and pairwise margins $\hat{\pi}_{s,s+1}(x_s, x_{s+1})$, which is estimated by hierarchical estimator for pairwise margins in section 5.2.1.

2. Solve the cubic function (5.15) and find all its real roots.

3. If there is only one real root, then use this root as the candidate estimate of $\phi_s(i)$, otherwise use the middle real root as the candidate estimate, denoted by $\hat{\phi}_s(i)$

4. Check if $\hat{\phi}_s(i)$ is in the interval (5.11). If yes, $\hat{\phi}_s(i)$ is the desired estimate. If no, the desired estimate should be the bound that is most closest to $\hat{\phi}_s(i)$ in the interval (5.11).

5. Calculate the estimate for other $\pi_{s,s+1,s+2}(x_s, i, x_{s+2})$'s through reparametrization in section 5.1.2.

5.2.3 Estimator for $(m+1)$-wise margins with $m \geq 2$

As discussed in section 5.1.3, to estimate $(m+1)$-wise margins for $m \geq 2$, the new parameters are $\phi_s(\underline{x})$, where $\underline{x} = (x_{s+1}, \ldots, x_{s+m-1})$, $x_{s+1} \in \{0, 1\}, \ldots, x_{s+m-1} \in$

$\{0,1\}$. Since the new parameters are separable in the order-m log MCCL, we can estimate $\phi_s(\underline{x})$ by maximizing the log likelihood function (5.4). Similar to estimating threewise margins, this optimization problem is equivalent to solving cubic equation

$$A_1^m \phi_s(\underline{x})^3 + A_2^m \phi_s(\underline{x})^2 + A_3^m \phi_s(\underline{x}) + A_4^m = 0, \tag{5.16}$$

where

$$\begin{aligned}
A_1^m &= \left(\sum_{x_s, x_{s+m}} n^s(x_s, \underline{x}, x_{s+m})\right)\left(\prod_{x_s, x_{s+m}} a^s_{x_s, x_{s+m}}\right), \\
A_2^m &= (n^s(0,\underline{x},0) + n^s(0,\underline{x},1) + n^s(1,\underline{x},0))a^s_{00}a^s_{01}a^s_{10}b^s_{11} \\
&\quad +(n^s(0,\underline{x},0) + n^s(0,\underline{x},1) + n^s(1,\underline{x},1))a^s_{00}a^s_{01}b^s_{10}a^s_{11} \\
&\quad +(n^s(0,\underline{x},0) + n^s(1,\underline{x},0) + n^s(1,\underline{x},1))a^s_{00}b^s_{01}a^s_{10}a^s_{11} \\
&\quad +(n^s(0,\underline{x},1) + n^s(1,\underline{x},0) + n^s(1,\underline{x},1))b^s_{00}a^s_{01}a^s_{10}a^s_{11}, \\
A_3^m &= (n^s(0,\underline{x},0) + n^s(0,\underline{x},1))a^s_{00}a^s_{01}b^s_{10}b^s_{11} + (n^s(0,\underline{x},0) + n^s(1,\underline{x},0))a^s_{00}b^s_{01}a^s_{10}b^s_{11} \\
&\quad +(n^s(0,\underline{x},0) + n^s(1,\underline{x},1))a^s_{00}b^s_{01}b^s_{10}a^s_{11} + (n^s(0,\underline{x},1) + n^s(1,\underline{x},0))b^s_{00}a^s_{01}a^s_{10}b^s_{11} \\
&\quad +(n^s(0,\underline{x},1) + n^s(1,\underline{x},1))b^s_{00}a^s_{01}b^s_{10}a^s_{11} + (n^s(1,\underline{x},0) + n^s(1,\underline{x},1))b^s_{00}b^s_{01}a^s_{10}a^s_{11}, \\
A_4^m &= n^s(0,\underline{x},0)a^s_{00}b^s_{01}b^s_{10}b^s_{11} + n^s(0,\underline{x},1)b^s_{00}a^s_{01}b^s_{10}b^s_{11} \\
&\quad +n^s(1,\underline{x},0)b^s_{00}b^s_{01}a^s_{10}b^s_{11} + n^s(1,\underline{x},1)b^s_{00}b^s_{01}b^s_{10}a^s_{11},
\end{aligned}$$

and

$$\begin{aligned}
n^s(x_s, \underline{x}, x_{s+m}) &= n_{s,\underline{x},s+m}(x_s, \underline{x}, x_{s+m}), \\
a^s_{x_s, x_{s+m}} &= (-1)^{x_s + x_{s+m}} \cdot (1-q)^m, \\
b^s_{x_s, x_{s+m}} &= f(x_s, \underline{x}, x_{s+m}) - a^m_{x_s, x_{s+m}} \phi_s(\underline{x}),
\end{aligned}$$

for $x_s, x_{s+m} \in \{0,1\}$.

Since the lemma and theorem in section 5.2.2 are still true for the cubic equation (5.16), the hierarchical estimator for $(m+1)$-wise margins, when $m \geq 2$ is

1. Fix all the corresponding estimates for lower marginal ancestral distribution.

2. Solve the cubic function (5.16) and find all its real roots.

3. If there is only one real root, then use this root as the candidate estimate of $\phi_s(\underline{x})$, otherwise use the middle real root as the candidate estimate, denoted by $\hat{\phi}_s(\underline{x})$

4. Check if $\hat{\phi}_s(\underline{x})$ is in the interval (5.3). If yes, $\hat{\phi}_s(\underline{x})$ is the desired estimate. If no, the desired estimate should be the bound that is most closest to $\hat{\phi}_s(\underline{x})$ in the interval (5.3).

5. Calculate the estimates for other $\pi_{s,\ldots,s+m}(x_s, \underline{x}, x_{s+m})$'s through reparametrization in section 5.1.3.

5.3 Comparison between left to right and hierarchical estimators

Both the left to right estimator and the hierarchical estimator are based on sequential maximization. That is, we estimate the marginal ancestral distribution sequentially by fixing the lower margins. The difference between these two estimators is that we fix the lower margins of the left (or right) neighbors when using the left to right estimator, but fix all the lower margins when using the hierarchical estimator.

As discussed at the beginning of this chapter, there are several cons for the left to right estimator, including the cumulative errors which could cause potential left-right effect, and multivariate optimization because the θ parameters are not separable in the likelihood function. Also, there is no natural hierarchy by which estimates from lower order m can be turned into higher order estimates. That is, the order m needs to be chosen. In the hierarchical method, one could derive the estimator so that the order m used was based on computation time.

For the hierarchical estimator the estimates of the marginal ancestral distribution at any one sequence segment are conditional on marginal to the left and right equally. Hence, there can be no such left-right effect exist for this hierarchical estimator. In addition, since no matter which order of margins we want to estimate, the optimization problem separates parameters, and so is always equivalent to solving a cubic equation problem. There is a closed form solution for solving the cubic equation, hence this estimator can be extended to any order m without multivariate optimization.

5.4 Simulation study

In this section, we will examine the performance of the hierarchical estimator through simulation studies. We will use the same simulated data sets as in chapter 4. We will compare the estimated marginal and joint ancestral distributions, obtained by using the hierarchical estimator, with the true marginal and joint ancestral distribution. In addition, we will also compare the performances of the hierarchical estimator with the left to right estimator by comparing the estimated margins, joint distribution, and the computation time. Let $\hat{\pi}^h$ denote the estimated marginal or joint distribution by applying the hierarchical estimator. To simplify the simulation report, when comparing with the left to right estimator, we only compare $\hat{\pi}^h$ with the best linear combination $\hat{\pi}^\alpha$.

5.4.1 Estimated marginal distributions

First, we compare the estimated margins $\hat{\pi}^h$, obtained by using the hierarchical estimator, with $\hat{\pi}^\alpha$, obtained by using the left to right estimator. Table 5.1 shows the comparison for estimated pairwise margins between two estimators in data set 1, where $L = 10$ and $q = 0.01$.

From the table we can see that there is not much difference between the two

estimators for estimating the marginal distributions. In addition, the biases and the standard errors are small for both two estimators. Since the similar conclusion can be drawn for higher order margins and other data sets (see Figure C.1 to C.8 in Appendix C for details), we focus on comparing the estimated margins $\hat{\pi}^h$ with the true margins τ in the following.

Next, we consider the effect of recombination factor for estimating marginal distribution. Figure 5.1 and Figure 5.2 show the estimated pairwise margins when $L = 10$, $q = 0.01, 0.05, 0.1$. From the figures we can see that similar to the left to right estimator, the recombination factor doesn't have a large effect when we applying the hierarchical estimator to do the inference. Similar results can be found for higher order marginal distribution and other data sets (see Figure C.9 to C.16 in Appendix C for details).

5.4.2 Estimated joint distributions

In this subsection, we compare the estimated joint ancestral distribution, $\hat{\pi}^h(\underline{x})$, obtained by using the hierarchical estimator, with the true ancestral distribution $\tau(\underline{x})$ and the estimates $\hat{\pi}^\alpha(\underline{x})$ obtained by using the left to right estimator.

We first check the ability to identify the true ancestors by using the hierarchical estimator. As discussed in section 5.4.1, our simulation results showed the performances of both the hierarchical and the left to right estimator for estimating marginal distribution are similar. In addition, since we still use the Markov chain property to reconstruct the joint distribution from the margins, we expect that the ability of the hierarchical estimator to identify the true ancestor would be similar to the left to right estimator.

Table 5.2 compares the estimated ancestral distributions obtained from $\hat{\pi}^\alpha$ and $\hat{\pi}^h$ by using order-m MCCL for $m = 1, 2, 3$, to the true ancestral distribution τ. The simulated data used is data set 1, where $L = 10$ and $q = 0.01$. There

Table 5.1: Estimated pairwise margins using the left to right estimator and the hierarchical estimator. Here $L = 10$, $q = 0.01$.

site	(1,2)	(2,3)	(3,4)	(4,5)	(5,6)	(6,7)	(7,8)	(8,9)	(9,10)
					$\tau_{s,s+1}(0,0)$				
true	0.0000	0.0000	0.0500	0.0500	0.0500	0.3917	0.5417	0.4583	0.7500
					$\hat{\pi}^h_{s,s+1}(0,0)$				
bias	0.0004	0.0000	0.0027	0.0022	0.0023	0.0060	-0.0060	-0.0059	-0.0049
sd	0.0020	0.0000	0.0234	0.0237	0.0238	0.0486	0.0522	0.0500	0.0479
					$\hat{\pi}^\alpha_{s,s+1}(0,0)$				
bias	0.0004	0.0000	0.0028	0.0023	0.0024	0.0058	-0.0049	-0.0053	-0.0045
sd	0.0021	0.0000	0.0233	0.0237	0.0238	0.0486	0.0522	0.0500	0.0479
					$\tau_{s,s+1}(0,1)$				
true	0.6250	0.0167	0.0083	0.0000	0.0000	0.0167	0.4417	0.0833	0.0083
					$\hat{\pi}^h_{s,s+1}(0,1)$				
bias	-0.0049	0.0001	0.0008	0.0009	0.0005	0.0000	0.0058	0.0001	-0.0009
sd	0.0461	0.0137	0.0097	0.0029	0.0021	0.0133	0.0512	0.0278	0.0084
					$\hat{\pi}^\alpha_{s,s+1}(0,1)$				
bias	-0.0051	0.0003	0.0006	0.0009	0.0005	0.0002	0.0047	0.0005	-0.0009
sd	0.0461	0.0137	0.0096	0.0028	0.0021	0.0134	0.0512	0.0279	0.0085
					$\tau_{s,s+1}(1,0)$				
true	0.0167	0.0583	0.0000	0.0000	0.3583	0.5917	0.0000	0.3000	0.1667
					$\hat{\pi}^h_{s,s+1}(1,0)$				
bias	-0.0003	0.0035	0.0004	0.0006	0.0038	-0.0062	0.0001	0.0001	0.0041
sd	0.0139	0.0241	0.0019	0.0028	0.0451	0.0488	0.0010	0.0465	0.0387
					$\hat{\pi}^\alpha_{s,s+1}(1,0)$				
bias	-0.0001	0.0034	0.0004	0.0006	0.0036	-0.0061	0.0001	-0.0001	0.0038
sd	0.0139	0.0240	0.0019	0.0027	0.0452	0.0488	0.0010	0.0465	0.0387
					$\tau_{s,s+1}(1,1)$				
true	0.3583	0.9250	0.9417	0.9500	0.5917	0.0000	0.0167	0.1583	0.0750
					$\hat{\pi}^h_{s,s+1}(1,1)$				
bias	0.0048	-0.0036	-0.0039	-0.0037	-0.0066	0.0001	0.0000	0.0058	0.0017
sd	0.0466	0.0292	0.0243	0.0237	0.0487	0.0010	0.0135	0.0402	0.0296
					$\hat{\pi}^\alpha_{s,s+1}(1,1)$				
bias	0.0048	-0.0037	-0.0038	-0.0038	-0.0065	0.0001	0.0002	0.0049	0.0016
sd	0.0466	0.0291	0.0242	0.0237	0.0487	0.0010	0.0135	0.0401	0.0295

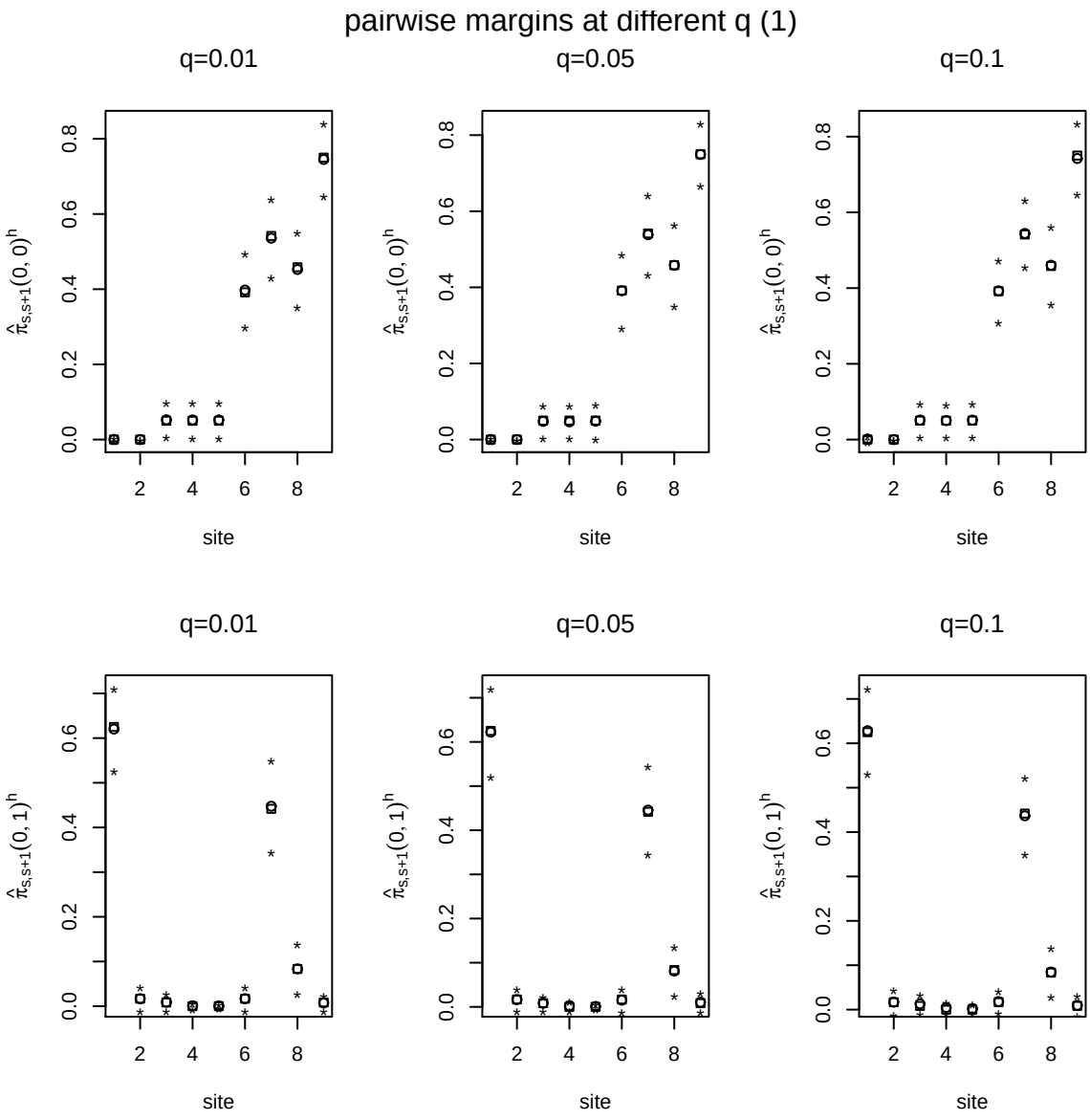


Figure 5.1: Estimated pairwise margins at different q using the hierarchical estimator when $L = 10$. Squares correspond to the true margins, circles are the average estimated margins $\hat{\pi}^h$, and stars above and below the circles represent $\hat{\pi}^h \pm 2s$, where s is the standard error.

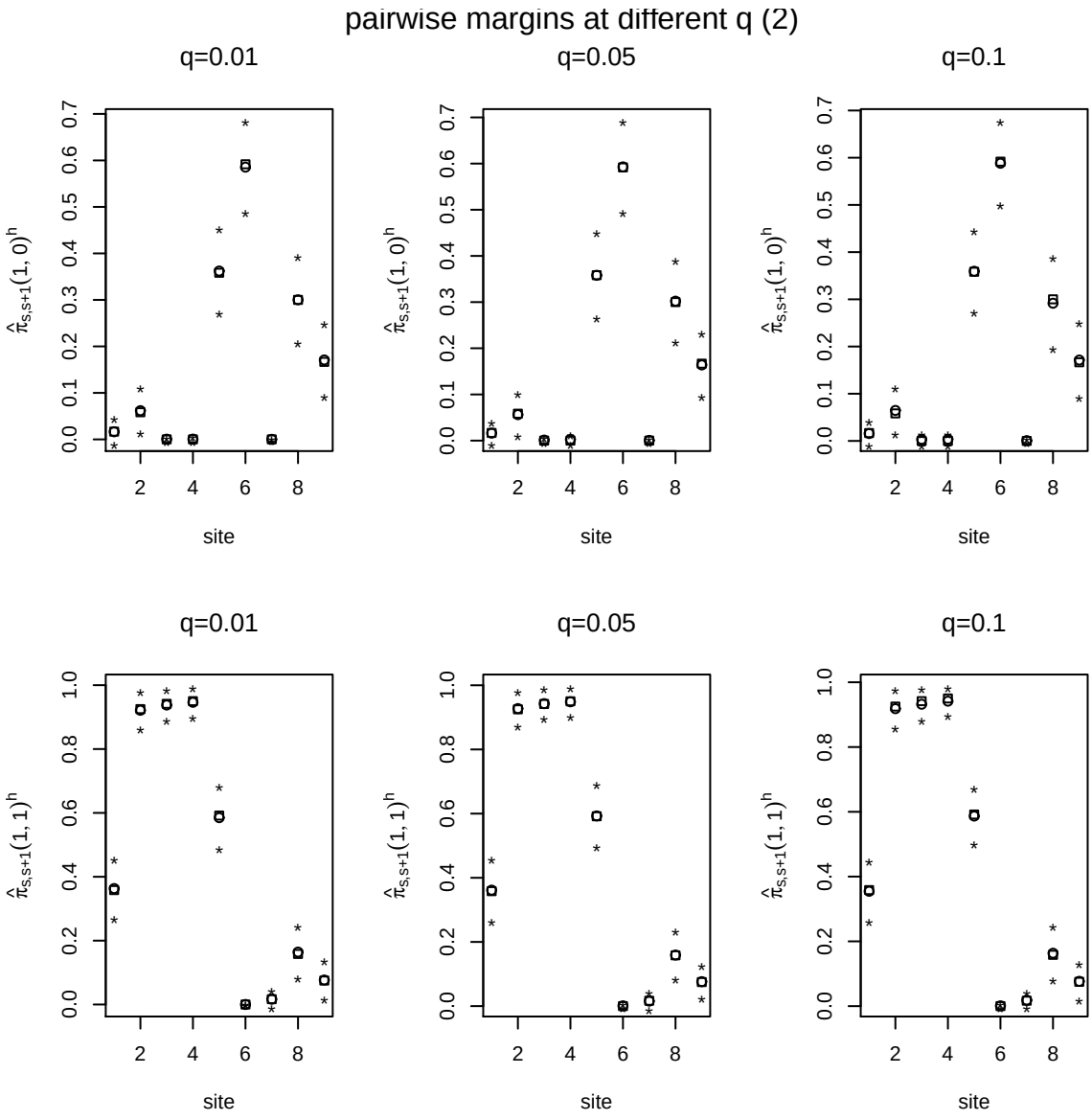


Figure 5.2: Estimated pairwise margins at different q using the hierarchical estimator when $L = 10$. Squares correspond to the true margins, circles are the average estimated margins $\hat{\pi}^h$, and stars above and below the circles represent $\hat{\pi}^h \pm 2s$, where s is the standard error.

are in total 13 true ancestors, denoted by μ^i for $i = 1, \ldots, 13$, and the table only lists the estimated probabilities at the true ancestors. From the table, the joint distribution reconstructed from $\hat{\pi}^h(\mu)$ is a little better than the one reconstructed from $\hat{\pi}^\alpha(\mu)$ because the total probability at true ancestors is larger for $\hat{\pi}^h(\mu)$ than $\hat{\pi}^\alpha(\mu)$, but this difference is not big.

Figure 5.3 shows the comparison between the estimated joint ancestral distribution $\hat{\pi}^h(\mu)$ and the true ancestral distribution $\tau(\mu)$ when $L = 10$ and $q = 0.01$. From both Figure 5.3 and Table 5.2, we can see that, similar to the estimated joint ancestral distribution $\hat{\pi}^\alpha$, there exists considerable bias for $\tau(\mu)$ in $\hat{\pi}^h(\mu)$. One possible explanation for that is that the true ancestral distribution is not a Markov chain.

Figure 5.4 shows the comparison between $\hat{\pi}^h(\mu_{MC})$ with $\tau^m_{MC}(\mu_{MC})$, the Markov chain reconstruction based on true ancestral distribution τ, where $m = 1, 2, 3$ is the order of Markov chain, μ_{MC} represents a binary sequence such that $\tau^m_{MC}(\mu_{MC}) > 0$, and $L = 10$, $q = 0.01$. We can see, if the true ancestral distribution really has the Markov property, our estimates are very good. In addition, in Table 5.3, we also list the sum of estimated probabilities at μ_{MC}. We can also draw the same conclusion from this table because most of the estimated probabilities are assigned to the correct sequences.

Figure 5.5 compares the estimated ancestral distribution $\hat{\pi}^h(\mu)$, which is reconstructed by using order-3 MCCL, with the true ancestral distribution $\tau(\mu)$ under the combinations of $L = 10, 20$ and $q = 0.01, 0.1$. Similar to the simulation results for $\hat{\pi}^\alpha(\mu)$, when L or q increases, the bias also increases. In Table 5.4 we list the sum of the estimated probability at the true ancestors μ. From the table we can see that when L increases, more probabilities are assigned to the non-ancestral sequences and hence the bias increases.

Similar as for $\hat{\pi}^\alpha$, when L increases, more probabilities are assigned to the

Table 5.2: Comparison between true, $\tau(\mu)$, and estimated, $\hat{\pi}^h(\mu)$ and $\hat{\pi}^\alpha(\mu)$, probability at true ancestor μ when $L = 10$ and $q = 0.01$

ancestors	μ^1	μ^2	μ^3	μ^4	μ^5	μ^6	μ^7	μ^8	μ^9	μ^{10}	μ^{11}	μ^{12}	μ^{13}	
	true ancestral distribution													total
$\tau(\mu)$	0.4250	0.2917	0.0750	0.0500	0.0417	0.0250	0.0250	0.0167	0.0167	0.0083	0.0083	0.0083	0.0083	1
	order-1 MC reconstruction													
$\hat{\pi}^\alpha(\mu)$	0.1658	0.0382	0.0218	0.0017	0.0137	0.0395	0.0181	0.0000	0.0950	0.0015	0.0622	0.0016	0.0124	0.4715
sd	0.0378	0.0160	0.0100	0.0017	0.0041	0.0106	0.0081	0.0001	0.0138	0.0018	0.0143	0.0022	0.0051	
$\hat{\pi}^h(\mu)$	0.1657	0.0382	0.0218	0.0018	0.0137	0.0395	0.0181	0.0001	0.0949	0.0016	0.0622	0.0017	0.0124	0.4720
sd	0.0378	0.0160	0.0100	0.0016	0.0041	0.0106	0.0081	0.0002	0.0138	0.0018	0.0143	0.0021	0.0051	
	order-2 MC reconstruction													
$\hat{\pi}^\alpha(\mu)$	0.2624	0.0904	0.0493	0.0081	0.0226	0.0047	0.0043	0.0002	0.1684	0.0036	0.1350	0.0041	0.0314	0.7844
sd	0.0444	0.0289	0.0177	0.0072	0.0074	0.0037	0.0035	0.0004	0.0205	0.0045	0.0218	0.0052	0.0102	
$\hat{\pi}^h(\mu)$	0.2620	0.0903	0.0489	0.0088	0.0230	0.0057	0.0049	0.0003	0.1683	0.0042	0.1348	0.0044	0.0312	0.7866
sd	0.0444	0.0281	0.0176	0.0066	0.0073	0.0033	0.0031	0.0004	0.0200	0.0045	0.0208	0.0052	0.0102	
	order-3 MC reconstruction													
$\hat{\pi}^\alpha(\mu)$	0.2602	0.1018	0.0492	0.0073	0.0154	0.0144	0.0153	0.0002	0.1669	0.0066	0.1526	0.0043	0.0313	0.8254
sd	0.0446	0.0302	0.0177	0.0061	0.0064	0.0075	0.0081	0.0005	0.0203	0.0071	0.0223	0.0051	0.0101	
$\hat{\pi}^h(\mu)$	0.2603	0.1030	0.0489	0.0058	0.0189	0.0186	0.0122	0.0000	0.1671	0.0067	0.1544	0.0043	0.0311	0.8313
sd	0.0445	0.0299	0.0176	0.0052	0.0073	0.0093	0.0071	0.0000	0.0198	0.0072	0.0224	0.0052	0.0101	

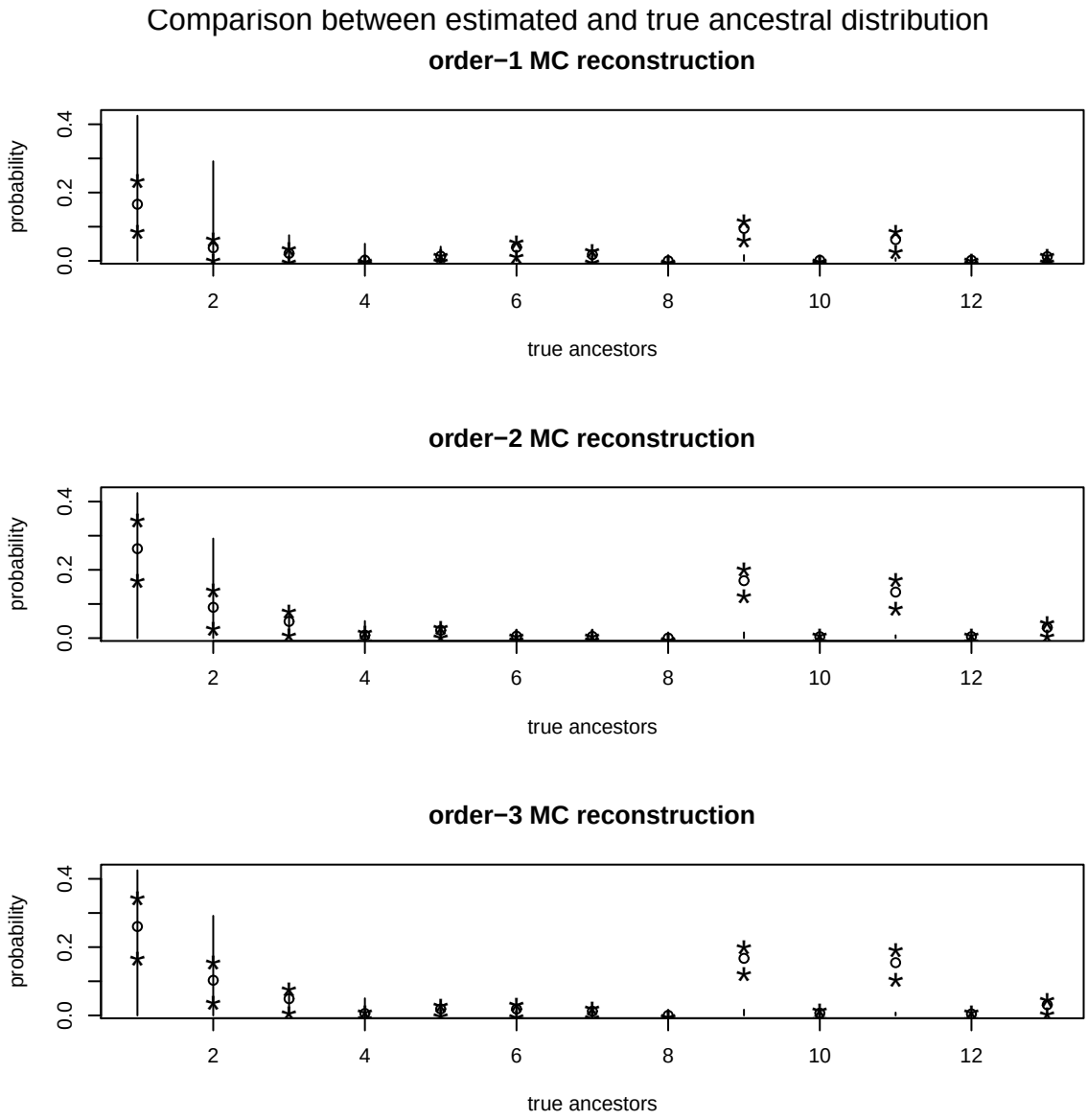


Figure 5.3: Estimated Ancestral Distribution. Here $L = 10$, $q = 0.01$ and order-1,2,3 MC are applied. Solid vertical line show the true probability at the true ancestors, drawn by a decreasing order. Circles are mean $\hat{\pi}^h$ at the true ancestors. Stars above and below the circles correspond to $\hat{\pi}^h \pm 2s$.

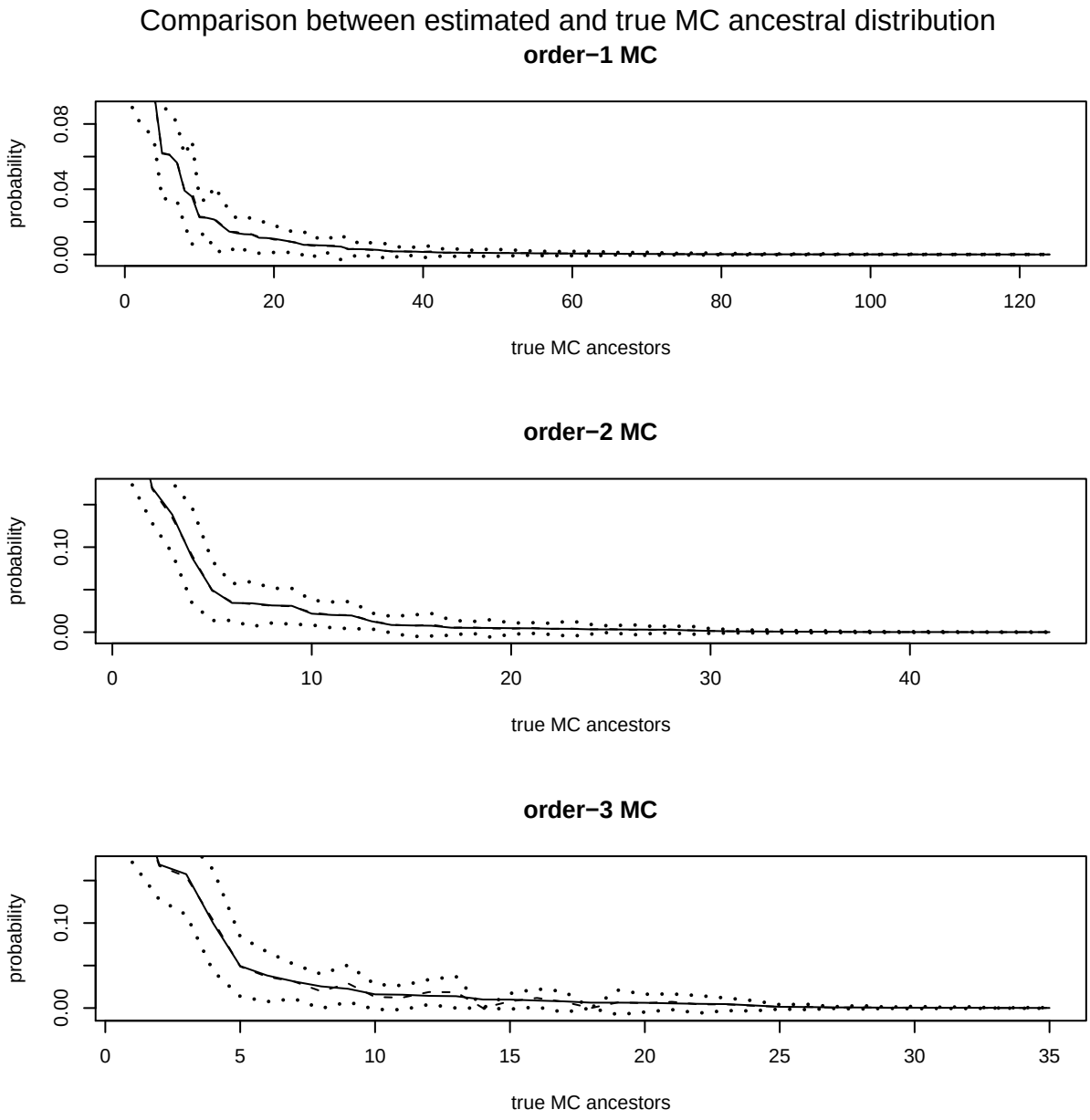


Figure 5.4: Estimated Ancestral Distribution. Here $L = 10$, $q = 0.01$ and order-1,2,3 MC are applied. Solid line is the true probability at the true MC ancestors ordered by frequency. Dashed line is $\hat{\pi}^h$ at the true ancestors. Dotted lines above and below the dashed line correspond to $\hat{\pi}^h \pm 2s$.

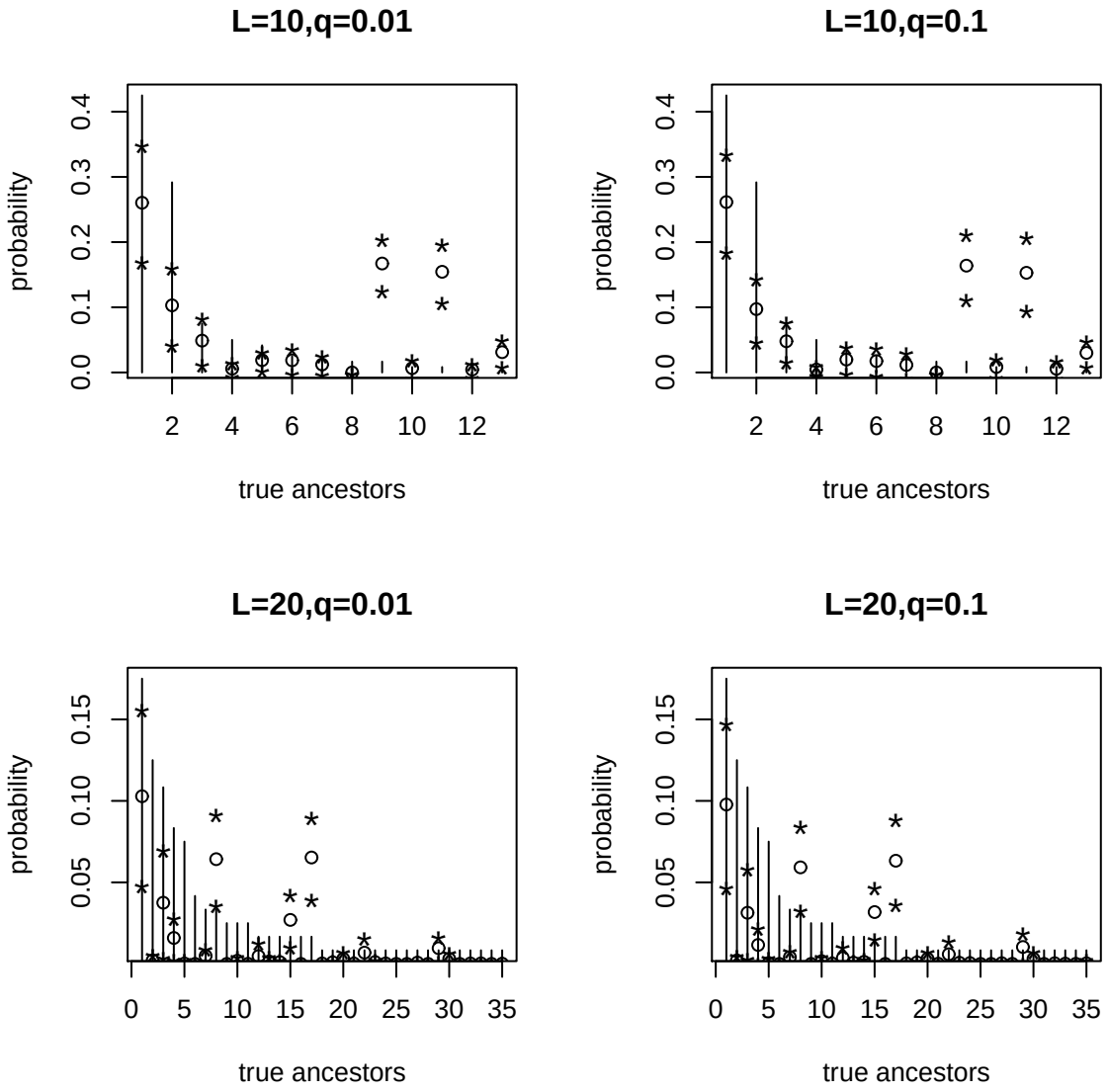


Figure 5.5: Estimated ancestral distribution, $\hat{\pi}^h$, for $L = 10, 20$ and $q = 0.01, 0.1$. Order-3 MCCL is applied. Solid vertical line is the true probability at the true ancestors, drawn by a decreasing order. Dashed line is mean $\hat{\pi}^h$ at the true ancestors. Dotted lines above and below the dashed line correspond to $\hat{\pi}^h \pm 2s$.

Table 5.3: Sum of estimated probabilities, $\hat{\pi}^h(\mu_{MC})$, over the binary sequence, μ_{MC}. Here $L = 10$ and $q = 0.01$.

	$m = 1$	$m = 2$	$m = 3$
$\sum_{\mu_{MC}} \hat{\pi}^h(\mu_{MC})$	0.9970	0.9933	0.9773

Table 5.4: Sum of estimated probabilities $\hat{\pi}^h(\mu)$ over the true ancestors μ.

	$q = 0.01$	$q = 0.05$	$q = 0.1$
$L = 10$	0.8313	0.8344	0.8207
$L = 15$	0.5148	0.4982	0.4733
$L = 20$	0.3563	0.3450	0.3356

neighbor sequences of the true ancestors. Thus, in Table 5.5, we summarize the sum of probabilities at true ancestors and the neighbor sequences with Hamming distance $d_h = 1$ where order-3 MCCL is used to do the reconstruction.

Table 5.5: Sum of estimated probabilities $\hat{\pi}^h$ at the true ancestors μ plus neighbor sequences with Hamming distance as 1.

	$q = 0.01$	$q = 0.05$	$q = 0.1$
$L = 10$	0.9428	0.9450	0.9393
$L = 15$	0.8136	0.8080	0.7970
$L = 20$	0.6525	0.6439	0.6325

5.4.3 Computation time

In Table 5.6, we report the computation time (in seconds) for estimating the marginal ancestral distribution by applying the hierarchical estimator in $N = 100$ replications. The computation was run by using R language on the computational

cluster Lion-XO. Compared with the computation time for the left to right estimator, the hierarchical estimator is faster than the left to right estimator when estimate the pairwise margins, but for the higher order margins, the hierarchical estimator is a little slower.

Table 5.6: Computation time (in seconds) for estimating margins in 100 samples using the hierarchical method.

	time			log_2 time		
	pairwise	threewise	fourwise	pairwise	threewise	fourwise
$L = 10, q = 0.01$	0.14	14.16	48.05	-2.84	3.82	5.59
$q = 0.05$	0.14	14.43	48.92	-2.84	3.85	5.61
$q = 0.1$	0.14	14.50	50.39	-2.84	3.86	5.66
$L = 15, q = 0.01$	0.22	25.27	88.80	-2.18	4.66	6.47
$q = 0.05$	0.24	27.59	96.71	-2.06	4.79	6.60
$q = 0.1$	0.24	26.45	92.93	-2.06	4.73	6.54
$L = 20, q = 0.01$	0.35	32.24	114.50	-1.51	5.01	6.84
$q = 0.05$	0.35	32.89	116.44	-1.51	5.04	6.86
$q = 0.1$	0.35	32.51	115.20	-1.51	5.02	6.85

Chapter 6

Gradient Methods for the Estimated Ancestral Distribution

In chapter 4 and chapter 5, we introduced two estimators which can be used to estimate the $(m+1)$-wise marginal ancestral distribution base on order-m MCCL (4.2). In addition, in the simulation studies, we found that the performance for these two estimators were very good for estimating marginal distributions. However, the estimates obtained from either of two estimators are not the maximum likelihood estimates for the MCCL. They are instead conditional maxima for the MCCL. We next consider how one might find the maximum composite likelihood estimate for MCCL. One could also ask if one could add some improvement steps so that the improved estimates are closer to the maximum composite likelihood estimate. In this chapter, we will introduce one potential method to improve the estimates obtained from two estimators discussed in chapter 4 and chapter 5.

6.1 Simplex gradient method

The method we proposed to use is gradient method. That is, we begin from the estimated joint ancestral distribution obtained by using either the left to right estimator or the hierarchical estimator, then we update the estimates by adding probability mass to the potential ancestral sequences that have positive gradient values and reweighting the probabilities for the other ancestral sequences, until no sequences have positive gradients. By the theorems in the monograph by Lindsay

(1995), if all the gradients are less than zero, we have found the MLE for the ancestral distribution.

Our interest parameters are the probability mass on all 2^L potential ancestors, denoted by $\pi(\mu^1), \ldots, \pi(\mu^{2^L})$, where μ^i is the ith potential ancestral sequence. These probability mass must satisfy

$$0 \leq \pi(\mu^i) \leq 1, \text{ and } \sum_{i=1}^{2^L} \pi(\mu^i) = 1,$$

which implies the parameter space is a simplex. Thus, instead of using a traditional gradient method, we proposed to use the simplex gradient method.

Let $D(\mu^i)$ denote the simplex gradient at ancestral sequence μ^i. Unlike traditional gradient method, which uses the partial derivatives of log likelihood with respect to the vector parameter $\pi(\mu^i)$, the simplex gradient take the derivative of log likelihood with respect to ϵ, where ϵ is the parameter for a path $(1-\epsilon)\pi(\mu^j)+\epsilon\delta_{\mu^j}(\mu^i)$ going from π to $\delta_{\mu^j}(\mu^i)$, where $\delta_{\mu^j}(\mu^i)$ is a degenerate density, $\delta_{\mu^j}(\mu^i) = 1$ when $\mu^i = \mu^j$ and zero otherwise. That is,

$$D(\mu^i) = \frac{d}{d\epsilon} \sum_{j=1}^{2^L} n(\mu^j) l((1-\epsilon)\pi(\mu^j) + \epsilon\delta_{\mu^j}(\mu^i)) \Big|_{\epsilon=0} \tag{6.1}$$

If $D(\mu^i) > 0$, the log likelihood will be increased by adding probability mass on the point μ^i by using some positive ϵ. In addition, since these paths are always within the parameter space, we do not need to take into account the 'sum to one' constraint on $\pi(\mu)$. For example, when we use order-1 MCCL (4.4), we have the simplex gradient for the sequence $\mu = (\mu_1, \ldots, \mu_L)$, where $\mu_i \in \{0, 1\}$, is

$$D(\mu) = \frac{d}{d\epsilon} \sum_{x_1} \cdots \sum_{x_L} n_{x_1,\ldots,x_L} \log \left\{ \frac{f^\epsilon(x_1, x_2) \cdots f^\epsilon(x_{L-1}, x_L)}{f^\epsilon(x_2) \cdots f^\epsilon(x_{L-1})} \right\} \Big|_{\epsilon=0},$$

where

$$\begin{aligned} f^{\epsilon}(x_s, x_{s+1}) &= (1-q)\pi^{\epsilon}_{s,s+1}(x_s, x_{s+1}) + q\pi^{\epsilon}_{s}(x_s)\pi^{\epsilon}_{s+1}(x_{s+1}), \\ f^{\epsilon}(x_s) &= \pi^{\epsilon}_{s}(x_s), \\ \pi^{\epsilon}_{s,s+1}(x_s, x_{s+1}) &= (1-\epsilon)\pi_{s,s+1}(x_s, x_{s+1}) + \epsilon\delta_{x_s,x_{s+1}}(\mu_s, \mu_{s+1}), \\ \pi^{\epsilon}_{s}(x_s) &= (1-\epsilon)\pi_{s}(x_s) + \epsilon\delta_{x_s}(\mu_s). \end{aligned}$$

By simple algebra, we can show that

$$\begin{aligned} & D(\mu) \qquad (6.2) \\ = \; & (1-q)\left[\frac{n(x_1=\mu_1, x_2=\mu_2)}{f(x_1=\mu_1, x_2=\mu_2)} + \frac{n(x_1=\mu_1, x_2=\mu_2)}{f(x_1=\mu_1, x_2=\mu_2)} + \cdots + \frac{n(x_{L-1}=\mu_{L-1}, x_L=\mu_L)}{f(x_{L-1}=\mu_{L-1}, x_L=\mu_L)}\right] \\ - \; & \left[\frac{n(x_2=\mu_2)}{f(x_2=\mu_2)} + \frac{n(x_3=\mu_3)}{f(x_3=\mu_3)} + \cdots + \frac{n(x_{L-1}=\mu_{L-1})}{f(x_{L-1}=\mu_{L-1})}\right] \\ + \; & q\Bigg[\sum_{x_1}\frac{n(x_1, x_2=\mu_2)}{f(x_1, x_2=\mu_2)}\pi_1(x_1) + \sum_{x_2}\frac{n(x_1=\mu_1, x_2)}{f(x_1=\mu_1, x_2)}\pi_2(x_2) \\ & \quad + \sum_{x_2}\frac{n(x_2, x_3=\mu_3)}{f(x_2, x_3=\mu_3)}\pi_2(x_2) + \sum_{x_3}\frac{n(x_2=\mu_2, x_3)}{f(x_2=\mu_2, x_3)}\pi_3(x_3) \\ & \quad + \cdots \\ & \quad + \sum_{x_{L-1}}\frac{n(x_{L-1}, x_L=\mu_L)}{f(x_{L-1}, x_L=\mu_L)}\pi_{L-1}(x_{L-1}) + \sum_{x_L}\frac{n(x_{L-1}=\mu_{L-1}, x_L)}{f(x_{L-1}=\mu_{L-1}, x_L)}\pi_L(x_L) \Bigg] \\ - \; & q\sum_{x_1}\cdots\sum_{x_L} n(x_1, \ldots, n_L)\left[\frac{\pi_1(x_1)\pi_2(x_2)}{f(x_1, x_2)} + \frac{\pi_2(x_2)\pi_3(x_3)}{f(x_2, x_3)} + \cdots + \frac{\pi_{L-1}(x_{L-1})\pi_L(x_L)}{f(x_{L-1}, x_L)}\right] \\ - \; & n, \end{aligned}$$

To update the estimated probability mass for all sequences, we can use the following updating equation

$$\pi^{new}(\mu_i) = \pi^{old}(\mu_i)(1 + \alpha D(\mu_i)), \qquad (6.3)$$

where $\alpha > 0$ is a tuning parameter and $i = 1, \ldots, 2^L$. There are two motivations for this equation. First, if the likelihood had a conventional mixture structure, and one used $\alpha = \frac{1}{n}$, this would be the EM algorithm. Second, if one uses updating equation

(6.3), the new estimates satisfy Lemma 6.1 below, so we know that $\sum_{i=1}^{2^L} \pi^{new}(\mu^i) = \pi^{old}(\mu^i)(1+\alpha D(\mu^i)) = 1$. Thus, as long as α is chosen small enough, and the initial $\pi(\mu^i)$ is positive everywhere, the updated $\pi(\mu^i)$'s are nonnegative and so in the parameter space.

Lemma 6.1.

$$\sum_{i=1}^{2^L} \pi(\mu^i) D(\mu^i) = 0.$$

Proof. This is a general property of a simplex gradient. Under differentiability of the objective function, the path derivative along the path $(1-\epsilon)\pi(\mu^i) + \epsilon g(\mu^i)$, for arbitrary density $g(\mu^i)$, is $\sum_i g(\mu^i) D(\mu^i)$. Clearly, if $g(\mu^i) = \pi(\mu^i)$, this derivative is zero. □

Therefore, beginning from the estimated joint distribution obtained from either the left to right estimator or the hierarchical estimator, we can improve the estimates by the following simplex gradient algorithm.

6.1.1 Simplex gradient algorithm

This algorithm assumes we wish to maximize the MCCL over a fixed support S.

1. Assume we have an initial estimated ancestral distribution $\hat{\pi}(\mu)$ with $\hat{\pi}(\mu) > 0$ on S. If μ is in S, calculate the gradient value $D(\mu)$. If for all μ in the S, we have $D(\mu) \leq tol = 0.005$, then stop the iteration and the $\hat{\pi}(\mu)$ is the final estimation. Otherwise, go to next step.

2. Find the tuning parameter value, α^*, such that $\hat{\pi}(\mu)(1 + \alpha D(\mu))$ for $\mu \in S$ maximizes the desired log likelihood. To find α^*, we can use Newton-Raphson method.

3. Calculate the new π by applying equation (6.3) with the tuning parameter value obtained from step 2, $\pi^{new}(\mu) = \hat{\pi}(\mu)(1 + \alpha^* D(\mu))$, and then go back to step 1 to check the new gradients.

To illustrate the idea of this simplex gradient algorithm, we did a simple example. We generated a simple data set from a known ancestral distribution, with $L = 5$, $n = 100$, and $q = 0.1$. We first estimated the pairwise margins by using both the left to right estimator, beginning from the left, and the hierarchical estimator. We then reconstructed the ancestral distribution by using simulated ancestors method. Denote the reconstructed joint distribution as $\hat{\pi}^l$ and $\hat{\pi}^h$ for two estimators respectively. We used π^l and π^h as the initial value and applied the above simplex gradient method to improve the estimated ancestral distribution.

In the example, we only check the gradients at the initial estimated support points. This is because we use equation (6.3) to update the estimated distribution. The probability masses of those potential ancestral sequences, which have no positive probability masses at the beginning, are always zeroes during the updating process.

The computation results show that by using the simplex gradient algorithm, the log likelihood was improved by 0.0038 from -198.0394 to -198.0356 for $\hat{\pi}^l$, and by 0.0058 from -198.0414 to -198.0356 for $\hat{\pi}^h$. Figure 6.1 show the estimates moved a little bit after the gradient improvement and all the gradients at the estimated ancestors were less than the tolerance. This example verified that both of our previous estimator can be very efficient at the maximizing the MCCL.

6.1.2 Discussion on simplex gradient method

Though we can improve our estimated ancestral distribution though the above simplex gradient method. However, there are several cons for this method.

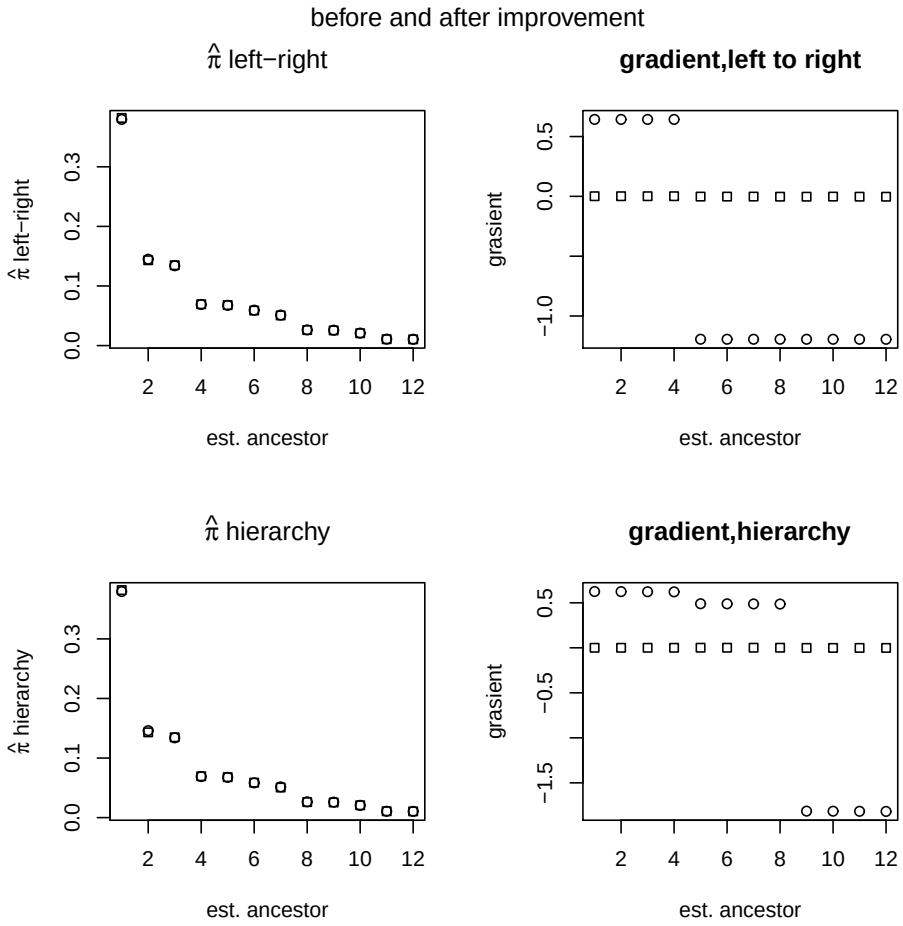


Figure 6.1: Simplex gradient improvement. Circles correspond to before improvement, and squares correspond to after improvement. Here $n = 100$, $q = 0.1$

Computation challenge

The first problem is the computation issue. Though, for the above toy example, the gradients on all estimated support points decrease below to the tolerance just after a few steps of iteration, this iteration procedure works very slowly even for a very short sequence length.

In practice, though by using the simplex gradient method, the log likelihood can be improved greatly in the first several steps of iterations, the algorithm then just adjusts the distribution of the probability masses to the support points

with very little improvement on the likelihood. Meanwhile, the value for tuning parameter α tends to be very small, which means the adjustment on the probability masses and gradients are very limited within each iteration. Thus, the positive gradients decrease even more slowly.

Possible missing support points

Another con for this simplex gradient algorithm is that we may fail to identify some support points for the distribution. When we use (6.3) to update the estimates, the estimated support points would stay the same as the initial estimates. Thus, we may miss those potential support points which have zero probability masses but should appear in a global solution.

6.2 Gradient method with neighbor searching

To solving the possible missing support points problem for the simplex gradient method discussed in previous section, we may calculate the gradients for all the potential points and then add those points with positive gradients back to the support points class. However, for our problem this universal search is very time consuming when L is larger, because we have 2^L potential ancestral sequences. One possible solution is that instead of search universally, we only search locally. That is, within each iteration, we only search the neighbor sequences of current estimated support points, and add the sequences with positive gradients but zero probabilities into the set of estimated support points. Then, we update the estimates based on this new support points set.

The neighbor sequences of a particular sequence, say μ, can be defined by using the concept of Hamming distance, which is defined as the number of positions where two strings a and b have different elements. For example, all the sequences have hamming distance 1 with μ are just the sequences that only have one different

element compared with μ, or, by switching only one symbol of μ at each time.

To using the neighbor search method, we also need to do the modification for updating equation (6.3), because this equation doesn't change the probability mass for the new added points. The new updating method could be

$$\pi^{new}(\mu) = \begin{cases} \pi^{old}(\mu) + \alpha D(\mu) & \text{if } \mu \in C_1 \\ \pi^{old}(\mu) + \beta D(\mu) & \text{if } \mu \in C_2 \end{cases} \tag{6.4}$$

where C_1 denote the class of sequences that have negative gradients, C_2 is the class of sequences with positive gradients, and α, β are two tuning parameters. Thus, the current estimated supports could be in both classes, while the new added potential support points are just in class C_2. As long as the tuning parameter $\beta \geq 0$, the new estimated probability masses for these new added potential supports are non-negative.

In order to make equation (6.4) an eligible updating equation for our problem, the updated $\pi^{new}(\mu)$ must satisfy the constraint $\sum_\mu \pi^{new}(\mu) = 1$. This restriction can be achieved by carefully choosing the values of α and β such that $\alpha > 0$, $\beta > 0$, and $\alpha \sum_{\mu \in C_1} D(\mu) + \beta \sum_{\mu \in C_2} D(\mu) = 0$, which implies

$$\beta = -\alpha \frac{\sum_{\mu \in C_1} D(\mu)}{\sum_{\mu \in C_2} D(\mu)}. \tag{6.5}$$

This formula can be plugged into (6.4) to obtain a path depending only on α

In addition, $\pi^{new}(\mu)$ also need to satisfy $0 \leq \pi^{new}(\mu) \leq 1$ for all the μ in the support. For $\mu \in C_1$, the gradient $D(\mu) < 0$, so $\pi^{new}(\mu) = \pi^{old}(\mu) + \alpha D(\mu) < \pi^{old}(\mu) \leq 1$. To solve inequality $\pi^{old}(\mu) + \alpha D(\mu) \geq 0$, we have $\alpha \leq -\pi^{old}(\phi)/D(\mu)$. Thus, for $\mu \in C_1$, α should be

$$\alpha \in \left[0, -\frac{\pi^{old}(\mu)}{D(\mu)}\right]. \tag{6.6}$$

By similar discussion, for $\mu \in C_2$, in order to satisfy the restriction $0 \leq \pi^{new}(\mu) \leq 1$, we must have

$$\beta \leq (1 - \pi^{old}(\mu))/D(\mu). \tag{6.7}$$

Combine equations (6.5), (6.6), and (6.7) if α is selected from interval

$$0 \leq \alpha \leq \min \left\{ \min_{\mu \in C_1} \left(-\frac{\pi^{old}(\mu)}{D(\mu)} \right), \min_{\mu \in C_2} \left(-\frac{(1-\pi^{old}(\mu))}{D(\mu)} \cdot \frac{\sum_{\mu \in C_2} D(\mu)}{\sum_{\mu \in C_1} D(\mu)} \right) \right\}, \quad (6.8)$$

the equation (6.4), with β given by (6.5), is an eligible updating equation for the modified gradient method with neighbor searching. Therefore, the modified gradient algorithm is

1. Given an initial estimated ancestral distribution $\hat{\pi}(\mu)$, where μ is an estimated ancestral sequence, i.e. $\mu \in S$. Calculate the gradient value $D(\mu)$. If we have $D(\mu) \leq tol = 0.005$ for all supports μ in S and their unique neighbors, then stop the iteration and the $\hat{\pi}(\mu)$ is the final estimation. Otherwise, go to next step.

2. Add the neighbor sequences that have positive gradients into the set of supports and call this new support set S^{new}.

3. Find the tuning parameter values α^* and β^*, where β^* is given by (6.5), with the restriction (6.8), such that π^{new} maximizes the desired log likelihood for $\mu \in S^{new}$, where π^{new} is obtained by equation (6.4).

4. Calculate the new π by applying equation (6.4) with the tuning parameters obtained from step 3, and then go back to step 1 to check the new gradients.

Of course, one of the biggest challenge for this modified gradient method is again the speed of computation.

Remark: In more standard mixture problems, this algorithm provides a simple extension of the EM algorithm that might be useful when the set of positive support is unknown. We intend to investigate this separately.

Chapter 7

Future Work

In previous three chapters, we discussed applying the MCCL to the recombination model. There are several research topics need further explore.

7.1 Performance of MCCL

We proposed to use MCCL to estimate both the marginal and the joint ancestral distributions. As illustrated in chapter 2, by using CL method instead of full likelihood, we can reduce the computation complexity but lose some efficiency. So, an interesting question is that compared with the MLE from full likelihood, what's the performance of our estimates obtained from MCCL.

Because our estimates, derived either from the left to right estimator or the hierarchical estimator, are just conditional maxima, not the maximum likelihood estimates for the MCCL, so to address the above question, we may need to begin from the study of the relative efficiency of our estimates compared with the MCCLE, and then compare the efficiency for MCCLE with the MLE derived from the full likelihood. In addition, because of the complex structures for both the MCCL and the full likelihood, to find the MLE, we need to limit the length L and apply some alternative computation methods, such as Monte Carlo likelihood (Thompson (1994)).

7.2 Model with both mutation and recombination

In the thesis, we only discussed applying MCCL to the model with recombination. Since our motivating problem is the inference for model with both mutation and recombination, the next step is to add the mutation factor back into the model and investigate the performance of MCCL when adding the mutation factor.

7.3 Hidden Markov chain model

In genetics, people are also interested in find the location of 'hotspots' in the genome, i.e. small regions in the genome where recombination occur more frequently. Our proposed statistical model with both mutation and recombination in this thesis can be used to address this question with some modification.

Suppose we observe the current descendent binary sequence X, and let recombination variable H be the hidden variable. Then, the complete data will be $(X, H) = (X_1, H_{1,2}, X_2, H_{2,3}, \ldots, H_{L-1,L}, X_L)$, where $H_{i,i+1} = 1$ if there is recombination between site i and site $i+1$ and 0 otherwise. Given the probability of recombination, q, fixed, the distribution of $H_{i,i+1}$ is just independent bernoulli trial with success probability as q for $i = 1, \ldots, L-1$. In addition, the conditional distribution of X given H can be modeled by our proposed model with mutation and recombination. If we assume that the ancestral distribution is a Markov chain, similar to what was done in this thesis, then the likelihood will have a hidden Markov chain structure. Hence, we can use the popular algorithms in the hidden Markov chain model, such as the forward and backward algorithms (Ewens and Grant (2005)), to estimate the ancestral distribution and the location of recombination as well.

APPENDICES

Appendix A

Proof for Lemmas in Chapter 5

Lemma A.1. *Functions* $\frac{d}{d\phi_s(i)}l(\phi_s(i))$ *and* $\frac{d^2}{d\phi_s(i)^2}l(\phi_s(i))$ *have no definitions when*

$$\phi_s(i) = -\frac{b^s_{00}}{a^s_{00}}, -\frac{b^s_{01}}{a^s_{01}}, -\frac{b^s_{10}}{a^s_{10}}, \text{ or } -\frac{b^s_{11}}{a^s_{11}}.$$

Proof. It is easy to see that when $\phi_s(i)$ take any one of the above four values, we have one of the $f(x_s, i, x_{s+1}) = a^s_{x_s,x_{s+1}}\phi_s(i) + b^s_{x_s x_{s+1}} = 0$. Hence the denominators of $\frac{d}{d\phi_s(i)}l(\phi_s(i))$ and $\frac{d^2}{d\phi_s(i)^2}l(\phi_s(i))$ are zeroes, which implies that they have no definitions at above four points. □

Lemma A.2. *The cubic equation* (5.15) *has no real root in the intervals*

$$\left(-\infty, \min\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\right) \text{ and } \left(\max\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}, \infty\right).$$

Proof. Because of Lemma 5.3, in the open interval $\left(-\infty, \min\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\right)$, both $\frac{d}{d\phi_s(i)}l(\phi_s(i))$ and $\frac{d^2}{d\phi_s(i)^2}l(\phi_s(i))$ exist. In addition, because $\frac{d^2}{d\phi_s(i)^2}l(\phi_s(i)) \leq 0$, $\frac{d}{d\phi_s(i)}l(\phi_s(i))$ is a decreasing function of $\phi_s(i)$. Moreover, the numerator of equation (5.14) is a cubic function of $\phi_s(i)$, and the denominator of equation (5.14) is a function of $[\phi_s(i)]^4$. Hence, $\frac{d}{d\phi_s(i)}l(\phi_s(i)) \to 0$ as $\phi_s(i) \to -\infty$. Thus, $\frac{d}{d\phi_s(i)}l(\phi_s(i))$ cannot achieve zero when $\phi_s(i)$ is some finite value in the open interval $\left(-\infty, \min\{\phi^{(1)}_s(i), \phi^{(4)}_s(i)\right)$, which implies there is no solution for score equation $\frac{d}{d\phi_s(i)}l(\phi_s(i)) = 0$ and hence the cubic equation (5.15) in this interval.

Similarly, we can prove that there could not be real root for the cubic equation (5.15) in the interval $\left(\max\{\phi^{(2)}_s(i), \phi^{(3)}_s(i)\}, \infty\right)$. □

Lemma A.3. *There is at most one real solution for the cubic equation* (5.15) *when* $\phi_s(i)$ *belongs to the open intervals* $\Big(\min\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}\Big)$, $\Big(\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$, *and* $\Big(\min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}, \max\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$.

Proof. From Lemma 5.3, in the open interval $\Big(\min\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}\Big)$, both $\frac{d}{d\phi_s(i)} l(\phi_s(i))$ and $\frac{d^2}{d\phi_s(i)^2} l(\phi_s(i))$ exist. In addition, because $\frac{d^2}{d\phi_s(i)^2} l(\phi_s(i)) \leq 0$, which implies $\frac{d}{d\phi_s(i)} l(\phi_s(i))$ is a decreasing function of $\phi_s(i)$ in this interval. Thus, there is at most one solution for $\frac{d}{d\phi_s(i)} l(\phi_s(i)) = 0$. Hence, the cubic function (5.15) has at most one real root in this open interval.

By same arguments, we can prove that there is at most one real root for the cubic function (5.15) in the open intervals $\Big(\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$ and $\Big(\min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}, \max\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$. □

Lemma A.4. *The parameter space for* $\phi_s(i)$, *i.e. interval* (5.11), *is a sub set of* $\Big[\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big]$.

Proof. Define

$$\Phi_1 \triangleq \max\{0, \hat{\pi}_{s,s+1}(0,i) + \hat{\pi}_{s+1,s+2}(i,0) - \hat{\pi}_{s+1}(i)\}, \tag{A.1}$$

$$\Phi_2 \triangleq \min\{\hat{\pi}_{s,s+1}(0,i), \hat{\pi}_{s+1,s+2}(i,0)\}. \tag{A.2}$$

The sample space for $\phi_s(i)$ is $\phi_s(i) \in \Phi = [\Phi_1, \Phi_2]$. Since $\Phi_1 \geq 0$, $\phi_s^{(1)}(i) \leq 0$, and $\phi_s^{(4)}(i) \leq 0$, we have $\Phi_1 \geq \max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}$.

In addition, for any $\phi_s(i) \in (\Phi_1, \Phi_2)$, we must have $\pi_{s,s+1,s+2}(x_s, i, x_{s+2}) \in (0,1)$, and hence $f(x_s, i, x_{s+1}) > 0$. Thus, $\phi_s^{(k)}(i) \notin (\Phi_1, \Phi_2)$ which implies $\Phi_2 \leq \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}$. Therefore, we prove that

$$\Phi \subseteq \left[\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\right].$$

□

Theorem A.5. *If there exist three real roots for cubic function* (5.15)*, there is only one real root is the eligible candidate estimate for* $\phi_s(i)$*, such that* $f(x_s, x_{s+1} = i, x_{s+2}) \geq 0$ *for* $x_s \in \{0,1\}$ *and* $x_{s+1} \in \{0,1\}$.

Proof. From Lemma 5.4 and Lemma 5.5, we know that if there are three real roots for cubic function (5.15), then these three roots must be in the intervals $\Big(\min\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\},\ \max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}\Big)$, $\Big(\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$, and $\Big(\min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\},\ \max\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$, with each real root in one interval.

When $\phi_s(i) \in \Big(\min\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}\Big)$, we must have one of the $f(0,i,0) = a_{00}^s \phi_s(i) + b_{00}^s$ and $f(1,i,1) = a_{11}^s \phi_s(i) + b_{11}^s$ is negative, because $f(0,i,0)$ and $f(1,i,1)$ are increasing functions with respect to $\phi_s(i)$, and they are achieve zeroes when $\phi_s(i) = \phi_s^{(1)}(i)$ and $\phi_s(i) = \phi_s^{(4)}(i)$ respectively. Thus, for all the $\phi_s(i) \in \Big(\min\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}\Big)$, including the real root for equation (5.15) that falls in this interval, is not an eligible estimate for $\phi_s(i)$.

By similar arguments, we can prove that the real root in the interval $\Big(\min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\},\ \max\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$ is also not an eligible estimate, because one of the $f(0,i,1) = a_{01}^s \phi_s(i) + b_{01}^s$ and $f(1,i,0) = a_{10}^s \phi_s(i) + b_{10}^s$ is negative for $\phi_s(i)$ in this interval. Moreover, only the real root in the interval $\Big(\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}$, $\min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$ can make all the observations probability is non-negative, i.e. $f(x_s, x_{s+1} = i, x_{s+2}) \geq 0$ for $x_s \in \{0,1\}$ and $x_{s+1} \in \{0,1\}$.

Therefore, if there exist three real roots for the cubic equation (5.15), there is only one root is eligible estimate in the sense of $f(x_s, i, x_{s+1}) \geq 0$ for all $x_s, x_{s+1} \in \{0,1\}$, and this root is in the interval $\Big(\max\{\phi_s^{(1)}(i), \phi_s^{(4)}(i)\}, \min\{\phi_s^{(2)}(i), \phi_s^{(3)}(i)\}\Big)$.

□

Appendix B

Additional Figures for Chapter 4

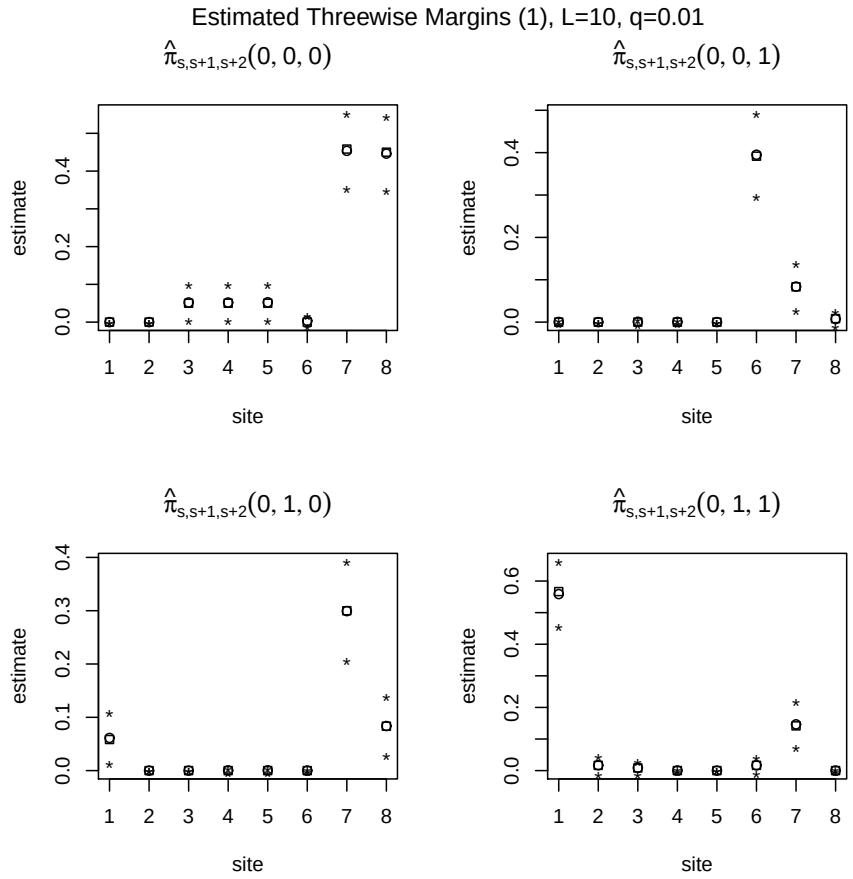


Figure B.1: Estimated Threewise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

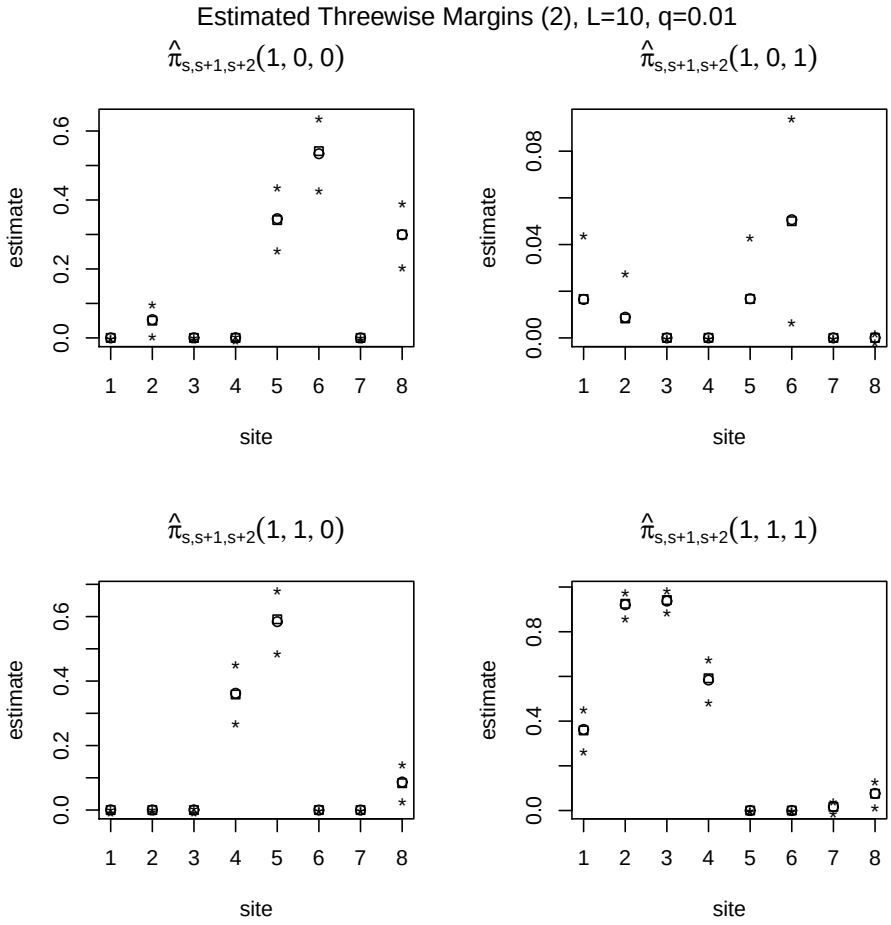


Figure B.2: Estimated Threewise Margins (2). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

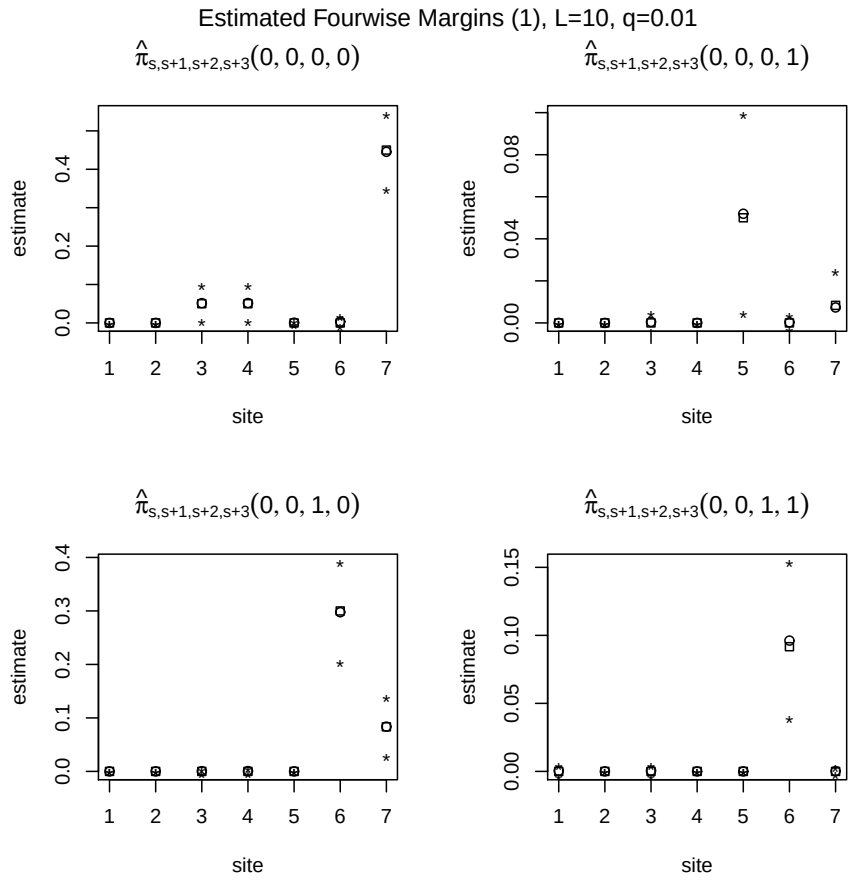


Figure B.3: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

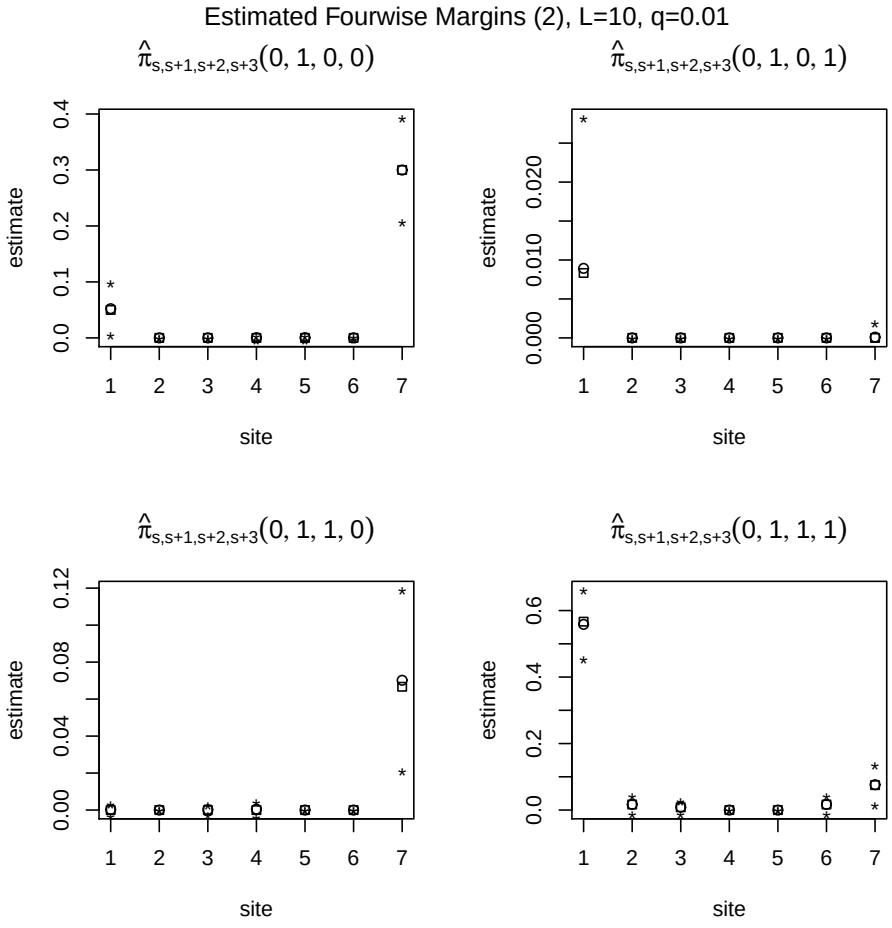


Figure B.4: Estimated Fourwise Margins (2). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

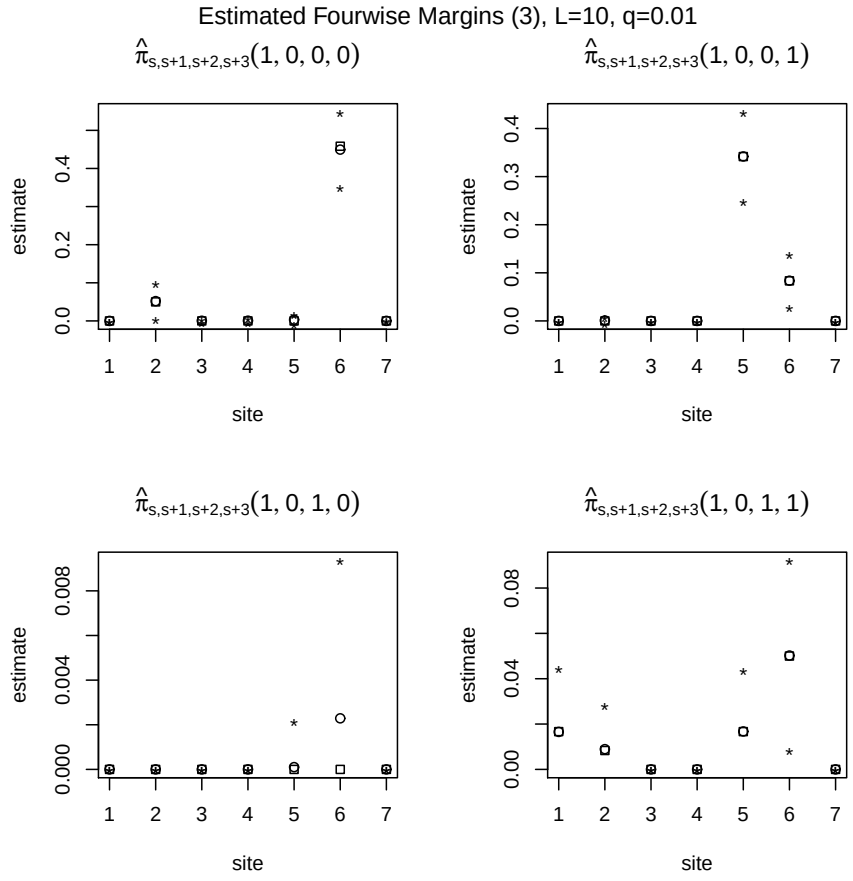


Figure B.5: Estimated Fourwise Margins (3). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

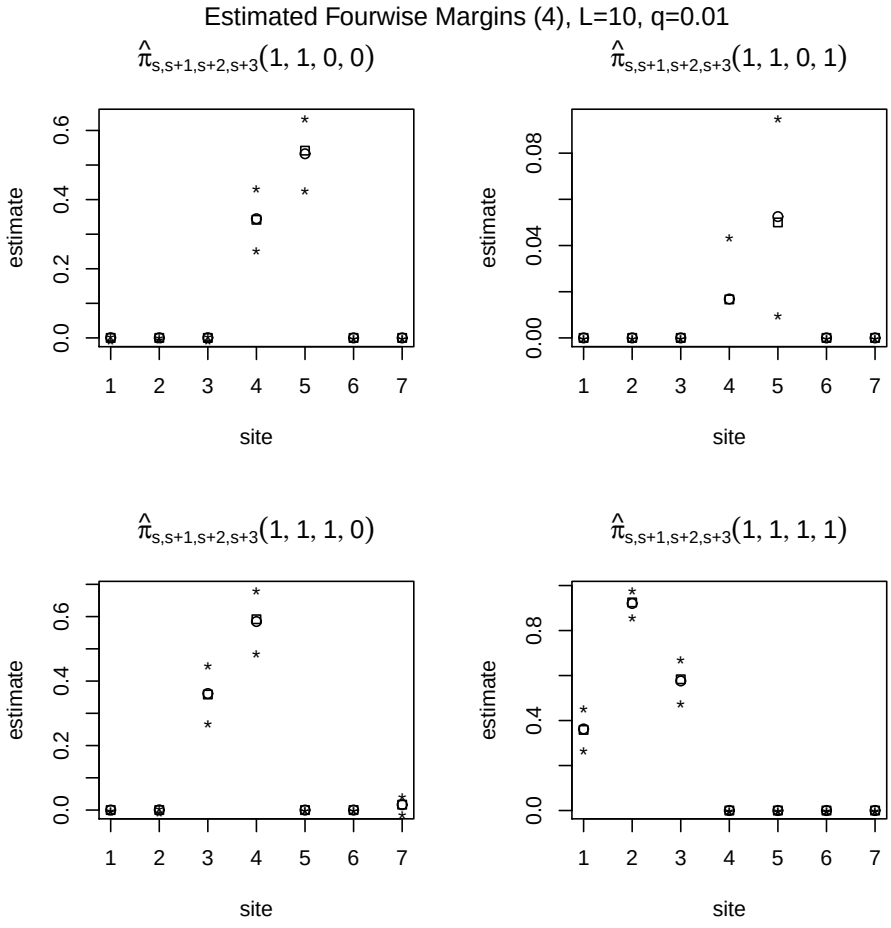


Figure B.6: Estimated Fourwise Margins (4). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

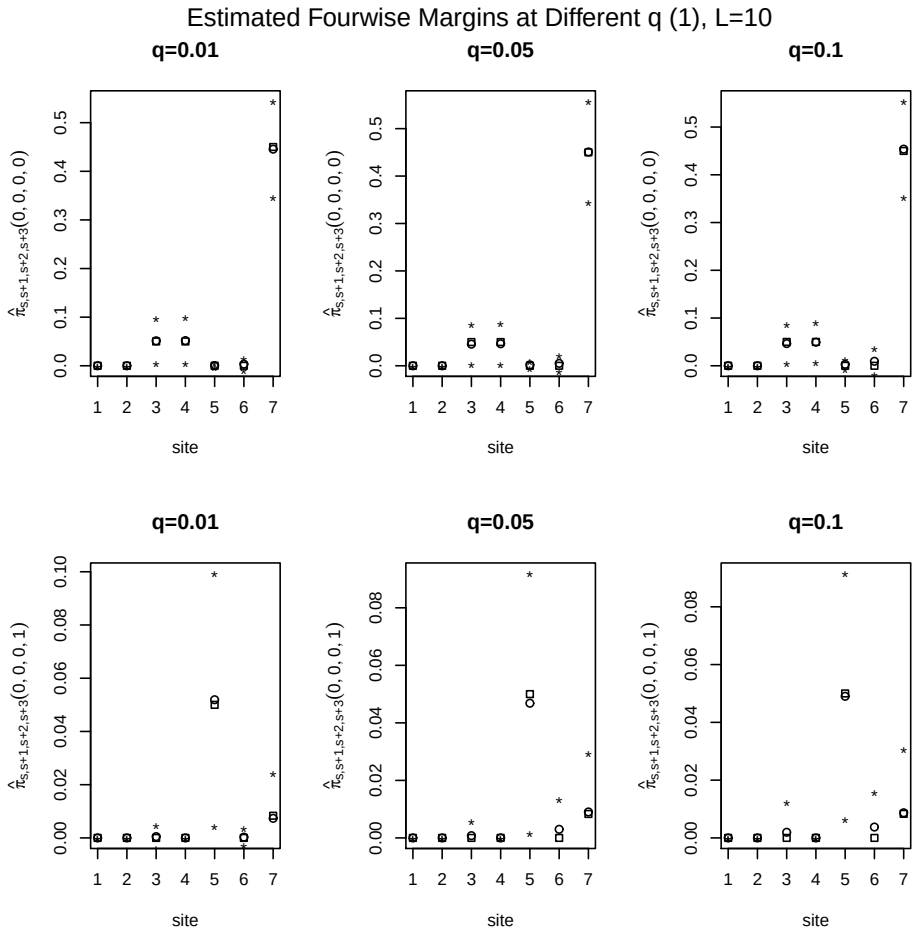


Figure B.7: Estimated Fourwise Margins at Different q (1). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

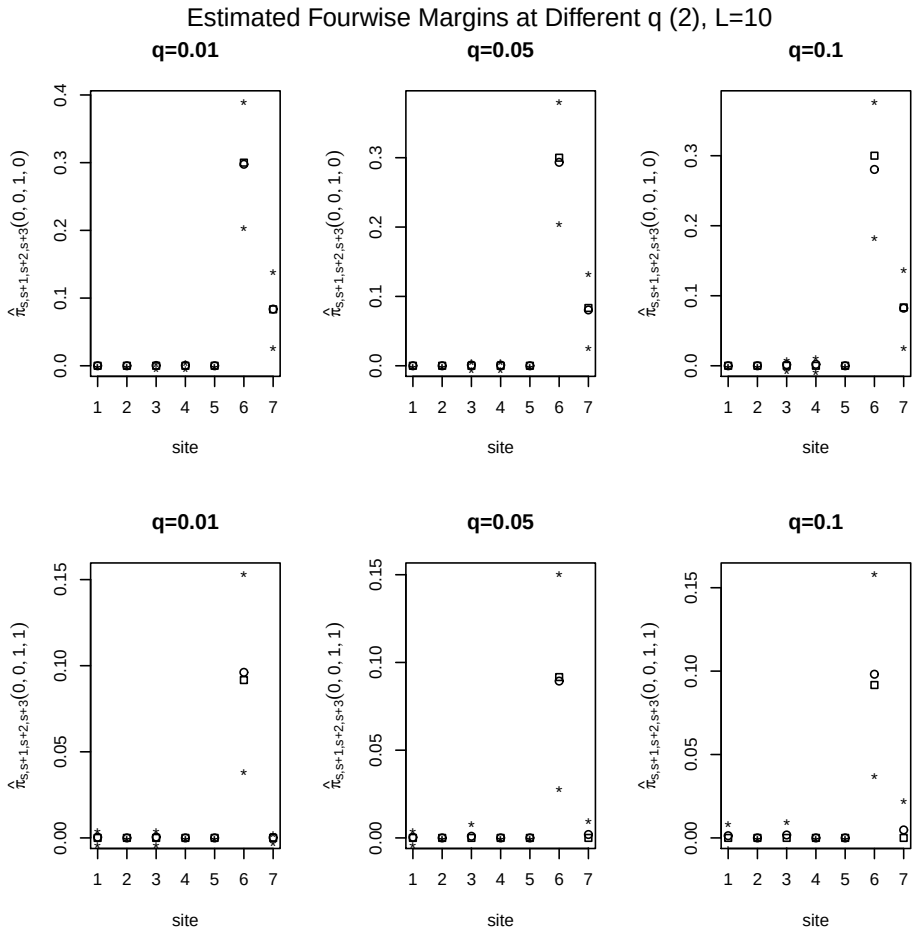


Figure B.8: Estimated Fourwise Margins at Different q (2). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

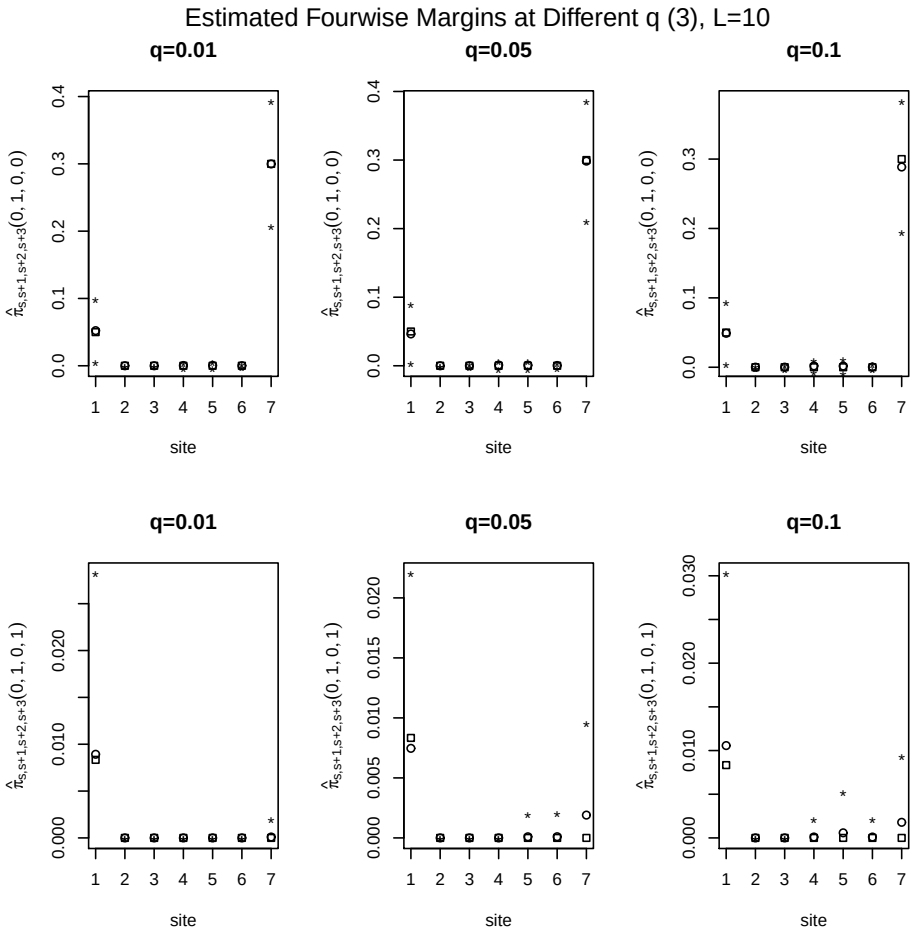


Figure B.9: Estimated Fourwise Margins at Different q (3). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

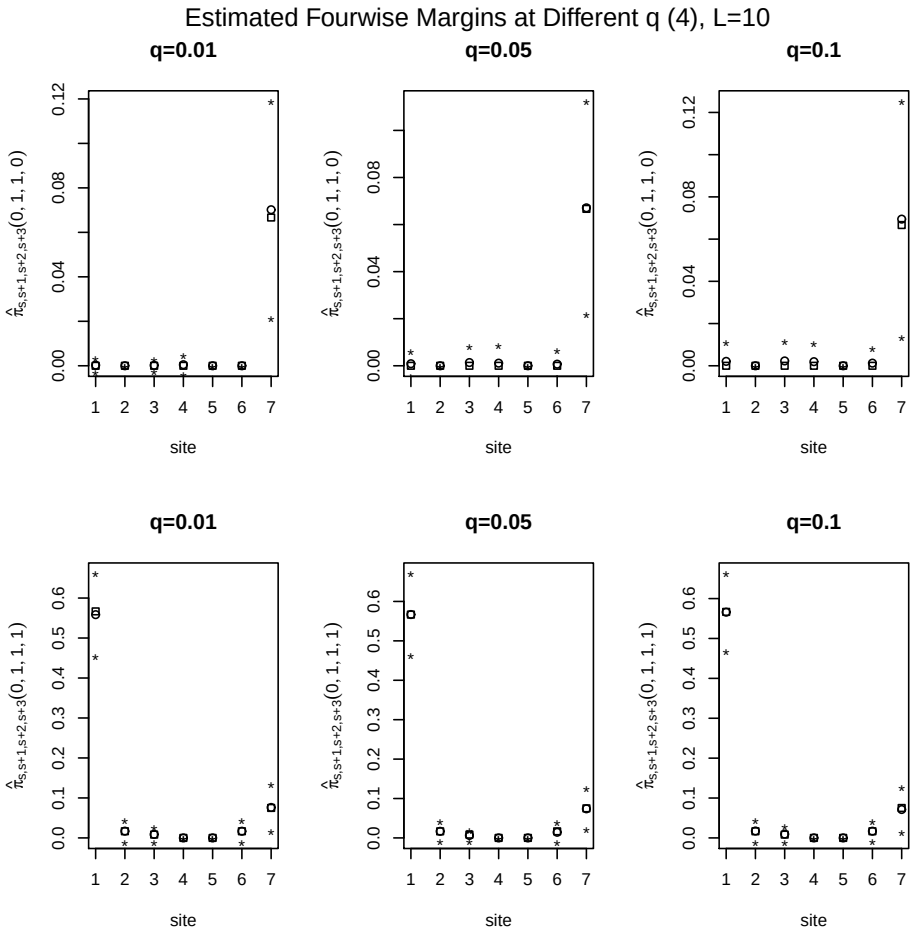


Figure B.10: Estimated Fourwise Margins at Different q (4). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

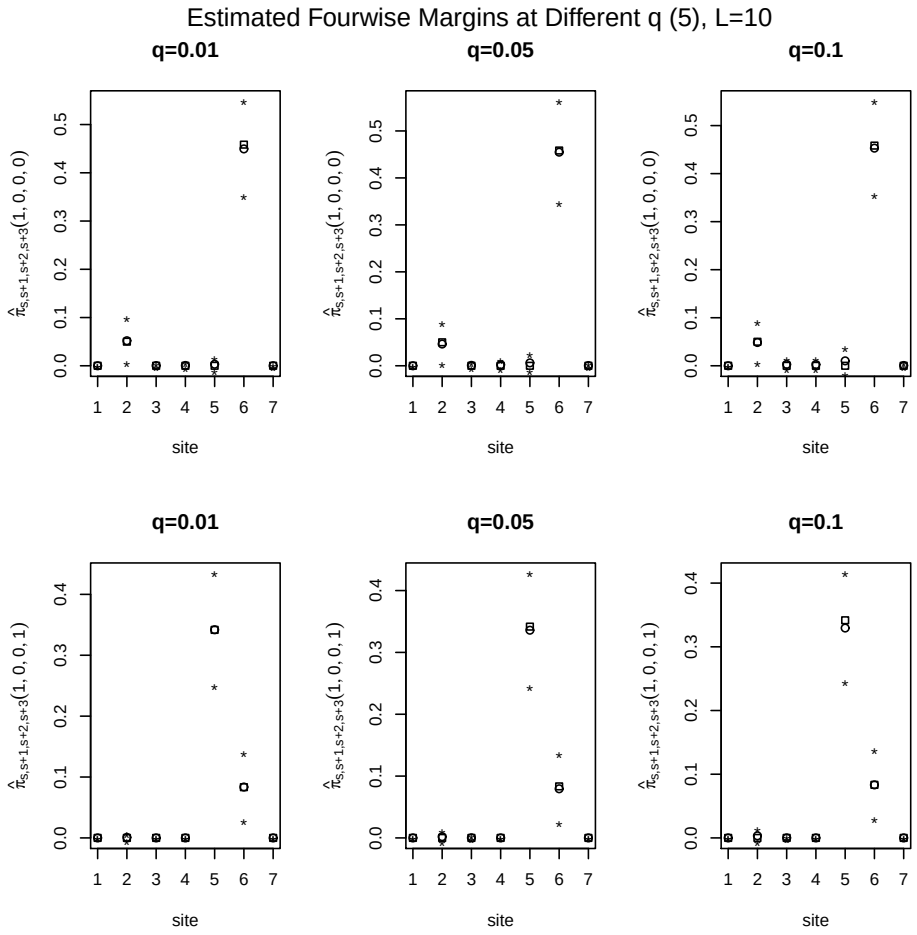


Figure B.11: Estimated Fourwise Margins at Different q (5). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

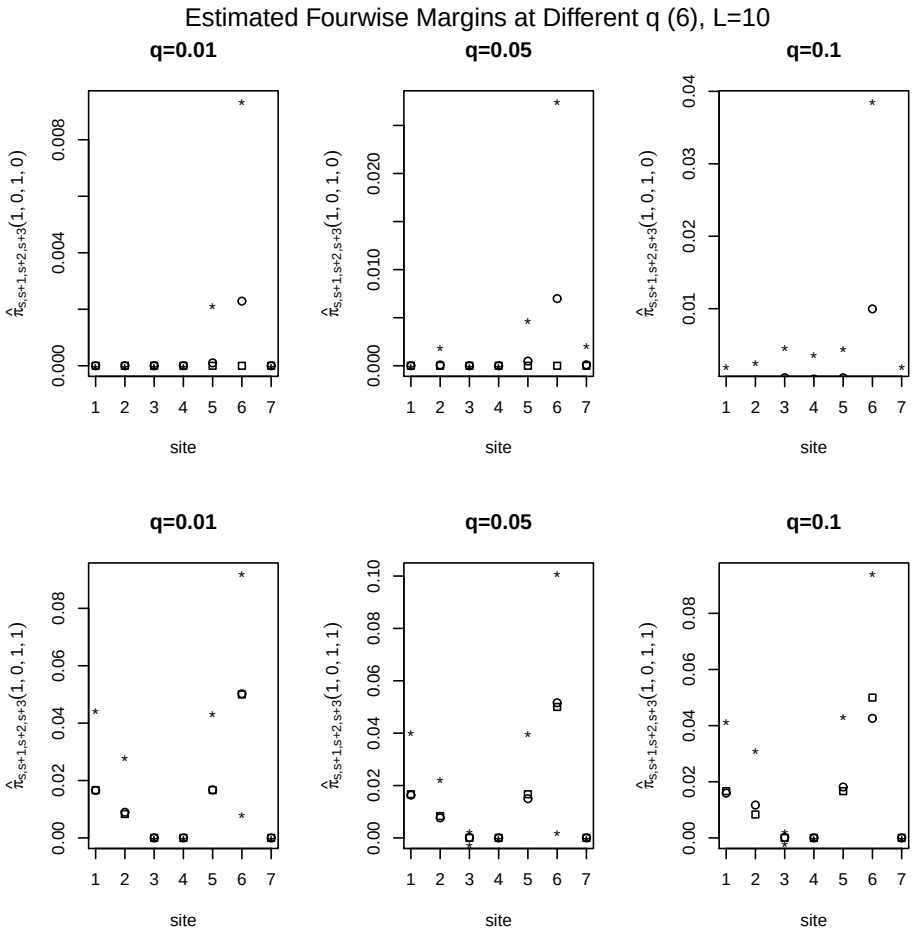


Figure B.12: Estimated Fourwise Margins at Different q (6). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

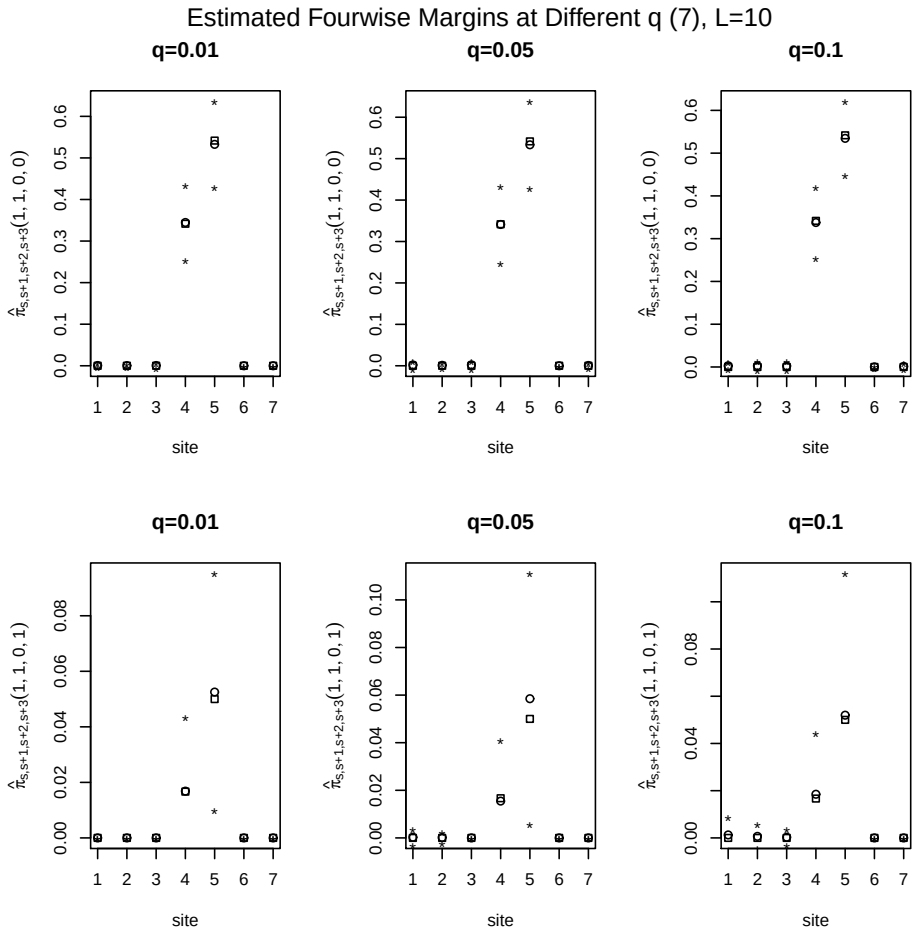


Figure B.13: Estimated Fourwise Margins at Different q (7). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

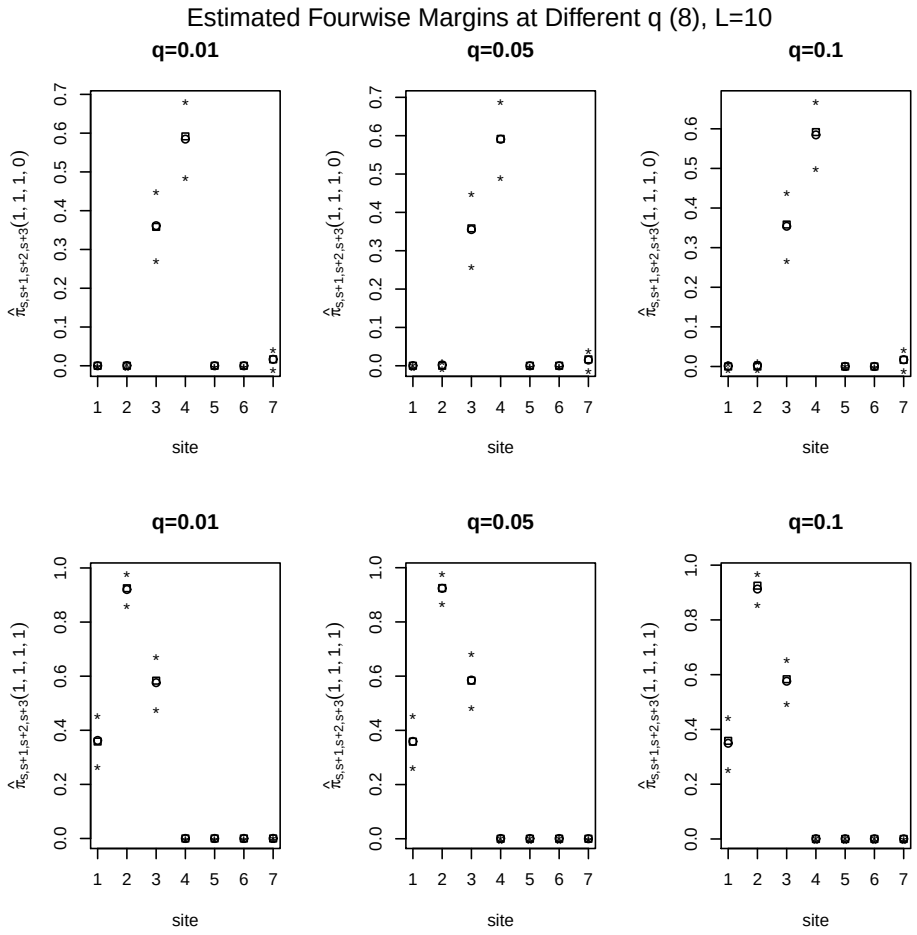


Figure B.14: Estimated Fourwise Margins at Different q (8). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

Appendix C

Additional Figures for Chapter 5

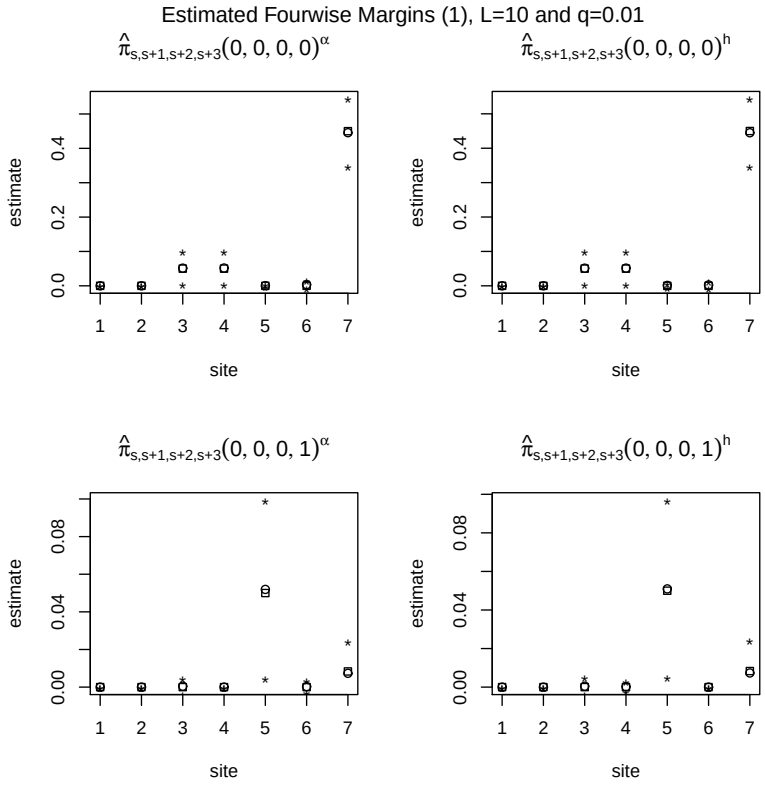


Figure C.1: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

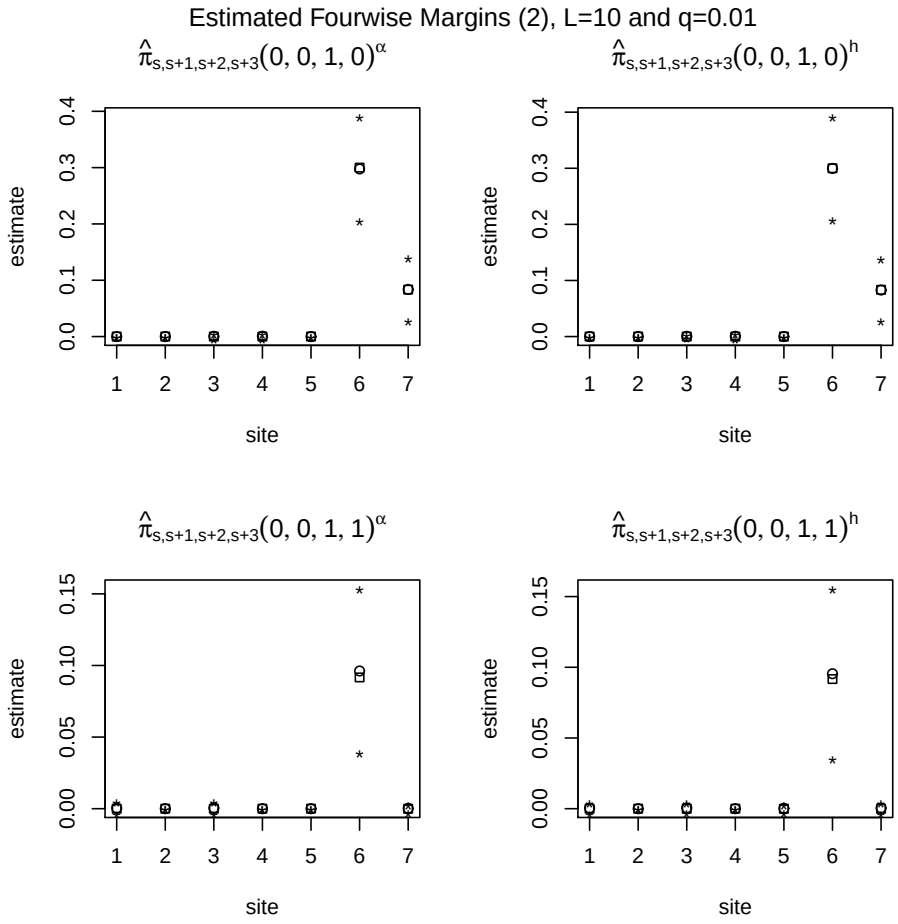


Figure C.2: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

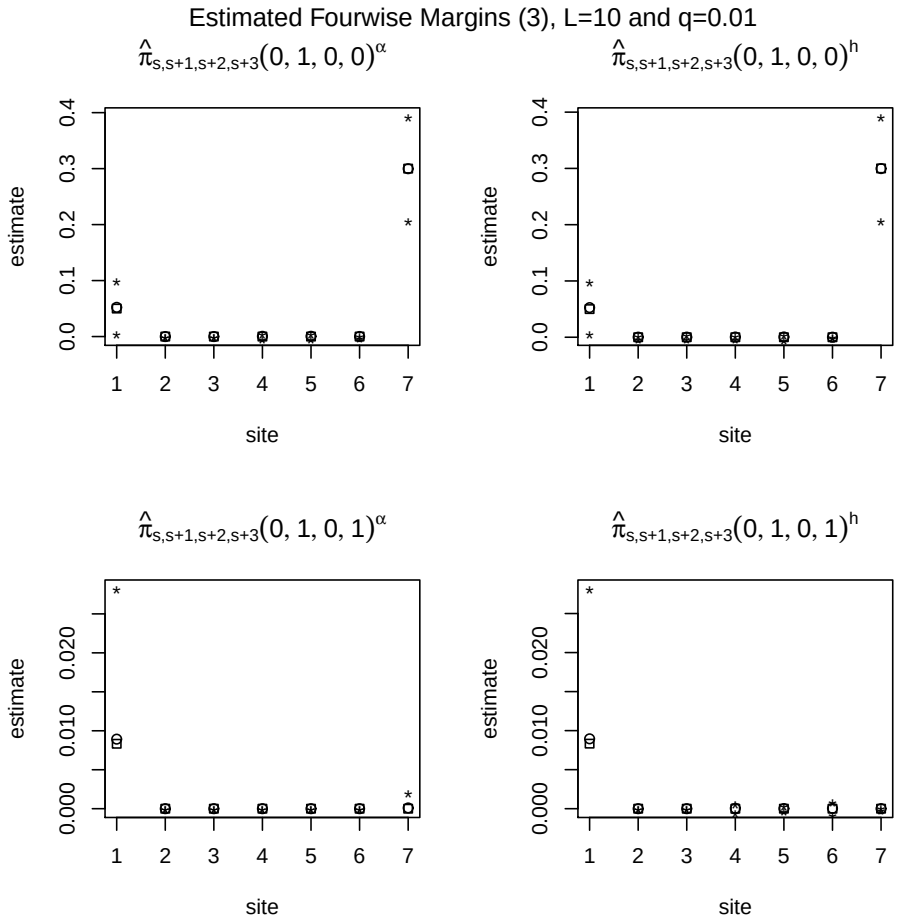


Figure C.3: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

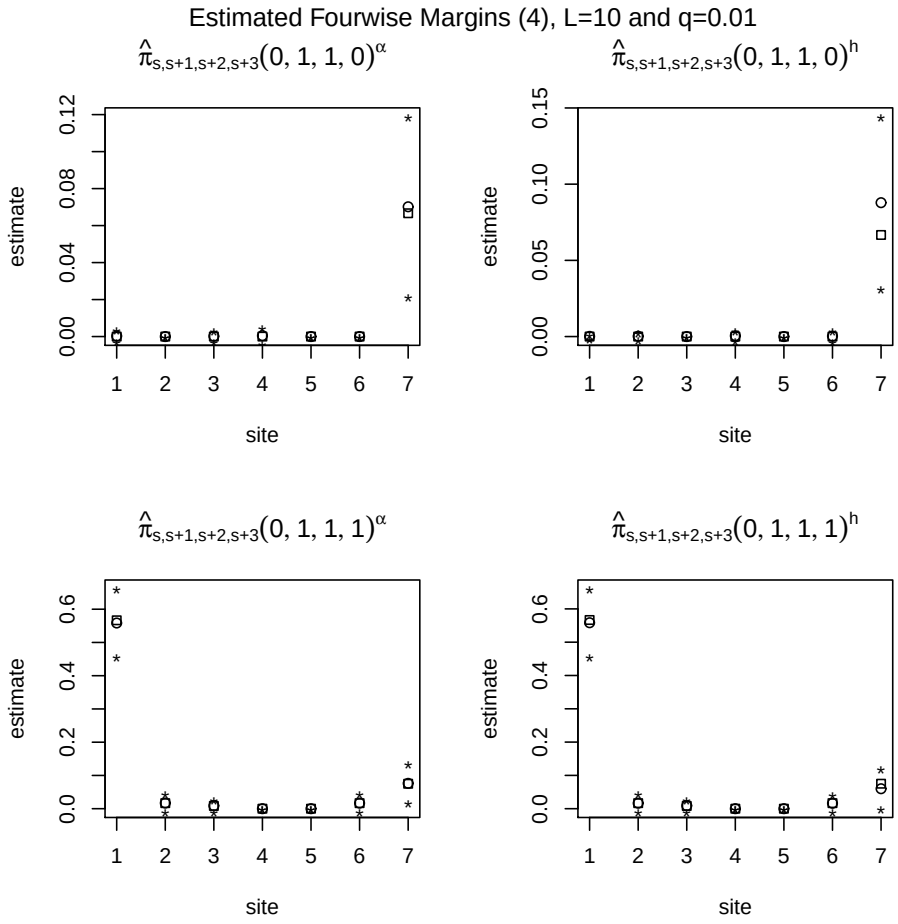


Figure C.4: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

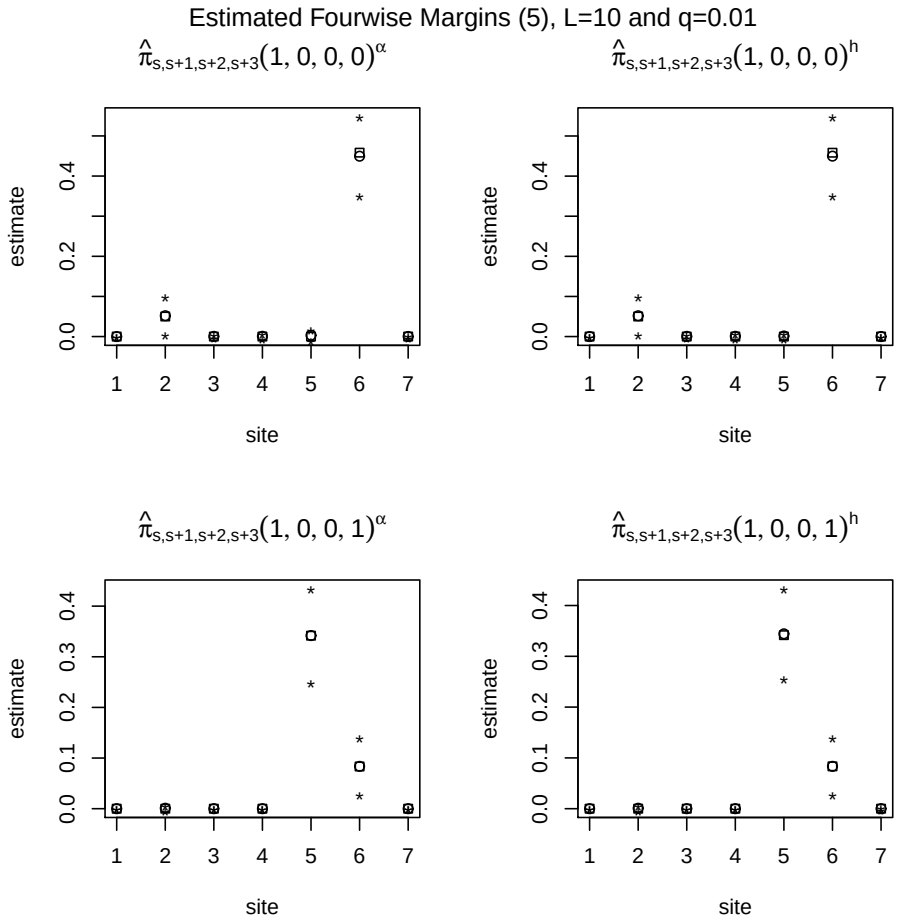


Figure C.5: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

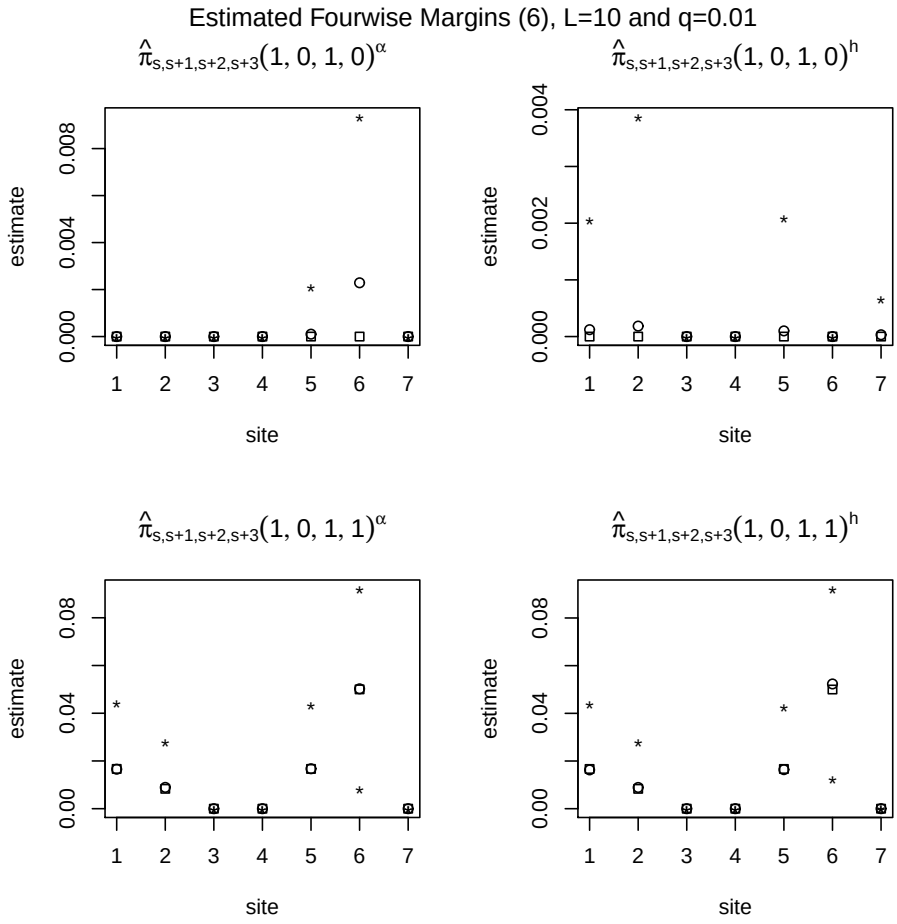


Figure C.6: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

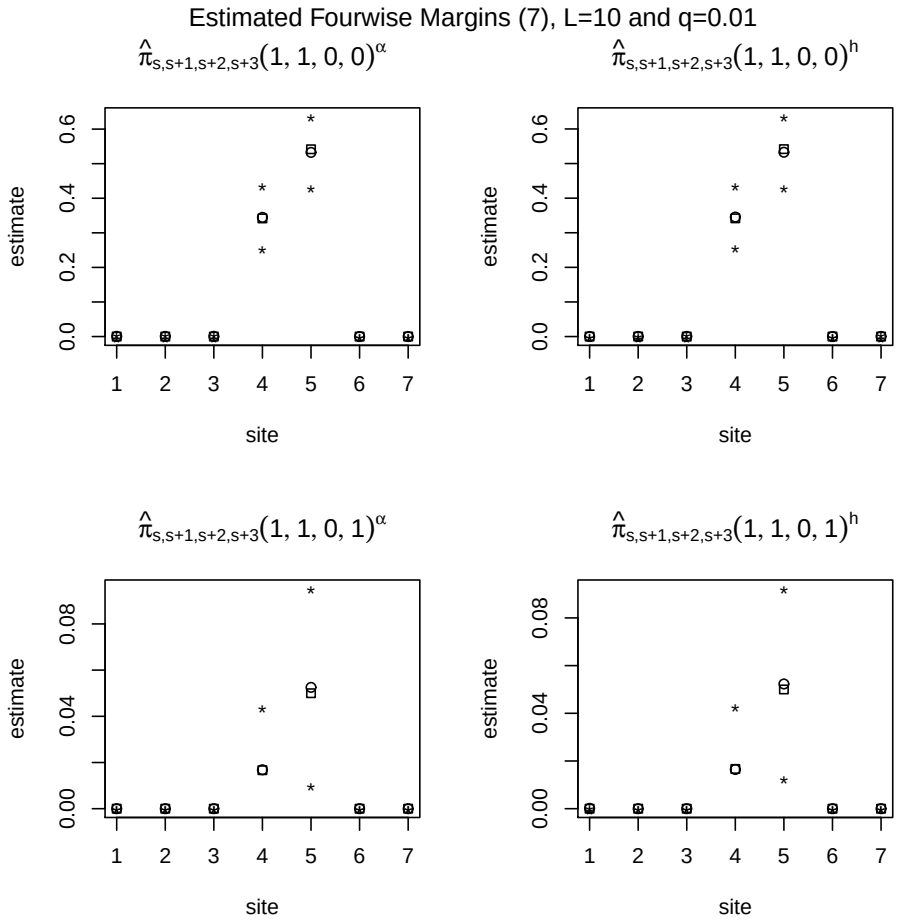


Figure C.7: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

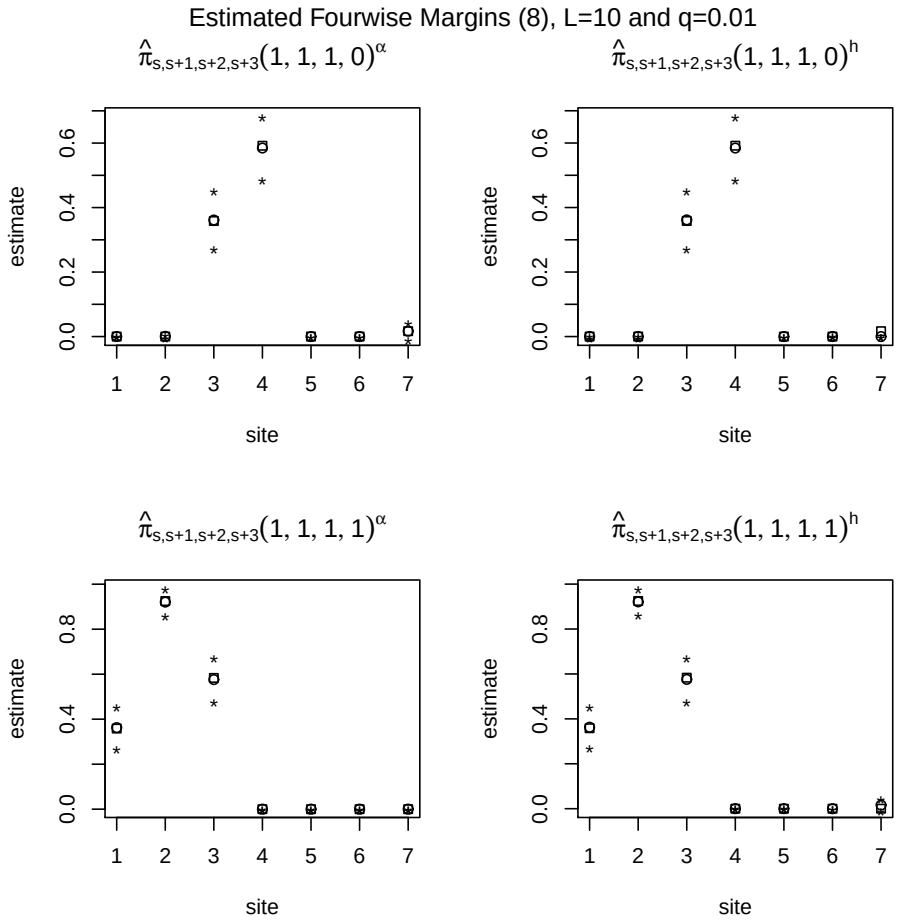


Figure C.8: Estimated Fourwise Margins (1). Here $L = 10$, $q = 0.01$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error. $\hat{\pi}^{\alpha}$ is the estimate obtained by using the left to right estimator and $\hat{\pi}^{h}$ is the estimate obtained by using the hierarchical estimator.

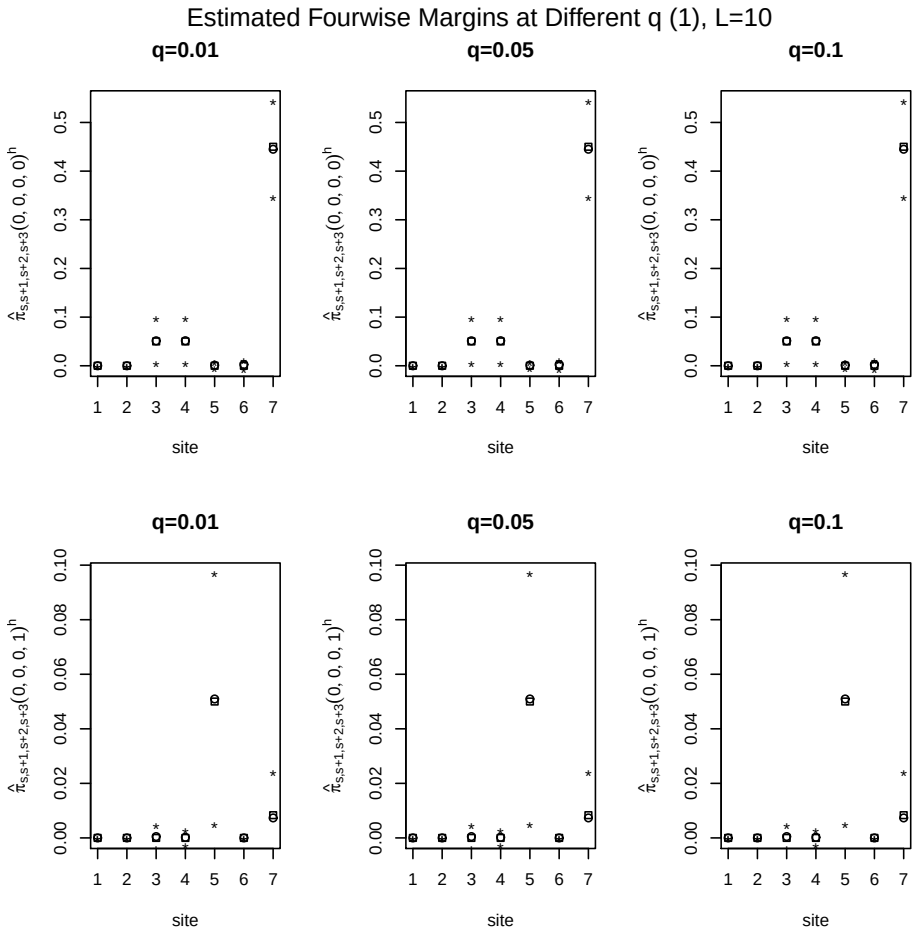


Figure C.9: Estimated Fourwise Margins at Different q (1). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

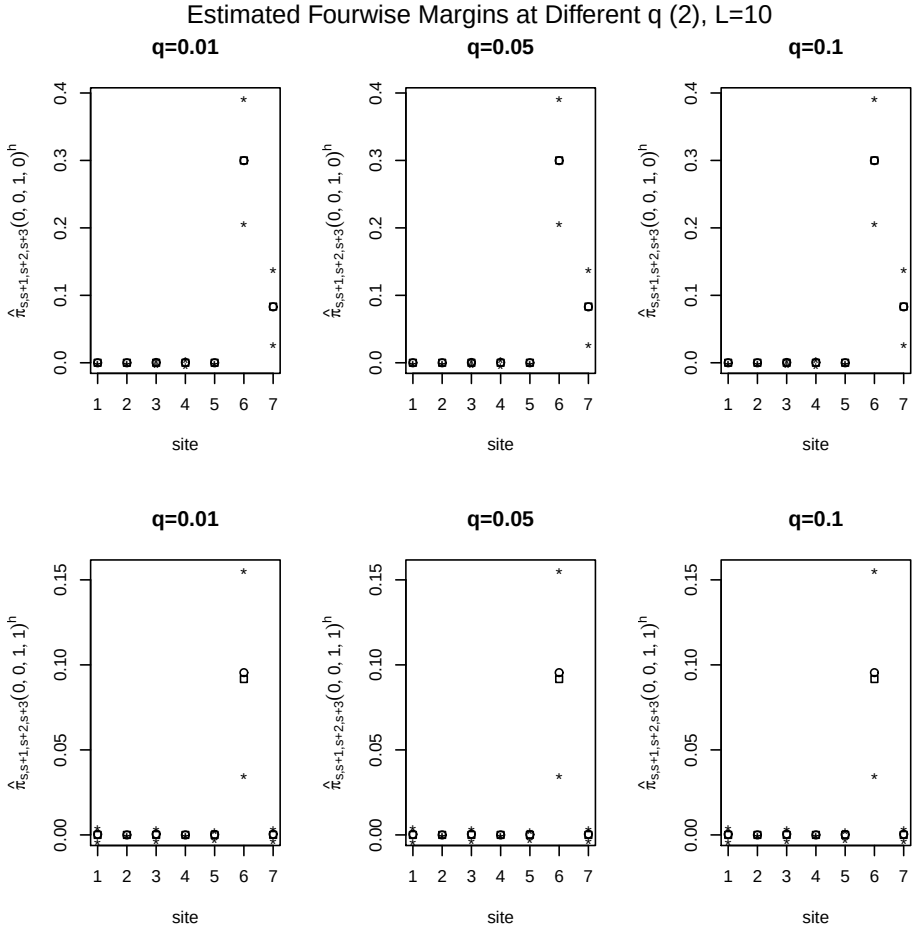


Figure C.10: Estimated Fourwise Margins at Different q (2). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

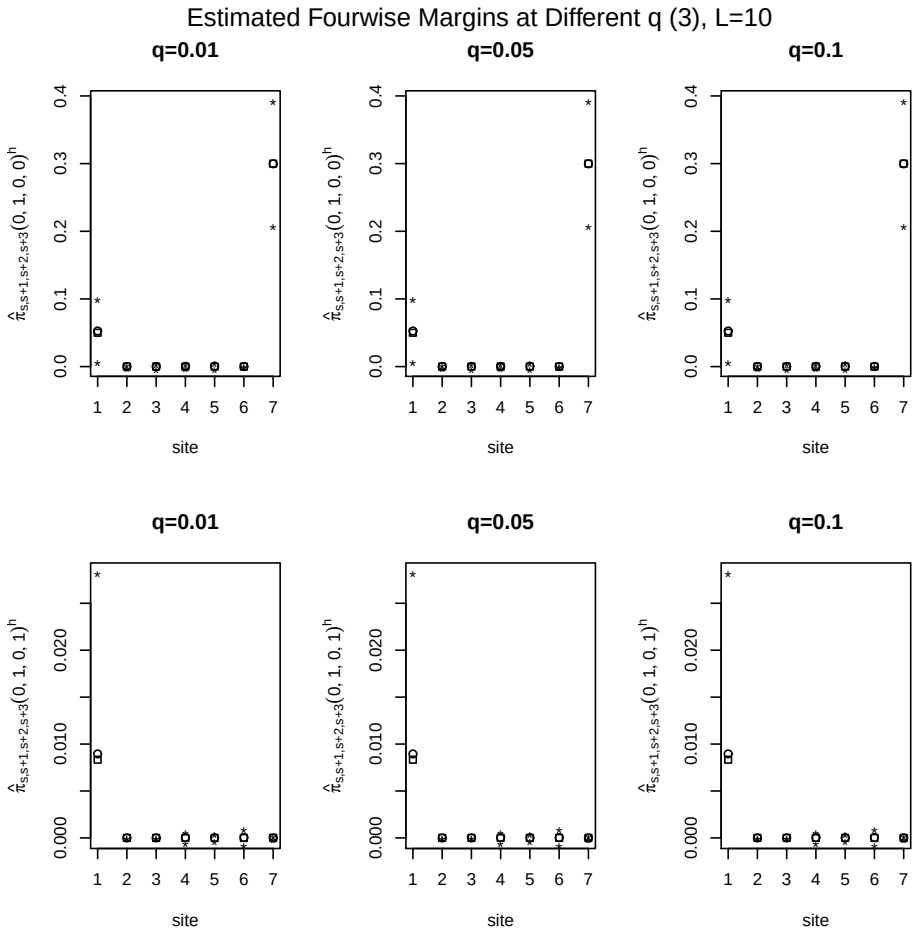


Figure C.11: Estimated Fourwise Margins at Different q (3). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

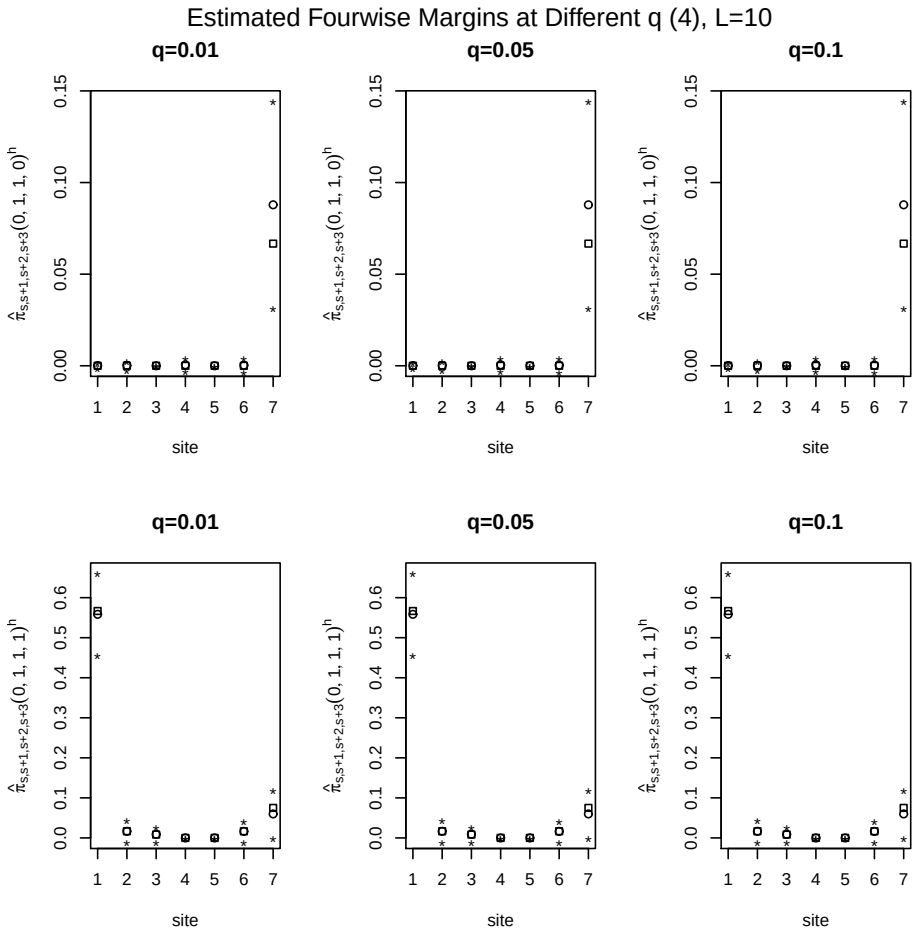


Figure C.12: Estimated Fourwise Margins at Different q (4). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

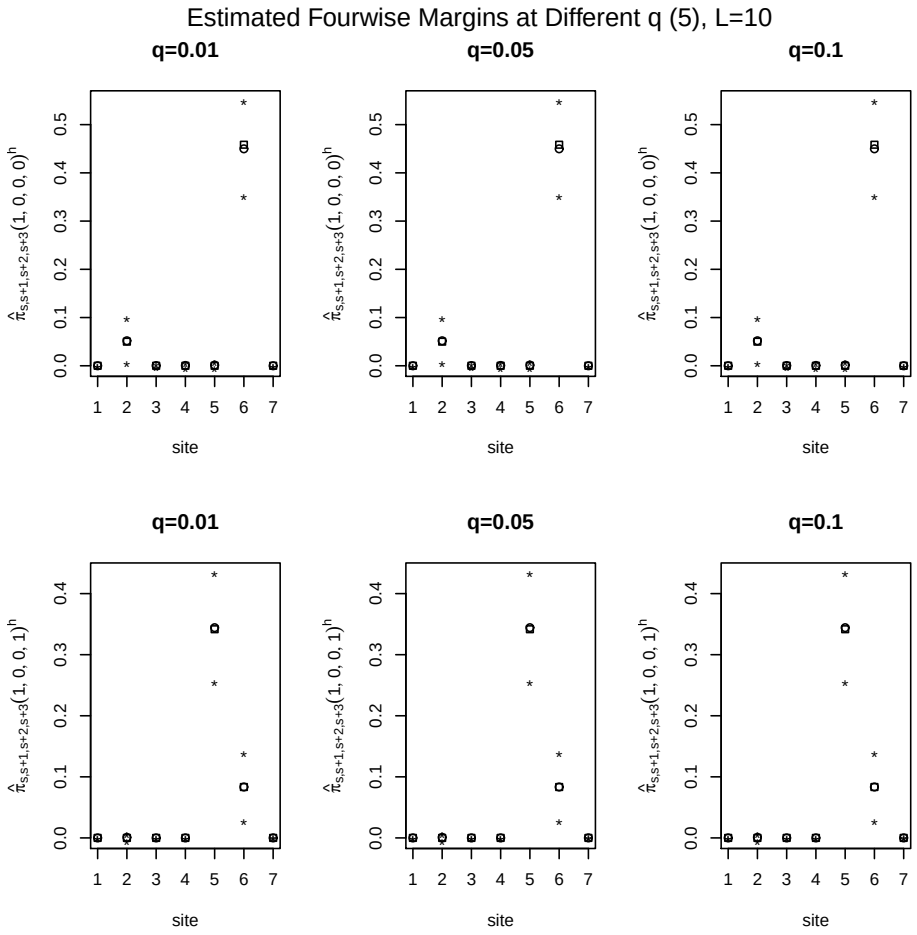


Figure C.13: Estimated Fourwise Margins at Different q (5). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

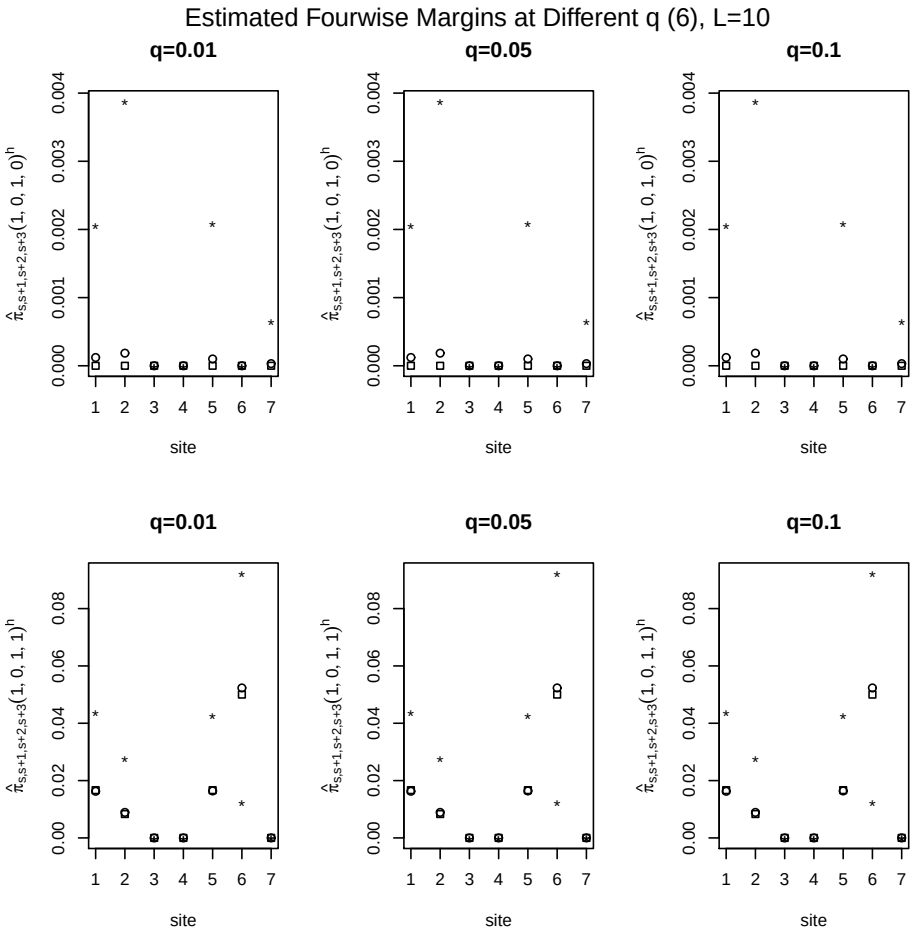


Figure C.14: Estimated Fourwise Margins at Different q (6). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

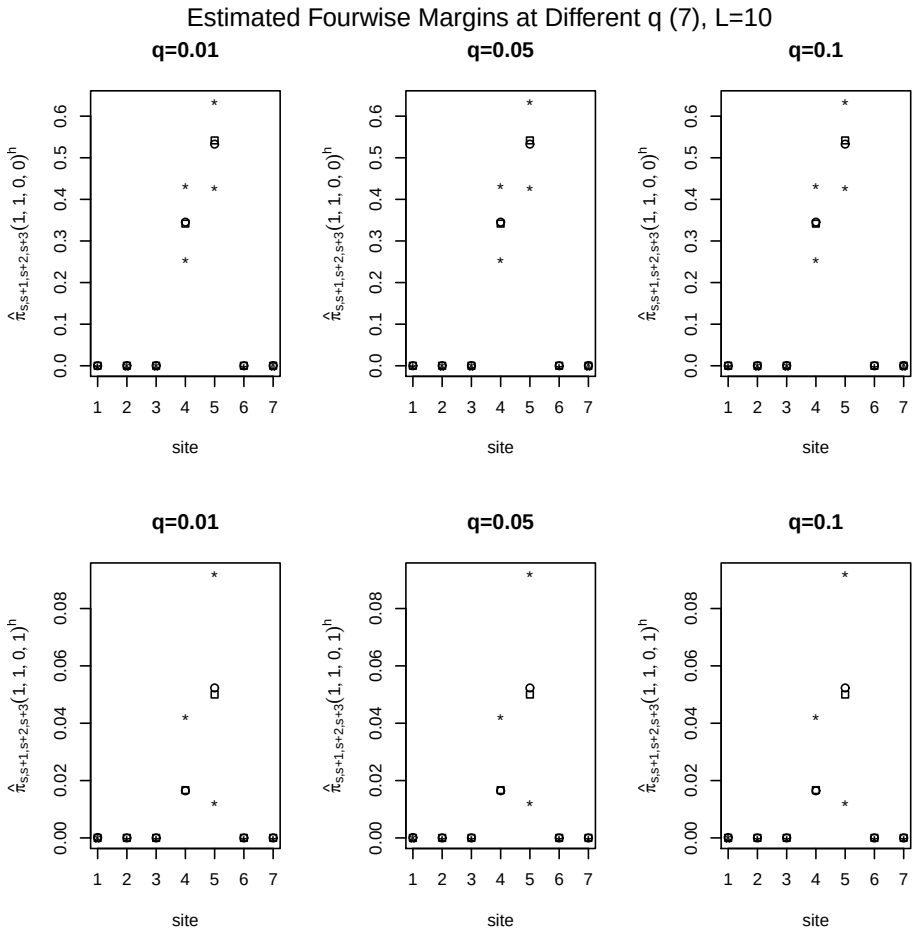


Figure C.15: Estimated Fourwise Margins at Different q (7). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

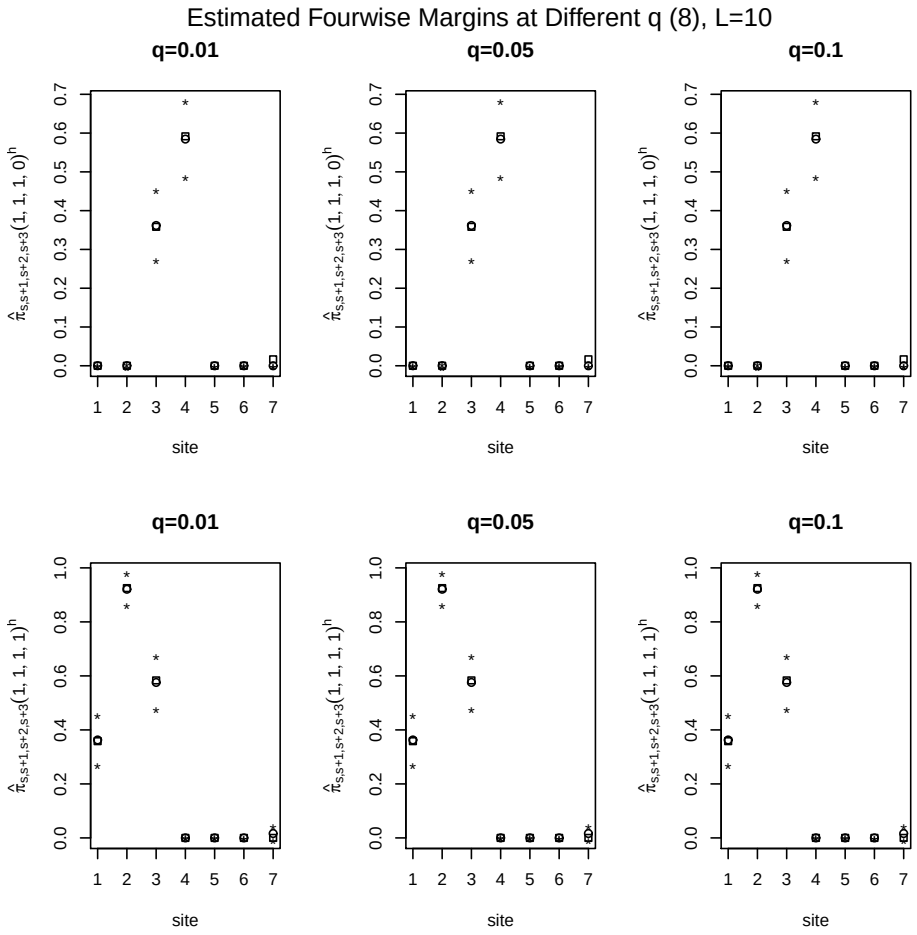


Figure C.16: Estimated Fourwise Margins at Different q (8). Here $L = 10$. Squares correspond to true margins, circles are the average estimated margins, and stars above and below the circles represent the average estimated margins plus or minus two times of standard error.

Bibliography

[1] Azzalini, A. (1983). *Maximum likelihood of order m for stationary stochastic processes*, Biometrika, 70, pp. 367-381.

[2] Besag, J.E. (1974). *Spatial interaction and the statistical analysis or lattice systems (with discussion)* Journal of the Royal Statistical Society: Series B, 36, pp. 192-236.

[3] Besag, J.E. (1977). *Efficiency of pseudolikelihood estimation for simple Gaussian fields*, Biometrika, 64, pp. 616-618.

[4] Chen, S. and Lindsay, B. G. (2006). *Building mixture trees from binary sequence data*, Biometrika, 93, 4. pp. 843-860.

[5] Cox, D. R. (1975). *Partial likelihood*, Biometrika, 62, pp. 269-276.

[6] Cox, D. R. and Reid, N. (2004). *Miscellanea-A note on pseudolikelihood constructed from marginal desities*, Biometrika, 91, 3, pp. 729-737.

[7] Ewens, W. and Grant G. (2005). *Statistical Methods in Bioinformatics : An Introduction*, Springer, New York c2005.

[8] Griffiths, R. C. and Tavaré, S. (1994). *Ancestral inference in population genetics*, Statistical Science, 9, 3. pp. 307-319.

[9] Hjort, N. L. and Varin, C. (2008). *ML, PL, QL in Markov Chain Models*, Scandinavian Journal of Statistics, 35, 1. pp. 64-82.

[10] Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules*, Academic Press, New York, pp. 21132.

[11] Kimura, M. (1980). *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences*, J. Mol. Evol. 16, pp. 111120.

[12] Kuk, A.Y.C. (2003) *Automatic choice of driving values in Monte Carlo likelihood approximation via posterior simulations*, Stat. Comput. 13, no. 2, pp. 101–109. Database Expansion Item

[13] Kuk, A.Y.C. (2007). *A hybrid pairwise likelihood method*, Biometrika, 94, pp. 939-952.

[14] Lindsay, B. G. (1988). *Composite likelihood method.* Contemporary Mathematics , 80, pp. 221-39.

[15] Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Application* Institute of Mathematical Statistics.

[16] Mao, X. (2010). *Density estimation and Modal based method for haplotyping and recombination*, Ph.D. dissertation.

[17] McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*, 2nd ed. London: Chapman and Hall.

[18] Small, C.G. and McLeish D.L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*, John Wiley & Sons (New York; Chichester)

[19] Stein, M. (2004). *Approximating likelihoods for large spatial data sets*, Journal of the Royal Statistical Society B, 66, pp. 79-86.

[20] Sung, Y.J and Geyer, C.J. (2007) *Monte Carlo likelihood inference for missing data models*, The Annals of Statistics, 35, no. 3, pp. 990-1011.

[21] Thompson, E.A. (1994) *Monte Carlo likelihood in genetic mapping*, Statistical Science, 9, no. 3, pp. 355-366.

[22] Thompson, E.A. (1996) *Likelihood and Linkage: From Fisher to the Future*, The Annals of Statistics, 24, no. 2, pp. 449-465.

[23] Varin, C. and Vidoni, P. (2005). *A note on composite likelihood inference and model selection*, Biometrika, 92, pp.519-528.

[24] Varin, C. (2008). *On composite marginal likelihoods*, AStA Advances in Statistical Analysis, 92, 1, pp.1-28.

Vita

Research Interests

- Composite Likelihood, Statistical Genetics, Statistical Computing.

Education

- Ph.D., Statistics, The Pennsylvania State University, May 2011.
- M.S., Statistics, Nankai University, China, July 2005.
- B.S., Probability and Statistics, Nankai University, China, July 2002.

Professional Experience

- Research Assistant, Department of Statistics, The Pennsylvania State University, July 2007 to December 2010.
- Student Consultant, Statistical Consulting Center, The Pennsylvania State University, January 2007 to December 2007.
- Instructor, Introduction to Probability, Fall 2010 and Fall 2009; Elementary Statistics, Summer 2009; Experimental Methods, Summer 2008, Department of Statistics, The Pennsylvania State University.

Awards

- Graduate students travel grants, Department of Statistics, The Pennsylvania State University, 2008, 2009, and 2010.