应用概率统计　第二十五卷
第四期　2009年8月

Chinese Journal of Applied Probability
and Statistics Vol.25 No.4 Aug. 2009

对含未知基因型个体的家系进行单倍型推断的EM方法 *

赵　红[1,2]　朱文圣[1⋆]　郭建华[1]

([1]东北师范大学应用统计教育部重点实验室&数学与统计学院, 长春, 130024)

([2]中国海洋大学数学系, 青岛, 266071)

摘　要

单倍型推断在现代连锁分析和关联分析中起着非常关键的作用. 目前的单倍型推断方法基本上是根据基因型去推断个体的单倍型, 而实际家系中某些个体的基因型经常是有部分缺失或者是完全未知的. 本文给出了当家系中含有部分缺失或者完全缺失基因型个体时的单倍型推断的EM方法, 并且给出了参数估计的标准差, 最后通过模拟研究证实了我们的方法的可行性.

关键词: 单倍型推断, 家系, EM算法.

学科分类号: O212.

§1. 引　　言

单倍型推断(haplotype inference)是现代遗传流行病学研究中非常重要的问题, 它是人们基于单倍型进行连锁分析(linkage analysis)和关联分析(association analysis)的前提和基础. 但是利用目前的生物技术我们很难直接得到个体的单倍型, 实际中观测到的数据都是未含连锁相(phase)信息的基因型(genotype)数据. 因此, 要想基于单倍型进行连锁分析和关联分析的研究, 首先必须去推断个体的单倍型.

许多作者都致力于单倍型推断方法的研究, 并且提出了大量的行之有效的方法. 这些方法大致可分为三大类: 第一类是分子实验室方法, 这类方法虽说结果比较准确但是花费高而且时间长, 所以不能被广泛应用; 第二类方法是利用家系中亲属之间的关系去推断单倍型, 但是这种方法要求家系中个体的基因型没有缺失, 并且当位点比较多时个体的单倍型仍旧不能唯一确定; 而利用统计方法进行单倍型推断是目前公认的最为简单有效的方法, 比如说Clark算法[1], EM算法[2–11], Bayesian方法[12]等等. 这些统计方法又可分为基于群体结构(个体之间相互独立)的推断方法和基于家系结构的推断方法.

许多研究表明, 基于家系结构的单倍型推断方法能够提高单倍型推断的精确性[6–9], 因为利用家系中亲属之间的关系可以减小个体单倍型的不确定性. 其中Zou等人[8]以及Zhu等人[9]讨论了基因型带有测量误差时的单倍型推断问题. 但是, 这些方法都没有考虑家系中

*国家自然科学基金项目(10701022, 10871038, 10826110)、国家973计划(2007CB311002)和吉林省杰出青年科学研究基金资助项目(20030113)资助.

⋆通讯作者, E-mail: wszhu@nenu.edu.cn.

本文2006年1月4日收到, 2007年11月23日收到修改稿.

个体基因型有缺失的情况, 而在实际中, 由于某些生物技术的缺陷或某些人为因素, 使得家系中某些个体的基因型数据无法得到或者得到的数据有部分缺失, 比如某些个体的迁移或死亡等, 这类个体我们统称为“未知基因型个体”. 由于在家系结构的单倍型推断过程中, 亲属间的关系是非常有用的信息, 它能够帮助我们减小个体单倍型的不确定性, 因此如果我们从家系中去掉这些个体, 则会损失许多重要的信息, 从而降低参数估计的准确性. 最近, Ding等人[10]考虑了核心家庭(只有一对父母及他们孩子组成的特殊家系)中父母仅有一个个体的基因型有缺失情况下的单倍型推断方法, 而Liu等人[11]随之提出了核心家庭中父母双方基因型全部缺失情况下的单倍型推断方法. 但是, 对于一般的大的家系中含有未知基因型个体的情况还没有提出有效的单倍型推断方法.

本文针对家系中含有未知基因型个体的情况, 充分利用家系中个体之间的亲属关系, 给出了单倍型概率估计的EM方法, 进而去推断家系中所有个体的单倍型, 同时我们还给出了参数估计的标准差, 最后通过模拟研究证实了我们方法的可行性.

§2. 基本定义和记号

假设我们总共观测到I个家系, 而每一个个体我们测定了其紧密连锁的L个位点的基因型. 在每一个家系中, 我们把没有亲代的个体称为该家系的祖先个体(founder), 而把有亲代的个体称为该家系的非祖先个体(non-founder). 对于第i个家系来说, 假设共有$N_i^{(1)}+N_i^{(2)}+N_i^{(n)}$个个体, 其中$N_i^{(1)}$是基因型完全已知的祖先个体数目; $N_i^{(2)}$是未知基因型的祖先个体的数目; 剩下的$N_i^{(n)}$个个体是非祖先个体, 其基因型可能完全已知也可能是有缺失的, 于是我们定义下面的记号, 对于第i个家系:

$$\begin{cases} G_{ij}^{(1)}: \text{第}j\text{个完全基因型祖先个体的基因型}, j=1,\cdots,N_i^{(1)};\\ G_{ik}^{(2)}: \text{第}k\text{个未知基因型祖先个体的基因型}, k=1,\cdots,N_i^{(2)};\\ G_{il}^{(n)}: \text{第}l\text{个非祖先个体的基因型}, l=1,\cdots,N_i^{(n)}. \end{cases}$$

其中$i=1,\cdots,I$. 我们假设各家系之间是相互独立的, 而同一家系中祖先个体之间也是相互独立的.

为简单起见, 我们仅考虑这样的家系, 即家系中所有的核心家庭的两代个体中, 最多只有一代个体的基因型有缺失的情况, 如果非祖先个体的基因型有缺失, 则其父母的基因型是完全已知的, 我们可以通过其亲代的基因型按照孟德尔遗传法则去推断其所有可能的基因型; 而如果祖先个体的基因型有缺失, 则我们把其所属的核心家庭视为一个整体, 作为一个特殊祖先来处理, 并且把该核心家庭中所有个体的基因型放在一起组成一个向量G_{is}^*, 称其为第i个家系中的第s个核心家庭基因型向量, 其中$s=1,\cdots,S_i$, S_i表示第i个家系中上述核心家庭的数目. 在我们的方法中, 假设每个个体只结婚一次, 因此$S_i\le N_i^{(2)}$. 进一步的, 在第i个家系中, 我们记这S_i个核心家庭以外的祖先个体和非祖先个体总数分别为R_i和T_i, 那么显然有$R_i\le N_i^{(1)}$, $T_i\le N_i^{(n)}$, 其中$i=1,\cdots,I$. 于是

我们相当于把第i个家系中所有个体的基因型重新进行了排列, 从而得到新的基因型向量$G_i = (G_{i1}^{(1)}, \cdots, G_{iR_i}^{(1)}, G_{i1}^{*}, \cdots, G_{iS_i}^{*}, G_{i1}^{(n)}, \cdots, G_{iT_i}^{(n)})'$. 我们把所有$I$个家系的观测数据记为$\mathbf{G} = (G_1, G_2, \cdot, G_I)'$.

§3. 单倍型推断的EM方法

如果某一给定的基因型共有L个杂合位点, 则与之相匹配的所有可能的双倍型有2^{L-1}种, 并且我们很容易就能给出这2^{L-1}种可能的双倍型. 令$d_{ia_r}^{(1)} = \{x, y\}$表示与祖先个体基因型$G_{ir}^{(1)}$相匹配的第$a_r$个双倍型, 记为$x \oplus y = G_{ir}^{(1)}$, $r = 1, \cdots, R_i$, $a_r = 1, \cdots, A_{ir}$, 其中A_{ir}表示与基因型$G_{ir}^{(1)}$相匹配的双倍型总数. 类似的, 令$d_{ic_t}^{(n)} = \{x, y\}$表示与非祖先个体基因型$G_{it}^{(n)}$相匹配的第$c_t$个双倍型, $t = 1, \cdots, T_i$, $c_t = 1, \cdots, C_{it}$, 其中C_{it}表示与基因型$G_{it}^{(n)}$相匹配的双倍型总数. 对于核心家庭基因型向量G_{is}^{*}来说, 其中的祖先个体的基因型有缺失, 我们不能直接确定与之相匹配的所有可能的双倍型向量, 因此我们有必要详细地讨论其确定方法.

3.1 确定与G_{is}^{*}相匹配的双倍型向量的算法

我们把G_{is}^{*}作为一个整体来考虑, 不妨假设该核心家庭中共有k个孩子, 并且根据前面的假设我们知道他们的基因型是完全已知的. 下面我们分两步去确定与G_{is}^{*}相匹配的双倍型向量.

1) 根据k个孩子的基因型确定与之相匹配的所有可能的双倍型向量, 而每一个可能的双倍型向量最多含有4种不同的单倍型, 因为孩子的单倍型是由父母单倍型传递下来的, 为了说明的简单, 记这k个孩子一种可能的双倍型向量为$(d_{is_1}^{c}, \cdots, d_{is_k}^{c})$.

2) 根据孩子的双倍型向量$(d_{is_1}^{c}, \cdots, d_{is_k}^{c})$去构造父母的所有可能的双倍型向量. 我们把满足$\prod_{l=1}^{k} \mathsf{P}(d_{is_l}^{c} | d_{isf}, d_{ism}) > 0$的双倍型对$d_{isp} = (d_{isf}, d_{ism})$作为父母的一种可能的双倍型向量, 于是便得到与$G_{is}^{*}$相匹配的一种可能的双倍型向量$(d_{isp}, d_{is_1}^{c}, \cdots, d_{is_k}^{c})$.

按照上面两步我们可以找到与G_{is}^{*}相匹配的所有的双倍型向量, 总数记为B_{is}, 与G_{is}^{*}相匹配的第b_s个可能的双倍型向量记为$d_{ib_s}^{*}$, 令$d_{ib_sp}^{*}$表示$d_{ib_s}^{*}$中相应的父母的双倍型向量, 其中$b_s = 1, \cdots, B_{is}$.

3.2 参数估计的EM算法

我们假设Hardy-Weinberg平衡成立, 以祖先个体为例, 即

$$\mathsf{P}(d_{ia_r}^{(1)} = \{x, y\}) = \begin{cases} 2\theta_x\theta_y, & x \neq y; \\ \theta_x^2, & x = y. \end{cases}$$

而对于其他个体来说类似的结论也成立, 其中θ_x, θ_y分别是单倍型x和y的群体概率. 进一步, 我们假设L个位点是紧密连锁的, 对于第r个子代来说, 在亲代的双倍型d_{f_r}, d_{m_r}已知的条件下, 转移概率$\mathsf{P}(d_{c_r}|d_{f_r}, d_{m_r}) = 0, 1/4, 1/2$或1, 其中$d_{c_r}$表示子代的双倍型. 令$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_H)$, H表示单倍型总数, 我们的目的是要估计参数$\boldsymbol{\theta}$.

我们把观测不到的双倍型作为缺失数据(missing data)处理, 利用EM算法去估计单倍型概率$\boldsymbol{\theta}$. 因为我们假设I个家系是相互独立的, 于是对E步来说我们有

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{I} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{I} \mathsf{E}_{D_i}[\log \mathsf{P}(G_i, D_i)|G_i, \boldsymbol{\theta}^{(t)}],$$

其中$D_i = (d^{(1)}_{ia_1}, \cdots, d^{(1)}_{ia_{R_i}}, d^{*}_{ib_1p}, \cdots, d^{*}_{ib_{S_i}p})$, 又因为每个家系中各祖先个体之间也是相互独立的, 于是有

$$\begin{aligned}
&\mathsf{E}_{D_i}[\log \mathsf{P}(G_i, D_i)|G_i, \boldsymbol{\theta}^{(t)}] \\
\propto\ & \sum_{a_1=1}^{A_{i1}} \cdots \sum_{a_{R_i}=1}^{A_{iR_i}} \sum_{b_1=1}^{B_{i1}} \cdots \sum_{b_{S_i}=1}^{B_{iS_i}} \log \mathsf{P}(d^{(1)}_{ia_1}, \cdots, d^{(1)}_{ia_{R_i}}, d^{*}_{ib_1p}, \cdots, d^{*}_{ib_{S_i}p}) \\
&\times \mathsf{P}(d^{(1)}_{ia_1}, \cdots, d^{(1)}_{ia_{R_i}}, d^{*}_{ib_1}, \cdots, d^{*}_{ib_{S_i}}|G_i, \boldsymbol{\theta}^{(t)}) \\
=\ & \sum_{r=1}^{R_i} \sum_{a_r=1}^{A_{ir}} \log \mathsf{P}(d^{(1)}_{ia_r}) \mathsf{P}(d^{(1)}_{ia_r}|G_i, \boldsymbol{\theta}^{(t)}) + \sum_{s=1}^{S_i} \sum_{b_s=1}^{B_{is}} \log \mathsf{P}(d^{*}_{ib_sp}) \mathsf{P}(d^{*}_{ib_s}|G_i, \boldsymbol{\theta}^{(t)}) \\
=\ & \sum_{r=1}^{R_i} \sum_{a_r=1}^{A_{ir}} \omega_{ia_r,(t)} \log \mathsf{P}(d^{(1)}_{ia_r}) + \sum_{s=1}^{S_i} \sum_{b_s=1}^{B_{is}} \eta_{ib_s,(t)} \log \mathsf{P}(d^{*}_{ib_sp}).
\end{aligned}$$

下面我们分别来计算上式中的$\omega_{ia_r,(t)}$与$\eta_{ib_s,(t)}$, 其中

$$\omega_{ia_r,(t)} = \mathsf{P}(d^{(1)}_{ia_r}|G_i, \boldsymbol{\theta}^{(t)}) = \sum_{a_1=1}^{A_{i1}} \cdots \sum_{a_{r-1}=1}^{A_{i(r-1)}} \sum_{a_{r+1}=1}^{A_{i(r+1)}} \cdots \sum_{a_{R_i}=1}^{A_{iR_i}} \mathsf{P}(d^{(1)}_{ia_1}, \cdots, d^{(1)}_{ia_{R_i}}|G_i, \boldsymbol{\theta}^{(t)}),$$

$$\eta_{ib_s,(t)} = \mathsf{P}(d^{*}_{ib_s}|G_i, \boldsymbol{\theta}^{(t)}) = \sum_{b_1=1}^{B_{i1}} \cdots \sum_{b_{s-1}=1}^{B_{i(s-1)}} \sum_{b_{s+1}=1}^{B_{i(s+1)}} \cdots \sum_{b_{S_i}=1}^{B_{iS_i}} \mathsf{P}(d^{*}_{ib_1}, \cdots, d^{*}_{ib_{S_i}}|G_i, \boldsymbol{\theta}^{(t)}).$$

为了记号的简单, 我们令$p^* = \mathsf{P}(d^{(1)}_{ia_1}, \cdots, d^{(1)}_{ia_{R_i}}, d^{*}_{ib_1}, \cdots, d^{*}_{ib_{S_i}}, d^{(n)}_{ic_1}, \cdots, d^{(n)}_{ic_{T_i}}|\boldsymbol{\theta}^{(t)})$, 于是

$$\mathsf{P}(d^{(1)}_{ia_1}, \cdots, d^{(1)}_{ia_{R_i}}|G_i, \boldsymbol{\theta}^{(t)}) = \frac{\sum_{b_1=1}^{B_{i1}} \cdots \sum_{b_{S_i}=1}^{B_{iS_i}} \sum_{c_1=1}^{C_{i1}} \cdots \sum_{c_{T_i}=1}^{C_{iT_i}} p^*}{\sum_{a_1=1}^{A_{i1}} \cdots \sum_{a_{R_i}=1}^{A_{iR_i}} \sum_{b_1=1}^{B_{i1}} \cdots \sum_{b_{S_i}=1}^{B_{iS_i}} \sum_{c_1=1}^{C_{i1}} \cdots \sum_{c_{T_i}=1}^{C_{iT_i}} p^*},$$

$$\mathsf{P}(d^{*}_{ib_1}, \cdots, d^{*}_{ib_{S_i}}|G_i, \boldsymbol{\theta}^{(t)}) = \frac{\sum_{a_1=1}^{A_{i1}} \cdots \sum_{a_{R_i}=1}^{A_{iR_i}} \sum_{c_1=1}^{C_{i1}} \cdots \sum_{c_{T_i}=1}^{C_{iT_i}} p^*}{\sum_{a_1=1}^{A_{i1}} \cdots \sum_{a_{R_i}=1}^{A_{iR_i}} \sum_{b_1=1}^{B_{i1}} \cdots \sum_{b_{S_i}=1}^{B_{iS_i}} \sum_{c_1=1}^{C_{i1}} \cdots \sum_{c_{T_i}=1}^{C_{iT_i}} p^*}.$$

因此为了计算$\omega_{ia_r,(t)}$和$\eta_{ib_s,(t)}$, 我们只需要计算p^*,

$$\begin{aligned} p^* &= \mathsf{P}(d_{ia_1}^{(1)},\cdots,d_{ia_{R_i}}^{(1)},d_{ib_1}^{*},\cdots,d_{ib_{S_i}}^{*},d_{ic_1}^{(n)},\cdots,d_{ic_{T_i}}^{(n)}|\boldsymbol{\theta}^{(t)}) \\ &= \prod_{r=1}^{R_i}\mathsf{P}(d_{ia_r}^{(1)}|\boldsymbol{\theta}^{(t)})\prod_{s=1}^{S_i}\mathsf{P}(d_{ib_sp}^{*}|\boldsymbol{\theta}^{(t)})\prod_{l=1}^{N_i^{(n)}}\mathsf{P}(d_{ic_l}^{(n)}|d_{if_l},d_{im_l}), \end{aligned}$$

其中d_{if_l},d_{im_l}是表示与第l个非祖先个体双倍型$d_{ic_l}^{(n)}$相对应的亲代父母的双倍型. 对M步来说, 我们利用拉格朗日乘子法很容易得到下面的迭代公式,

$$\widehat{\theta}_x^{(t+1)}=\frac{1}{2N}\sum_{i=1}^{I}\Big(\sum_{r=1}^{R_i}\sum_{a_r=1}^{A_{ir}}\omega_{ia_r,(t)}\delta_{ia_r,x}+\sum_{s=1}^{S_i}\sum_{b_s=1}^{B_{is}}\eta_{ib_s,(t)}\varepsilon_{ib_s,x}\Big),$$

其中$\delta_{ia_r,x}$表示双倍型$d_{ia_r}^{(1)}$中含有单倍型x的个数, 其可能取值是0, 1, 2; $\varepsilon_{ib_s,x}$表示父母双倍型向量$d_{ib_sp}^{*}$中含有单倍型x的个数, 其可能取值是0, 1, $\cdots$, 4; $N=\sum_{i=1}^{I}(N_i^{(1)}+N_i^{(2)})$是$I$个家系中所有祖先个体总数.

3.3　个体单倍型推断规则

将已得到的单倍型概率估计记为$\widehat{\boldsymbol{\theta}}$, 下面按照最大似然的原则分别推断三类个体的单倍型:

1) 对于基因型完全已知的祖先个体来说, 比如其基因型为$G_{ir}^{(1)}$, 我们把与$G_{ir}^{(1)}$相匹配并且使$\mathsf{P}(d_{ia_r}^{(1)}=\{x,y\}|\widehat{\boldsymbol{\theta}})$达到最大的单倍型对$\{x,y\}$作为该祖先个体的双倍型;

2) 对于基因型有缺失的祖先个体来说, 因为我们把其放入某一核心家庭来考虑, 比如相应的核心家庭基因型向量为G_{is}^{*}, 我们首先确定与G_{is}^{*}相匹配的双倍型向量, 然后把使得$\mathsf{P}(d_{ib_sp}^{*}=(\{x,y\},\{x',y'\})|\widehat{\boldsymbol{\theta}})$达最大的双倍型向量$(\{x,y\},\{x',y'\})$作为核心家庭中相应父母的双倍型向量;

3) 对于所有的非祖先个体, 我们利用其亲代的双倍型来确定他的双倍型. 具体地, 因为我们假设没有重组发生, 子代的两条单倍型分别来自父亲和母亲, 于是按照孟德尔遗传法则根据父母的双倍型可以得到子代所有可能的双倍型, 然后再利用极大似然原则确定子代的双倍型.

§4.　计算参数估计标准差的Louis方法

根据Louis方法[13]我们去计算参数估计的标准差, 我们首先计算观测数据的信息阵$I(\boldsymbol{\theta})$, 由Louis的讨论我们知道观测信息阵$I(\boldsymbol{\theta})$可以分开来算, 即

$$I(\boldsymbol{\theta})=I_1(\boldsymbol{\theta})-I_2(\boldsymbol{\theta})+I_3(\boldsymbol{\theta}),$$

其中

$$
\begin{aligned}
I_1(\boldsymbol{\theta}) &= \sum_{i=1}^{I}\Big(\sum_{r=1}^{R_i}\sum_{a_r=1}^{A_{ir}}(-\omega_{ia_r})\frac{\partial^2\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}+\sum_{s=1}^{S_i}\sum_{b_s=1}^{B_{is}}(-\eta_{ib_s})\frac{\partial^2\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big),\\
I_2(\boldsymbol{\theta}) &= \sum_{i=1}^{I}\Big[\sum_{r=1}^{R_i}\sum_{a_r=1}^{A_{ir}}\omega_{ia_r}\frac{\partial\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big(\frac{\partial\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big)'\\
&\quad+\sum_{s=1}^{S_i}\sum_{b_s=1}^{B_{is}}\eta_{ib_s}\frac{\partial\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big(\frac{\partial\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big)'\Big],\\
I_3(\boldsymbol{\theta}) &= \sum_{i=1}^{I}\Big[\sum_{r=1}^{R_i}\Big(\sum_{a_r=1}^{A_{ir}}\omega_{ia_r}\frac{\partial\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big)\Big(\sum_{a_r=1}^{A_{ir}}\omega_{ia_r}\frac{\partial\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big)'\\
&\quad+\sum_{s=1}^{S_i}\Big(\sum_{b_s=1}^{B_{is}}\eta_{ib_s}\frac{\partial\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big)\Big(\sum_{b_s=1}^{B_{is}}\eta_{ib_s}\frac{\partial\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big)'\Big].
\end{aligned}
$$

根据Fisher公式我们可以得到

$$
\begin{aligned}
I_1(\boldsymbol{\theta}) &\approx -\sum_{i=1}^{I}\Big[\sum_{r=1}^{R_i}\sum_{a_r=1}^{A_{ir}}\omega_{ia_r}\frac{1}{\mathsf{P}^2(d_{ia_r}^{(1)})}\frac{\partial\mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big(\frac{\partial\mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big)'\\
&\quad+\sum_{s=1}^{S_i}\sum_{b_s=1}^{B_{is}}\eta_{ib_s}\frac{1}{\mathsf{P}^2(d_{ib_sp}^{*})}\frac{\partial\mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big(\frac{\partial\mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big)'\Big]\\
&= \sum_{i=1}^{I}\Big[\sum_{r=1}^{R_i}\sum_{a_r=1}^{A_{ir}}\omega_{ia_r}\frac{\partial\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big(\frac{\partial\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\boldsymbol{\theta}}\Big)'\\
&\quad+\sum_{s=1}^{S_i}\sum_{b_s=1}^{B_{is}}\eta_{ib_s}\frac{\partial\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big(\frac{\partial\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\boldsymbol{\theta}}\Big)'\Big]\\
&= I_2(\boldsymbol{\theta}).
\end{aligned}
$$

于是计算观测信息阵只需计算$I_3(\boldsymbol{\theta})$, 进一步地, 我们很容易求得

$$
\begin{aligned}
\frac{\partial\log \mathsf{P}(d_{ia_r}^{(1)})}{\partial\theta_x} &= \frac{\delta_{ia_r,x}}{\theta_x}-\frac{\delta_{ia_r,1}}{\theta_1},\\
\frac{\partial\log \mathsf{P}(d_{ib_sp}^{*})}{\partial\theta_x} &= \frac{\varepsilon_{ib_s,x}}{\theta_x}-\frac{\varepsilon_{ib_s,1}}{\theta_1},
\end{aligned}
$$

$i=1,\cdots,I$, $r=1,\cdots,R_i$, $a_r=1,\cdots,A_{ir}$, $s=1,\cdots,S_i$, $b_s=1,\cdots,B_{is}$, $x=2,\cdots,H$.

于是, 我们得到$\widehat{\boldsymbol{\theta}}$的渐近协方差阵的估计为

$$
I^{-1}(\widehat{\boldsymbol{\theta}})\approx\{I_3(\widehat{\boldsymbol{\theta}})\}^{-1}\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}},
$$

其中$\widehat{\boldsymbol{\theta}}$是$\boldsymbol{\theta}$的估计值.

§5. 模拟研究

本节通过模拟研究去验证我们所提方法的可行性. 尽管我们的方法理论上可以处理一般的家系结构, 为了说明问题的简单, 这里我们基于“祖代–父代–子代”三代的家系结构进行模拟研究, 同时我们假设三代中仅观测到一个个体的基因型, 我们称之为“祖–父–子”家系, 如图1中的个体1、3、5. 显然, 传统的EM方法不能直接推断“祖–父–子”家系结构的单倍型, 因此我们把祖代和父代中另一个观测不到的个体的基因型(如图1中的个体2和4)作为完全缺失数据来处理, 利用我们提出的方法去估计参数单倍型概率, 并推断个体1、3、5的单倍型.

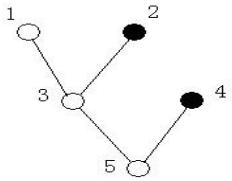

图1

5.1 模拟设计

首先介绍如何产生模拟数据. 我们从某一给定的真实的单倍型概率分布中随机地抽取600条单倍型, 接着两两配对得到300个个体的双倍型; 进一步我们把这300个个体随机地分成100组, 每组三个个体的双倍型分别代表图1所示家系中个体1、2、4的双倍型数据; 然后按照孟德尔遗传法则产生个体3和个体5的双倍型(假定没有重组发生); 最后我们扔掉个体2和个体4的双倍型, 同时忽略所有个体双倍型的连锁相信息, 从而得到100个“祖–父–子”家系的基因型模拟数据.

我们把个体2和个体4的基因型作为缺失数据, 基于得到的100个“祖–父–子”家系的模拟数据利用我们的方法去估计参数单倍型概率, 并且与给定的真实的单倍型的概率进行比较. 同时, 根据得到的单倍型概率估计, 按照极大似然原则去推断个体的双倍型. 为了评价个体双倍型推断的准确度, 我们引进招回率(call rate)[7, 9]为

$$\mathrm{CR} = \frac{N_{\mathrm{true}}}{N},$$

其中N表示100个“祖–父–子”家系中个体的总数, N_{true}表示双倍型被推断正确的个体数. 我们总共模拟1000次, 计算1000次模拟中每条单倍型概率估计的平均值以及1000次模拟的平均招回率. 因为我们的方法还能够计算参数估计的标准差, 为了进一步说明所提方法的正确性, 我们给出了95%的经验覆盖概率(empirical coverage probability).

5.2 模拟结果

为了便于说明问题, 我们考虑2个SNPs位点的模拟研究, 因为SNPs位点是人类染色体上最常见的多态性位点. 又因为每个SNP位点上有两种可能的等位基因, 不妨记为1和2, 因此共有4种可能的单倍型: (1 1)、(1 2)、(2 1)和(2 2), 相应的单倍型概率记为: θ_1、θ_2、θ_3和θ_4. 我们分别考虑三种不同的单倍型概率分布: (1): 0.25, 0.25, 0.25, 0.25; (2): 0.40, 0.30, 0.20, 0.10; (3): 0.40, 0.10, 0.10, 0.40.

表1中给出了基于三种不同单倍型概率分布的模拟结果. 从表1我们可以看出, 在三种不同的单倍型概率分布情况下, 利用我们的方法得到的单倍型概率的估计均接近真实的单倍型概率, 这表明我们的方法确实能够有效的处理家系中某些个体基因型有缺失情况下的单倍型推断问题, 同时我们还给出了1000次模拟结果的标准差. 为了进一步验证我们方法的可行性, 我们通过计算1000次模拟的平均招回率来衡量对个体双倍型推断的准确性, 结果见表2, 三种情况下个体单倍型的推断精度均超过了90%.

因为我们的方法还能够计算参数估计的标准差, 因此我们可以构造参数θ_i的95%的置信区间$(\widehat{\theta}_i \pm 1.96\widehat{\sigma}_i)$, $i = 2, 3, 4$ (因为$\sum_{i=1}^{4} \theta_i = 1$), 其中$\widehat{\theta}_i$是$\theta_i$的估计, 而$\widehat{\sigma}_i$是$\widehat{\theta}_i$的标准差的估计. 表3中给出了1000次模拟中$\theta_i$ $(i = 2, 3, 4)$的95%的经验覆盖概率. 从结果可以看出, 经验覆盖率均接近95%, 说明我们的方法在实际中是可行的.

表1 基于“祖–父–子”家系数据的单倍型概率估计

	单倍型			
	(1 1)	(1 2)	(2 1)	(2 2)
真实概率分布(1)	0.25	0.25	0.25	0.25
估计概率 (标准差)	0.2508 (0.0006)	0.2495 (0.0005)	0.2503 (0.0006)	0.2494 (0.0005)
真实概率分布(2)	0.40	0.30	0.20	0.10
估计概率 (标准差)	0.4025 (0.0007)	0.2975 (0.0006)	0.1990 (0.0005)	0.1010 (0.0003)
真实概率分布(3)	0.40	0.10	0.10	0.40
估计概率 (标准差)	0.3993 (0.0007)	0.1009 (0.0002)	0.1017 (0.0002)	0.3981 (0.0008)

表2 1000次模拟的平均招回率

分布种类	(1)	(2)	(3)
招回率	95.6%	97.9%	92.3%

表3　θ_2、θ_3和θ_4的95%的经验覆盖概率

	95%的经验覆盖概率		
分布种类	θ_2	θ_3	θ_4
(1)	96.0%	95.5%	95.5%
(2)	94.6%	94.9%	94.8%
(3)	96.3%	95.6%	93.1%

§6. 总　　结

本文针对家系中某些个体的基因型部分缺失或者是完全未知的情况提出了一种推断个体单倍型的EM算法, 并且通过基于"祖–父–子"家系数据的模拟研究评价了我们方法的可行性. 但是我们没有考虑亲代和子代的基因型同时有缺失的情况, 这也正是我们以后要进一步考虑的问题. 另外当位点数目比较多时, 计算过程中占用的计算机内存会非常大, 并且计算时间也是惊人的, 此时我们可以利用PL方法[14]先将染色体分成小的片段(包含5-7位点)逐个进行研究, 然后再把这些片段整合成整个单倍型. 同时为了减少计算量, 可以考虑利用基于规则的rule-based方法和单倍型消去法[15–17], 尽量减少参与E步计算的双倍型, 这样处理不仅解决了计算机存储的问题, 还可以提高计算的效率和精度.

参 考 文 献

[1] Clark, A.G., Inference of haplotypes from PCR-amplified samples of diploid populations, *Mol. Biol. Evol.*, **7**(1990), 111–122.

[2] Excoffier, L. and Slatkin, M., Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, *Mol. Biol. Evol.*, **12**(1995), 921–927.

[3] Dempster, A.P., Laird, N.M. and Rubin, D.B., Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Stat. Soc. Ser. B*, **39**(1977), 1–38.

[4] Zhu, W.S. and Guo, J.H., A likelihood-based method for haplotype association studies for case-control data with genotyping uncertainty, *Sci. China Ser. A*, **49**(2006), 130–144.

[5] 朱文圣, 郭建华, 病例－对照研究中基因型不确定时单倍型关联分析的似然方法, 中国科学A辑, **36**(2006), 403–417.

[6] Zhang, K., Sun, F. and Zhao, H., HAPLOE: a program for haplotype reconstruction in general pedigrees without recombination, *Bioinfoematics*, **21**(2005), 90–103.

[7] Becker, T. and Knapp, M., Maximum-likelihood estimation of haplotype frequencies in nuclear families, *Genet. Epid.*, **27**(2004), 21–32.

[8] Zou, G.H. and Zhao, H.Y., Haplotype frequency estimation in the presence of genotyping errors, *Hum. Hered.*, **56**(2003), 131–138.

[9] Zhu, W.S., Fung, W.K. and Guo, J.H., Incorporating genotyping uncertainty in haplotype frequency estimation in pedigree studies, *Hum. Hered.*, **64**(2007), 172–181.

[10] Ding, X.D., Zhang, Q., Flury, C. and Henner, S., Haplotype reconstruction and estimation of haplotype frequencies from nuclear families with only one parent available, *Hum. Hered.*, **62**(2006), 12–19.

[11] Liu, P.Y., Lu, Y. and Deng, H.W., Accurate haplotype inference for multiple linked single-nucleotide polymorphisms using sibship data, *Genetics*, **174**(2006), 499–509.

[12] Niu, T., Qin, Z.S., Xiu, X. and Liu, J.S., Bayesian haplotype inference for mulitiple linked SNPs, *Am. J. Hum. Genet.*, **70**(2002), 157–169.

[13] Louis, T.A., Finding the observed information matrix when using the EM algorithm, *J. Roy. Stat. Soc. Ser. B*, **44**(1982), 226–233.

[14] Qin, Z.S., Niu, T. and Liu, J.S., Partitioning-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide Ploymorphisms, *Am. J. Hum. Genet.*, **71**(2002), 1242–1247.

[15] O'Connell, J.R. and Weeks, D.E., An optimal algorithm for automatic genotype elimination, *Am. J. Hum. Genet.*, **65**(1999), 1733–1740.

[16] Lander, E.S. and Green, P., Construction of mulitilocus genetic linkage maps in humans, *Proc. Natl. Acad. Sci. USA*, **84**(1987), 2363–2367.

[17] Lange, K. and Goradia, T.M., An algorithm for automatic genotype elimination, *Am. J. Hum. Genet.*, **40**(1987), 250–256.

EM Algorithm for Haplotype Inference Incorporating Ungenotyped Individuals in Pedigree Studies

Zhao Hong[1,2] Zhu Wensheng[1] Guo Jianhua[1]

([1]*KLAS and School of Mathsmatics and Statistics, Northeast Normal University, Changchun, 130024*)

([2]*Department of Mathsmatics, Ocean University of China, Qingdao, 266071*)

Haplotype inference and reconstruction have become an essential step in linkage analysis and association analysis. The methods for haplotype inference are all based on the complete genotype data sets. However, genotypes of some individuals in each pedigree may be partly missing or ungenotyped. In this article, we propose a new EM algorithm to perform haplotype inference incorporating ungenotyped individuals in pedigree studies. We also give the standard errors of estimated parameters and evaluate the performance of our method by simulation studies.

Keywords: Haplotype inference, pedigree, EM algorithm.

AMS Subject Classification: 62P10.