Original Paper

Human Heredity

Hum Hered 2001;51:192–198

Received: January 10, 2000
Accepted: January 27, 2000

Effects of Implicit Parameters in Segregation Analysis

John Jen Tai Chuhsing Kate Hsiao

Division of Biostatistics, Institute of Epidemiology, National Taiwan University, Taipei, Taiwan, ROC

Key Words

Ascertainment · Biased estimation · Design parameter · Intractability · Nuisance parameter · Pseudo-mixture · Sampling scheme

Abstract

In human genetic analysis, data are collected through the so-called 'ascertainment procedure'. Statistically this sampling scheme can be thought of as a multistage sampling method. At the first stage, one or several probands are ascertained. At the subsequent stages, a sequential sampling scheme is applied. Sampling in such a way is virtually a nonrandom procedure, which, in most cases, causes biased estimation which may be intractable. This paper focuses on the underlying causes of the intractability problem of ascertained genetic data. Three types of parameters, i.e. target, design and nuisance parameters, are defined as the essences to formulate the true likelihood of a set of data. These parameters are also classified into explicit or implicit parameters depending on whether they can be expressed explicity in the likelihood function. For ascertained genetic data, a sequential scheme is regarded as an implicit design parameter, and a true pedigree structure as an implicit nuisance parameter. The intractability problem is attributed to loss of information of any implicit parameter in likelihood formulation. Several approaches to build a likelihood for estimation of the segregation ratio when only an observed pedigree structure is available are proposed.

Introduction

In human genetic analysis, collection of familial data can be thought of as a multistage sampling procedure. At the first stage, families are ascertained through one or several affected probands in each family by a prespecified ascertainment. At the subsequent stages, other relatives of the probands are sampled to form a pedigree of some structure by a fixed or sequential sampling scheme [1]. The sampling scheme may be proband-independent or proband-dependent, which results in either fixed or varied observed pedigree structures (OPS) [2, 3].

From the statistical point of view, since in a population there are only families with at least one affected member that are ascertained, the ascertainment procedure is in fact nonrandom rather than random. This particular nonrandom sampling method in genetic analysis causes ascertainment bias for estimation of genetic parameters (e.g., segregation ratio). Many methods for correction of this ascertainment bias have been proposed [1, 4–14]. However, from a different aspect, by thinking that pedigree structure is a relevant element in construction of the likelihood and should be included in analysis, Vieland and Hodge [2] contended that likelihood estimation conditioning on OPS under proband-dependent sampling is in general not correct. They concluded that in practice, because one has little choice but to condition on the OPS, the problem of genetic modeling for ascertainment data sets is, in most situations, intractable.

What supports the arguments of Vieland and Hodge about the intractability of the ascertainment problem is their definitions of true and observed likelihoods. They

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/01/0514–0192$17.50/0

Accessible online at:
www.karger.com/journals/hhe

Dr. John Jen Tai, Division of Biostatistics
Institute of Epidemiology, National Taiwan University
No. 1, Jen-Ai Rd., Sec. 1, Rm.1544
Taipei 100 (Taiwan, ROC)
Tel. +886 2 23970800 (ext. 8340)

argued that in pedigree analysis establishment of true likelihood should condition on two fundamental elements of information: true pedigree structure and sampling scheme. Incorrect estimation comes from the incorrect formulation of an observed likelihood conditioning on the observed structure as well as under the situation of a lack of information for sampling scheme. Unavoidable shortage of the relevant information in a practical study creates an inherent intractability problem.

In view of the above arguments, addressing such a problem is interesting and noteworthy. However, there still are a couple of questions that deserve further study. First, Vieland and Hodge's conclusion was based on their definitions of true and observed likelihoods. If their definitions are questioned, their conclusion needs justification. Second, if the intractability problem exists for ascertained genetic data, the next question is whether this problem is unique to ascertained genetic data, or whether it is a general problem for other than genetic data. If the problem is general, then showing it occurs for ascertained genetic data is less interesting than developing methods to reduce its severity – even if these methods may only be applicable to some simple familial data structures at present.

In this paper, we will (1) introduce the concept of target, design and nuisance parameters to define the true, conditional and incomplete conditional likelihoods. The conditional and incomplete conditional likelihoods are the likelihoods used for estimation of the target parameter in segregation analysis. (2) We will also use the concept of implicit parameters to indicate that the intractability problem is indeed a general problem in usual data analysis. (3) Finally, we will use the example given by Vieland and Hodge [2] to illustrate our idea. Three strategies, observed pedigree structure (OPS), maximum pedigree (MP) and pseudo-mixture (PM), will be proposed to try and handle the likelihood estimation procedure.

Target, Design and Nuisance Parameters

Consider a set of genetic data that is collected through a multistage sampling scheme as mentioned above. Establishment of the likelihood for this type of genetic data can be viewed from three aspects: (1) the target parameters, (2) the ascertainment and sampling schemes and (3) nuisance parameters. Target parameters are parameters of interest in a study. They are defined according to the problem which an investigator intends to resolve. In segregation and linkage analysis, segregation ratio and recombination fraction are the target parameters, respectively. We will use θ to represent target parameters. In familial studies different ascertainment and sampling schemes (e.g., sampling through ascertainment or not; proband independent or dependent sampling) will generate different genetic data structures, and consequently different likelihoods. For example, if families are drawn through a complete ascertainment procedure, truncated genetic data will be generated. The truncated genetic data have a different likelihood formulation from the random genetic data [15–17]. Although both the ascertainment and sampling procedures have an influence on likelihood formulation, the types of their influence are different. The influence of an ascertainment procedure on data structure can be expressed explicity by introducing the ascertainment probability into the corresponding likelihood. For this reason, we will call the ascertainment procedure an explicit design parameter and denote it by D_1. For a sampling procedure its influence is on the sampled structure of each pedigree. Practically, because it is difficult to formulate a distribution to model the variation of pedigree structures, its influence cannot be expressed explicitly as a parameter in likelihood function. We will call the sampling procedure an implicit design parameter and denote it by D_2. As before, since both D_1 and D_2 affect the likelihood formulation, shortage of their information causes bias in estimation of target parameters. They hence also play a nuisance role in likelihood estimation. In addition to D_1 and D_2, as a matter of fact, there are some nuisance parameters in genetic analysis. For example, in linkage analysis it is known that misspecification of genetic parameters, such as degree of dominance and penetrance, may result in calculation of wrong lod scores [18, 19]. Therefore, these parameters can be viewed as (explicit) nuisance parameters for recombination fraction. In segregation analysis, if we follow the derivation of Vieland and Hodge [2], the true pedigree structure designed in a sampling scheme implicity affects the likelihood formulation. Hence, we may consider the true pedigree structure as an implicit nuisance parameter. In the following discussion we denote the true pedigree structure by τ. The following table summarizes our concepts discussed above.

Parameters	Explicit	Implicit
Target	θ = segregation ratio, recombination fraction	–
Design	D_1 = ascertainment procedure	D_2 = segregation sampling scheme
Nuisance	δ = degree of dominance, penetrance	τ = true pedigree structure

Definitions of the Likelihoods

Now consider the segregation problem where segregation ratio θ is the target parameter that we want to estimate and the ascertainment probability τ is treated as an explicit nuisance parameter. According to the previous discussion, there are five inputs (D_1, D_2, τ, π, θ) that will affect the likelihood formulation. In usual estimation procedure, a likelihood is set conditional on how data are collected (i.e., we know the design information). Accordingly, we may define the so-called true 'likelihood' as

$$L(\theta, \pi, \tau \mid D_1, D_2). \quad (1)$$

Based on this definition, three points are worth remarking:

(a) Because τ is an implicit parameter, we cannot estimate it jointly with θ from (1). Nevertheless, if there is a way, as discussed in the following section, that allows us to obtain an estimate of τ from the familial data separately, then we may use the conditional likelihood

$$L(\theta, \pi \mid \hat{\tau}, D_1, D_2) \quad (2)$$

to estimate θ. Of course, if $\hat{\tau}$ does not coincide with τ, estimate of θ from (2) is biased.

(b) Either D_2 is unknown or too complicate to be formulated as a simple essence for establishment of the conditional likelihood (2), in this situation we have to use the incomplete conditional likelihood

$$L(\theta, \pi \mid \hat{\tau}, D_1) \quad (3)$$

for estimation of θ. Here, 'incomplete' means loss of D_2 information. Specifically, if the OPS is taken as the estimate for τ, (3) becomes

$$L(\theta, \pi \mid \hat{\tau} = \text{OPS}, D_1), \quad (4)$$

which is the observed likelihood (L_{OBS} in their notation) defined by Vieland and Hodge [2]. We will discuss more about (4) in the next section.

(c) Here, we defined $L(\theta, \pi, \tau \mid D_1, D_2)$ as the true likelihood. We may also directly define $L(\theta, \pi \mid \tau, D_1, D_2)$ as the true likelihood, as Vieland and Hodge [2] did. Moreover, if we know that there exists another explicit nuisance parameter δ that influences the likelihood formulation, we may also define $L(\theta, \pi, \tau, \delta \mid D_1 D_2)$ or $L(\theta, \pi, \delta \mid \tau, D_1, D_2)$ as the true likelihood. No matter which definition is used, the conditional or incomplete conditional likelihood (2), (3) or (4) will usually not be equivalent to any defined true likelihood unless an independent unbiased estimate $\hat{\tau}$ is obtained. This implies that the difficulty in obtaining an unbiased estimate for τ due to its implicit characteristic makes the ascertainment problem intractable. Therefore, the term 'true' likelihood should merely mean a definition about how to organize data information for (θ, π, τ). It is a contrast for indicating the difficulty in handling the estimation problem due to implicit nuisance parameters. As long as the difficulty cannot be overcome, the problem remains intractable and is felt inherent in the task. On the other hand, because the above discussion is also valid for general likelihood formulation, the intractability problem is a general problem in statistics but not a unique estimation problem occurring for ascertained genetic data. Hence, in the discussion of the ascertainment problem, the central point is not in emphasizing the intractability of the problem per se, but in developing methods improve the severity of the intractability – even if these methods may only work for some simple familial data structures at present. In the following section, we will utilize the example of Vieland and Hodge [2] to illustrate our ideas on the estimation of implicit parameters i.e., pedigree structure τ here). From a practical point of view, their example was quite unrealistic. Nevertheless, as an illustrative example, it has the advantage to clarify their notion of intractability and our ideas on the estimation of implicit parameters. Our methods can be extended to other realistic cases directly.

Estimation of Implicit Parameter τ

To simplify the discussion, we skip the review of the theoretical derivation of the true and observed likelihood functions in the paper of Vieland and Hodge [2]. We will directly use their example to address our theory on the intractability problem. The example involves ordered sibships of size $s = 3$ and a two-stage sampling scheme. At the first stage, affected individuals are independently ascertained with a common probability π. At the second stage, any previously unobserved sibling(s) adjacent to any probands in this ordering are sampled (i.e., a proband-dependent sampling). This sampling scheme generates an OPS sibship size either $s = 2$ or $s = 3$. In figure 1 of their paper, they demonstrated the probability distribution for the observed pedigree structures $s = 2$ and $s = 3$ if the true pedigree structure is known $s = 3$ and with the sampling scheme as mentioned above. Readers interested in knowing the details can refer to that figure. Following the example, we now go through the six different formulations of the likelihood: (1) true likelihood, (2) incomplete conditional likelihood with OPS strategy, (3) incomplete conditional likelihood with observed maximum pedigree structure (MPS) strategy, (4) incomplete conditional likelihood

Table 1. Probabilities of sibship size s and number of affected sibs r for the true situation ($\tau = +$, $D_1 = +$, $D_2 = +$) and observed situation ($\hat{\tau} = \text{OPS}$, $D_1 = +$, $D_2 = -$) under L_{OBS} strategy and ($\hat{\tau} = \text{MPS}$, $D_1 = +$, $D_2 = -$) under MP strategy

Observation		Probability of observation		
s	r	True $P(s, r; \theta, \pi \mid \tau = +, D_1 = +, D_2 = +)$	OPS $P(s, r; \theta, \pi \mid \hat{\tau} = \text{OPS} = s, D_1 = +, D_2 = -)$	MP $P(s, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = -)$
2	1	$2\theta\pi(1-\theta\pi)(1-\theta)/D_3$	$2\theta\pi(1-\theta)/D_2$	$2\theta\pi(1-\theta)/E_3$
2	2	$2\theta^2\pi(1-\theta\pi)(1-\pi)/D_3$	$\theta^2\pi(2-\pi)/D_2$	$\theta^2\pi(2-\pi)/E_3$
3	1	$\theta\pi(1-\theta)^2/D_3$	$3\theta\pi(1-\theta)^2/D_3$	$3\theta\pi(1-\theta)^2/E_3$
3	2	$\theta^2\pi(1-\theta)(2+\pi)/D_3$	$3\theta^2\pi(1-\theta)(2-\pi)/D_3$	$3\theta^2\pi(1-\theta)(2-\pi)/E_3$
3	3	$\theta^3\pi[1+\pi(1-\pi)]/D_3$	$\theta^3[1-(1-\pi)^3]/D_3$	$\theta^3[1-(1-\pi)^3]/E_3$

$D_2 = 1-(1-\theta\pi)^2$, $D_3 = 1-(1-\theta\pi)^3$, $E_3 = 2-(1-\theta\pi)^2-(1-\theta\pi)^3$.

with PM strategy, (5) conditional likelihood with MPS strategy, and (6) conditional likelihood with PM strategy.

True Likelihood $L(\theta, \pi \mid \tau = +, D_1 = +, D_2 = +)$

Table 1 lists the probability distribution $P(s, r; \theta, \pi \mid \tau = +, D_1 = +, D_2 = +)$ for the data in the form of (s, r), where r is the number of affected individuals in a sibship and ($\tau = +$, $D_1 = +$, $D_2 = +$) represents that the true pedigree structure (τ), ascertainment scheme (D_1) and sampling scheme (D_2) are all known (+). Denote the number of families of data (s, r) by n_{sr}. The true likelihood is established following a multinomial distribution.

$$L(\theta, \pi \mid \tau = +, D_1 = +, D_2 = +)$$

$$= L_{\text{True}}(\theta) \propto \prod_{s,r} \{P(s, r; \theta, \pi \mid \tau = +, D_1 = +, D_2 = +)\}^{n_{sr}},$$

where the phenotypic probabilities of the siblings are assumed to be independent and θ is the probability that an individual is affected, which is the target parameter in this example.

Incomplete Conditional Likelihood $L(\theta, \pi \mid \hat{\tau} = \text{OPS}, D_1 = +, D_2 = -)$

In the situation that both τ and D_2 are unknown (in our notation $\tau = -$, $D_2 = -$), we may take the OPS as the true one for establishing the likelihood. That is to say, when the sibship sizes $s = 2$ and $s = 3$ are both observed in a sample, the data are split as if they were from two independent systems, one from the true pedigree structure of sibship size 2 (denoted by $\hat{\tau} = \text{OPS} = 2$) and the other from the true pedigree structure of sibship size 3 (denoted by $\hat{\tau} = \text{OPS} = 3$). In the system of $\hat{\tau} = 3$, when the sampling scheme is unknown, one cannot assure whether a sibship size less than 3 (i.e., $s = 2$ here) has been generated. Therefore, we suggest to take a conservative strategy that only sibships of size $s = 3$ are included in the system. Similarly, there are only sibships of size $s = 2$ included in the system of $\hat{\tau} = 2$. With this strategy the probability distributions of observing the family data (s, r) are calculated depending on the observed structure $s = 2$ or $s = 3$, and expressed as $P(s = 2, r; \theta, \pi \mid \hat{\tau} = s = 2, D_1 = +, D_2 = -)$ and $P(s = 3, r; \theta, \pi \mid \hat{\tau} = s = 3, D_1 = +, D_2 = -)$, respectively (table 1). The observed likelihood function turns out to be

$$L(\theta, \pi \mid \hat{\tau} = \text{OPS} = s, D_1 = +, D_2 = -) = L_{--,\text{OPS}}(\theta, \pi)$$

$$\propto \prod_{r=1}^{2} (s = 2, r; \theta, \pi \mid \hat{\tau} = 2, D_1 = +, D_2 = -)^{n_{2r}} \times$$

$$\prod_{r=1}^{3} \{P(s = 3, r; \theta, \pi \mid \hat{\tau} = 3, D_1 = +, D_2 = -)\}^{n_{3r}}$$

where the first and the second subscripts '–' in $L_{--,\text{OPS}}$ represent lack of the information of τ and D_2, respectively; the subscript OPS represents that the observed sibship size s is taken as an estimate for τ. Note that this is also the likelihood (4) defined by Vieland and Hodge.

The above strategy for the situation ($\hat{\tau} = s$, $D_1 = +$, $D_2 = -$) indicates that the intractability of the ascertainment problem comes from: (1) a nonrandom ascertainment procedure, (2) lack of τ information, (3) lack of D_2 information, and (4) the strategy of handling the lack of τ and D_2 information. It is known that in segregation analysis a nonrandom ascertainment procedure does interfere with the correct estimation of θ, and its influence goes off as the ascertainment probability π tends to 0 (i.e., single ascertainment). The effect of lack of τ and D_2 information on estimation of θ has been discussed in the previous sec-

Table 2. Probabilities of sibship size s and number of affected sibs r for the observed situation ($\hat{\tau} = -$, $D_1 = +$, $D_2 = -$) under PM strategy

Observation		Probability of observation	
s	r	$P(s, r; \theta, \pi \mid \hat{\tau} = \text{MPS} - 1 = 2, D_1 = +, D_2 = -)$	$P(s, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = -)$
2	1	$2\theta\pi(1 - \theta)/D_2$	$2\theta\pi(1 - \theta)/E_3$
2	2	$\theta^2\pi(2 - \pi)/D_2$	$\theta^2\pi(2 - \pi)/E_3$
3	1	0	$3\theta\pi(1 - \theta)^2/E_3$
3	2	0	$3\theta^2\pi(1 - \theta)(2 - \pi)/E_3$
3	3	0	$\theta^3[1 - (1 - \pi^3]/E_3$

$D_2 = 1 - (1 - \theta\pi)^2$ and $E_3 = 2 - (1 - \theta\pi)^2 - (1 - \theta\pi)^3$.

tion. In the following, two other strategies, MP and PM, will be introduced to indicate that taking $\hat{\tau} = \text{OPS}$ is just a strategy for estimation of the implicit nuisance parameter τ. In other words, we want to point out that the conclusion of Vieland and Hodge [2] about the intractability of the ascertainment problem was based on their definition for the likelihood L_{OBS} (θ) (in their notation). There is no unquestionable reason to believe that an observed likelihood must be defined like that.

Incomplete Conditional Likelihood $L(\theta, \pi \mid \hat{\tau} = \text{MPS}, D_1 = +, D_2 = -)$

If we can be certain that various OPS in a sample are all generated from a simple true pedigree structure by a stipulated sampling scheme, then it is reasonable to believe that the observed MPS should be the true pedigree structure or close to it. Therefore, instead of taking the OPS strategy, we can take the observed MPS as the true pedigree structure. We call this strategy the MP strategy. For the above example, because $\tau = \text{MPS}$ is taken, all the possible outcomes ($s = 2$, r) and ($s = 3$, r) are pooled into one system to calculate the probability distribution $P(s, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = -)$. The results are also listed in table 1. The observed likelihood function under the MP strategy is

$$L(\theta, \pi \mid \hat{\tau} = \text{MPS}, D_1 = +, D_2 = -) = L_{--,\text{MP}}(\theta, \pi)$$

$$\propto \prod_{s, r} \{P(s, r; \theta, \pi \mid \hat{\tau} = \text{MPS}, D_1 = +, D_2 = -)\}^{n_{sr}}$$

Note that in practical investigations, OPS may not all be generated from a true pedigree structure. The following strategy may be considered as a useful alternative for these complicated situations.

Incomplete Conditional Likelihood $L(\theta, \pi, \delta \mid \hat{\tau} = \text{PM}, D_1 = +, D_2 = -)$

The OPS strategy treats the observed pedigree structures $s = 2$ and $s = 3$ as sampled independently from the $\tau = 2$ and $\tau = 3$ systems, respectively, and in each system only this structure appears. Obviously, whether or not the observed pedigree structures are all generated from a true pedigree structure, this strategy will not coincide with the most realistic cases and hence causes estimation bias. (Note that the defined L_{OBS} of Vieland and Hodge has the same formulation as the OPS strategy, so this criticism also applies to their derivation. In other words, if an observed likelihood is not defined as they did, thier conclusion for intractability may be different.) Alternatively, we can treat the two OPS $s = 2$ and $s = 3$ as two true pedigree structures $\tau = s = 2$ and $\tau = s = 3$ by the same sampling scheme. This idea is acceptable from the practical point of view because: (1) there is no strong reason to believe that the observed pedigree data must be generated from a single true pedigree structure of certain sibship size when we have no information about τ, and (2) even if there is exactly one true pedigree structure of certain sibship size, since the OPS strategy still represents a method of handling the lack of information for τ, it is worth trying the mixture strategy rather than treating them as from two independent systems. Therefore, under $\tau = s = 2$, definitely only structures of $s = 2$ are observed (here, we assume structures of $s = 1$ are not included in sampling); under $\tau = s = 3$, both structures of $s = 2$ and $s = 3$ may be observed. Thus, when we observe $s = 2$ we really do not know whether this structure is from $\tau = 2$ or $\tau = 3$. It is natural to use the mixture technique to deal with this uncertainty. We call this strategy the pseudo-mixture (PM) strategy, indicating that we simply take the strategy as a means for solving the uncertainty problem of the true pedigree structure.

Table 3. Probabilities of sibship size s and number of affected sibs r for the observed situation ($\hat{\tau} = -$, $D_1 = +$, $D_2 = +$) under PM strategy

Observation		Probability of observation	
s	r	$P(s, r, \theta, \pi \mid \hat{\tau} = \text{MPS} - 1 = 2, D_1 = +, D_2 = +)$	$P(s, r, \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = +)$
2	1	$2\theta\pi(1 - \theta)/D_2$	$2\theta\pi(1 - \theta\pi)(1 - \theta)/D_3$
2	2	$\theta^2\pi(2 - \pi)/D_2$	$2\theta^2\pi(1 - \theta\pi)(1 - \pi)/D_3$
3	1	0	$\theta\pi(1 - \theta)^2/D^3$
3	2	0	$\theta^2\pi(1 - \theta)(2 + \pi)/D_3$
3	3	0	$\theta^3\pi[1 + \pi(1 - \pi)]/D_3$

$D_2 = 1 - (1 - \theta\pi)^2$ and $D_3 = 1 - (1 - \theta\pi)^3$.

The probability distributions $P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} - 1 = 2, D_1 = +, D_2 = -)$, $P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = -)$ and $P(s = 3, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = -)$ are listed in table 2. The likelihood function is

$$L(\theta, \pi, \delta \mid \hat{\tau} = \text{PM}, D_1 = +, D_2 = -) = L_{--,\,\text{PM}}(\theta, \pi, \delta)$$

$$\propto \prod_{r=1}^{2} \{\delta P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} - 1 = 2, D_1 = +, D_2 = -) +$$

$$(1 - \delta) P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = -)\}^{n_{2r}} \times$$

$$\prod_{r=1}^{3} \{(1 - \delta) P(s = 3, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = -)\}^{n_{3r}}$$

where $\hat{\tau} = \text{PM}$ represents that the PM strategy is taken under the situation ($\tau = -$, $D_1 = +$, $D_2 = -$), and δ is the mixture weight for $\hat{\tau} = 2$ when observing $s = 2$.

Conditional Likelihood $L(\theta, \pi \mid \hat{\tau} = \text{MPS}, D_1 = +, D_2 = +)$

When the sampling scheme is known, an OPS $s = 2$ can be generated by taking either true pedigree structure $s = 2$ or $s = 3$ so that it is clear that the OPS strategy is not applicable (this is true for most reaslistic cases). Therefore, we will not discuss the likelihood $L_{-+,\text{OPS}}$ for this situation. The probability distribution $P(s, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = +)$ under MP strategy is the same as the true one. $P(s, r; \theta, \pi \mid \tau = +, D_1 = +, D_2 = +)$ as listed in table 1. The corresponding likelihood function is

$$L(\theta, \pi \mid \tau = \text{MPS}, D_1 = +, D_2 = +) = L_{-+,\text{MP}}(\theta, \pi)$$

$$\propto \prod_{s, r} \{P(s, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = +)\}^{n_{sr}}$$

which is the same as L_{True}.

Conditional Likelihood $L(\theta, \pi, \delta \mid \hat{\tau} = \text{PM}, D_1 = +, D_2 = +)$

If the true pedigree structure is unknown but the sampling scheme is known and the PM stragegy is taken, the probability distribution $P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} - 1 = 2, D_1 = +, D_2 = +)$, $P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = +)$ and $P(s = 3, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = +)$ are calculated as in table 3. The mixture likelihood function is

$$P(s, \pi, \delta \mid \hat{\tau} = \text{PM}, D_1 = +, D_2 = +) = L_{-+,\text{PM}}(\theta, \pi, \delta)$$

$$\propto \prod_{r=1}^{2} \{\delta P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} - 1 = 2, D_1 = +, D_2 = +) +$$

$$(1 - \delta) P(s = 2, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = +)\}^{n_{2r}} \times$$

$$\prod_{r=1}^{3} \{(1 - \delta) P(s = 3, r; \theta, \pi \mid \hat{\tau} = \text{MPS} = 3, D_1 = +, D_2 = +)\}^{n_{3r}}$$

Conclusion

In this paper we have discussed the intractability of the ascertainment problem addressed by Vieland and Hodge from two aspects: (1) the roles of the true pedigree structure and sampling scheme in likelihood estimation procedure for a target parameter; (2) the strategies to build a likelihood for estimation of the target parameter θ when there is only observed information for pedigree structure that can be used. For the first point we have used implicit characteristics of nuisance and design parameters to illustrate why loss of information for the true pedigree structure or sampling scheme may introduce the concept of an intractability problem. We point out that intractability is not unique to ascertained genetic data. Mainly, we separate the concepts of ascertainment and sampling from

each other, and distinguish between 'target', 'design', and 'nuisance' parameters. Our approach is helpful for resolving the difficult intractability problem discussed by Vieland and Hodge.

For the second point, we propose three approaches to demonstrate that the way to define an observed likelihood may not necessarily be as proposed by Vieland and Hodge. There are some alternatives. Our approach is essentially a two-stage estimation procedure. At the first stage, we can take the OPS or MP approach to obtain an estimate of the true pedigree structure; then, given the estimated pedigree structure, the target parameters are estimated through formulation of the likelihood. If both OPS and MP are considered unsuitable in a realistic analysis, the PM approach is suggested. To simplify our discussion and illustrate our approach, we have used the example of Vieland and Hodge [2]. Though this example merely considers sibships of 2 or 3 sibs, there is no difficulty in extending our approaches to sibships of many different sizes. Generalization of our approaches to more general situations, for example, to arbitrary pedigrees larger than sibships or to more general problems rather than estimating the segregation ratio, deserves further study.

Acknowledgments

This research was supported by NSC grant no. 88-2118-M-002-010, Republic of China.

References

1 Cannings C, Thompson EA: Ascertainment in the sequential sampling of pedigrees. Clin Genet 1977;12:208–212.
2 Vieland VJ, Hodge SE: Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. Am J Hum Genet 1995;56:33–43.
3 Vieland VJ, Hodge SE: The essence of single ascertainment. Genetics 1996;144:1215–1223.
4 Elandt-Johnson RC: Segregation analysis for complex modes of inheritance. Am J Hum Genet 1970;22:129–144.
5 Elston RC, Sobel E: Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 1979;31:62–69.
6 Boehnke M, Greenberg DA: The effects of conditioning on probands to correct for multiple ascertainment. Am J Hum Genet 1984;36: 1298–1308.
7 Ewens WJ, Shute NCE: A resolution of the ascertainment sampling problem. I. Theory. Theor Popul Biol 1986;30:523–542.
8 Hodge SE: Conditioning on subsets of the data: Applications to asceretainment and other genetic problems. Am J Hum Genet 1988;43: 364–373.
9 Shute NCE, Ewens WJ: A resolution of the ascertainment sampling problem. II. Generalizations and numerical results. Am J Hum Genet 1988;43:374–386.
10 Shute NCE, Ewens WJ: A resolution of the ascertainment sampling problem. III. Pedigrees Am J Hum Genet 1988;43:387–395.
11 Elston RC: Twixt cup and lip: How intractable is the ascertainment problem? Am J Hum Genet 1995;56:15–17.
12 Rabinowitz D: A pseudolikelihood approach to correcting for ascertainment bias in family studies. Am J Hum Genet 1996;59:726–730.
13 Karunaratne PM, Elston RC: Likelihood calculation conditional on observed pedigrees structure. Am J Hum Genet 1998;62:738–739.
14 Bonney GE: Ascertainment corrections based on smaller family units. Am J Hum Genet 1998;63:1202–1215.
15 Li CC: A method of subdividing genetic data into self-contained subsets. Ann Hum Genet 1986;50:259–270.
16 Tai JJ: Self-contained subsets method for estimation of gene frequencies of truncated genetic data. Genet Epidemiol 1997;14:465–477.
17 Yao YC, Tai JJ: Bias correction for segregation ratio estimation in human genetics. Biometrics, in press.
18 Clerget-Darpoux F, Bonaiti-Pellié C, Hochez J: Effects of misspecifying genetic parameters in lod score analysis. Biometrics 1986;42:393–399.
19 Hodge SE, Elston RC: Lods, wrods, and mods: The interpretation of lod scores calculated under different models. Genet Epidemiol 1994; 11:329–342.