中国科学 : 数学　2010 年　第 40 卷　第 10 期 : 971 ∼ 984
www.springerlink.com　math.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

约束下多子女家系数据重组率的最大似然估计

周影[①②], 韩国牛[③④], 史宁中[①], 冯荣锦[⑤], 郭建华[①*]

① 应用统计教育部重点实验室, 东北师范大学数学与统计学院, 长春 130024;
② 黑龙江大学数学科学学院, 哈尔滨 150080;
③ 南开大学组合数学中心, LPMC, 天津 300071;
④ 法国国家科研中心, 斯特拉斯堡, 法国;
⑤ 香港大学统计与精算系, 香港
E-mail: zhouy577@yahoo.com.cn, guoniu@math.u-strasbg.fr, shinz@nenu.edu.cn, wingfung@hku.hk, jhguo@nenu.edu.cn

收稿日期: 2007-09-24; 接受日期: 2010-02-04; * 通信作者
国家自然科学基金 (批准号: 10431010, 10701022)、国家重点基础研究发展计划 (973 计划) (批准号: 2007CB311002)、新世纪优秀人才支持计划 (批准号: NCET-04-0310)、教育部优秀青年教师和吉林省杰出青年科学研究基金 (批准号: 20030113)、大学创新研究团队 (PCSIRT) (批准号: #IRT0519) 和数学天元基金 (批准号: 10926174) 资助项目

摘要　本文针对相型信息未知的三回交家系, 讨论了在自然的序约束下重组率的估计问题. 考虑了多后代数据的后代表型分类问题, 给出了后代表型分类数的一个具体公式. 基于表型分类所得数据, 采用约束 EM 算法 (REM) 估计了两位点重组率. 鉴于交换干扰的存在可能会影响到基因定位的精度, 基于该估计, 进一步考虑了有关生物体基因组中交换干扰的统计推断问题. 实例和模拟研究均显示 REM 算法要优于无约束算法, 并证实了多后代家庭会提供更多连锁信息这一观点.

关键词　约束参数问题　连锁分析　重组率　约束 EM 算法
MSC (2000) 主题分类　97K80, 92D99

1　引言

连锁分析在分子遗传学中发挥着重要作用, 它指的是给染色体上的基因位点定序, 并且估计它们之间的遗传距离, 这里的距离由一种统计现象所决定. 当前, 人们已经在许多生物体的基因组上建立了高密度的连锁图用以分析家系数据和探测连锁 [1−6]. 连锁的程度可以由重组率, 即一次重组发生的概率来衡量. 具体来讲, 重组率是指由一个双杂合型的亲本所潜在产生的重组单倍型 (或后代) 的比例. 借助 Haldane[7]、Morgan[8] 和 Felsenstein[9] 所提出的图谱函数, 可以使重组率和遗传距离相互转化. 某些遗传性状的遗传变异通常由多个基因所控制, 要想找到控制这些性状的基因就需要探测一个基因与已知的标记位点之间的连锁. 因此精确地估计重组率在基因定位中非常重要. 同时, 影响基因定位的那些主要因素必须相应地予以考虑.

研究人员已经证实, 三位点分析在连锁分析中非常重要 [5,10−14], 因为它是多位点分析的基础. Lathrop 等 [15] 及 Lathrop[16] 提出了探测多位点连锁和估计重组的方法. Kruglyak 等 [4] 描绘了怎样从中等大小的一般家系中去提取完全的多位点遗传信息. 然而, 随着科技的进步, 需要新的方法与之同行. 近年来, Ott[5] 针对每家有两个后代的、相型信息未知的回交家庭给出了两位点重组率的估计方

引用格式: Zhou Y, Han G N, Shi N Z, et al. Maximum likelihood estimates of recombination fractions under restrictions for family data with multiple offspring (in Chinese). Sci Sin Math, 2010, 40(10): 971–984, doi: 10.1360/012007-482

法. 遗憾的是, 在 Ott 的分析中没有充分地考虑到两位点重组率应当满足的一些自然的约束, 因此可能会出现不合理的估计结果, 从而给实际研究带来一定的负面影响 [5,17].

在方差分量分析 [18]、约束最小二乘回归 [19] 和有序数据的统计建模 [20] 中均可见到约束参数问题. Roberston 等 [21] 详细地讨论了序约束下的统计推断. 最近, Liu[22] 和 Shi 等 [23] 利用 EM 算法得到了约束参数的最大似然估计. Zhou 等 [17] 重新考虑了文献 [5] 中的问题, 针对两后代、相型信息未知的三回交家庭, 提出了在一些自然的序约束下估计两位点重组率的约束 EM 算法 (REM). 所得结论是 REM 算法比文献 [5] 中的无约束算法具有优势. 值得注意的是, 每个家庭中后代越多会给连锁分析提供越多的信息, 而且不同的观测家庭中的孩子数目可能还会不同, 这就需要更为有效的方法把观测分到相应的表型类中, 之后才能使用 REM 算法进行参数估计. 因此, 本研究的目标是把文献 [17] 中的方法推广至针对每个家庭有多个后代的数据情形的参数估计方法, 并在获得重组率的估计之后, 进一步讨论基因组中的交换干扰问题.

我们首先考虑了多后代家庭的表型分类, 提出了一个后代表型分类数的具体公式. 其次, 我们演示了如何利用 REM 算法去处理通过表型分类所获得的家庭数据. 再次, 基于所获得的估计, 我们考虑了有关干扰参数的推断问题. 最后, 我们把算法应用到了真实数据和模拟数据中. 我们发现 REM 算法要优于无约束方法, 并且它是稳健和有效的. 我们同时也通过模拟考查了样本量 (家庭数), 以及家庭中后代数目对估计的影响情况.

2 记号、参数的自然的不等式约束

考虑三回交家庭, A, B 和 C 表示三个有序的位点 (顺序为 A-B-C), 它们各自的等位基因分别为 A, a; B, b 以及 C, c. 家庭中三纯和型亲本的基因型为 abc/abc, 三杂合型的亲本基因型可表示为 $Aa/Bb/Cc$, 在连锁平衡的假设下其具有四种等可能的连锁相 (相型): (I) ABC/abc, (II) ABc/abC, (III) AbC/aBc, (IV) Abc/aBC.

这里的主要记号同文献 [17] 中的记号相同: θ_{AB}, θ_{BC} 和 θ_{AC} 分别表示三个两位点重组率; $g_{ij}\,(i,j=0,1)$ 表示联合重组率, 其中下标 1 表示相应的区间内发生重组, 0 则表示非重组. 例如: g_{10} 表示在区间 AB 内发生重组而在 BC 内不发生重组的概率. 根据概率论的知识可得, 两类重组率间有如下的等式关系:

$$\theta_{AB} = g_{11} + g_{10}, \quad \theta_{BC} = g_{11} + g_{01}, \quad \theta_{AC} = g_{10} + g_{01}. \tag{1}$$

表 1 给定杂合型亲本连锁相时产生各种单倍型的条件概率

		连锁相			
i	单倍型	I	II	III	IV
1	ABC	$0.5g_{00}$	$0.5g_{01}$	$0.5g_{11}$	$0.5g_{10}$
2	ABc	$0.5g_{01}$	$0.5g_{00}$	$0.5g_{10}$	$0.5g_{11}$
3	AbC	$0.5g_{11}$	$0.5g_{10}$	$0.5g_{00}$	$0.5g_{01}$
4	Abc	$0.5g_{10}$	$0.5g_{11}$	$0.5g_{01}$	$0.5g_{00}$
5	aBC	$0.5g_{10}$	$0.5g_{11}$	$0.5g_{01}$	$0.5g_{00}$
6	aBc	$0.5g_{11}$	$0.5g_{10}$	$0.5g_{00}$	$0.5g_{01}$
7	abC	$0.5g_{01}$	$0.5g_{00}$	$0.5g_{10}$	$0.5g_{11}$
8	abc	$0.5g_{00}$	$0.5g_{01}$	$0.5g_{11}$	$0.5g_{10}$
合计		1	1	1	1

在回交家庭中, 三纯和型的亲本只产生单倍型 abc, 而三杂合型亲本则可能产生 8 种可能的单倍型 (见表 1 的第 2 列). 为了方便, 我们分别用表 1 中第 1 列的代号去表示它们. 在显性遗传的假设下, 这 8 种单倍型恰与后代的表现性相互对应. 在给定杂合型亲本连锁相时产生各种单倍型的条件概率均列于表 1 的最后 4 列.

鉴于参数约束在参数估计中的重要性, 我们回顾一下两位点重组率应当满足的一些自然而又必要的不等式约束. 首先, $\theta_{AB} \leqslant \theta_{BC} + \theta_{AC}$, $\theta_{BC} \leqslant \theta_{AB} + \theta_{AC}$, $\theta_{AC} \leqslant \theta_{AB} + \theta_{BC}$ 和 $0 \leqslant \theta_i \leqslant 1/2\,(i = AB, BC, AC)$ 是必须满足的; 其次, 对于给定的位点顺序 A-B-C, 此时 $\theta_{AB} \leqslant \theta_{AC}$, $\theta_{BC} \leqslant \theta_{AC}$ 也应满足. 我们把这些不等式整理成了下述形式 [17]:

$$\begin{cases} \theta_{AB} \leqslant \theta_{AC}, \\ \theta_{BC} \leqslant \theta_{AC}, \\ \theta_{AC} \leqslant \theta_{AB} + \theta_{BC}, \\ \theta_{AC} \leqslant 1/2. \end{cases} \tag{2}$$

由于 $(\theta_{AB}, \theta_{BC}, \theta_{AC})$ 与 (g_{10}, g_{01}, g_{11}) 是两组等价的参数 (见等式 (1)), 我们这里仅考虑后者, 同时上面的约束 (2) 变成如下的约束 (3):

$$\begin{cases} g_{11} \leqslant g_{01}, \\ g_{11} \leqslant g_{10}, \\ g_{11} \geqslant 0, \\ g_{01} + g_{10} \leqslant 1/2. \end{cases} \tag{3}$$

这样, 上面的三位点分析中的参数估计问题就变为寻找 $\mathbf{g} = (g_{10}, g_{01}, g_{11})$ 的约束的 MLE $\hat{\mathbf{g}}^R$, 使其在约束 (3) 下满足 $l(\hat{\mathbf{g}}^R) = \max_{\mathbf{g}} l(\mathbf{g})$.

3 表型分类及 REM 算法

在统计遗传学中, 样本量是影响连锁分析的一个重要因素. 注意到每个观测家庭中后代越多给连锁分析提供的信息也越多这一观点 [5,11], 在这一节中我们将把文献 [17] 中的 REM 算法推广至每家有多个后代的情形. 然而, 随着后代数的增多, 表型分类又变得相对复杂. 下面我们首先解决这个困难.

3.1 表型分类的方法

在每个家庭中两后代单倍型对出现的联合概率容易算得. 具有相同概率的单倍型对可按连锁分析规则分入一类, 因而每个家庭有两个后代时表型总共有四类 [5].

随着每家中后代数的增加, 单倍型组合的数目也相应增加. 例如, 当每个观测家庭中有三个后代时, 总共有 120 种单倍型的组合. 当收集表型数据时, 首先就应确定总共的表型类数. 令每个家庭中的后代数为 r, 相应的表型分类数为 $f(r)$. 这里我们考虑如何将任一个包含 r 条单倍型的单倍型组合分入某一类, 并给出 $f(r)$ 关于 r 的表达式.

我们定义四个等价类: $E_1 = \{1, 8\}, E_2 = \{2, 7\}, E_3 = \{3, 6\}$ 以及 $E_4 = \{4, 5\}$. 这是由于单倍型 1 和 8 在给定任一种相型时出现的概率都是相等的, 即它们在所有的单倍型中处于相同的地位, 其它三对的情况也是如此. 在同一等价类中的单倍型称作是等价的. 这样为了方便, 我们仅考虑单倍型 1, 2,

3 和 4 即可. 令 x_1, x_2, x_3 和 x_4 分别表示 $0.5g_{00}$, $0.5g_{01}$, $0.5g_{11}$ 和 $0.5g_{10}$. 事实上, 表 1 中的条件概率是 $\{x_1, x_2, x_3, x_4\}$ 的某个对称置换. 令 $\{x_{ij}\}$ 表示这个置换.

进一步, 当每个家庭中后代数为 r 时, 每一个表型类概率恰好是 $(\sum_{i=1}^4 x_i)^r$ 的展开式中某些项之和, 并且在 $\sum_{j=1}^4(\sum_{i=1}^4 x_{ij})^r$ 的展开式中形如 $\sum_{j=1}^4\prod_{i=1}^4 x_{ij}^{r_i}$ 的每一项均对应于每一个表型类概率 $\frac{r!}{r_1!r_2!r_3!r_4!}(x_1^{r_1}x_2^{r_2}\ x_3^{r_3}x_4^{r_4}+x_2^{r_1}x_1^{r_2}x_4^{r_3}x_3^{r_4}+x_3^{r_1}x_4^{r_2}x_1^{r_3}x_2^{r_4}+x_4^{r_1}x_3^{r_2}x_2^{r_3}x_1^{r_4})$, 这种概率形式正说明了这个类中任一个单倍型组合必然包含 r_1 条 E_1 中的单倍型, r_2 条 E_2 中的单倍型, r_3 条 E_3 中的单倍型, 和 r_4 条 E_4 中的单倍型; 或者是 r_1 条 E_2 中的单倍型, r_2 条 E_1 中的单倍型, r_3 条 E_4 中的单倍型和 r_4 条 E_3 中的单倍型等等, 这里 $\sum_{i=1}^4 r_i = r$. 这样所有表型类的总数 $f(r)$ 恰好就等于 $\sum_{j=1}^4(\sum_{i=1}^4 x_{ij})^r$ 的展开式中不同项 $\sum_{j=1}^4\prod_{i=1}^4 x_{ij}^{r_i}$ 的数目. 下面的定理给出了 $f(r)$ 的表达式.

定理 若 $\{x_{ij}\}$ 是表 1 中所给出的 $\{x_1, x_2, x_3, x_4\}$ 的对称置换, 则

$$\sum_{j=1}^4\left(\sum_{i=1}^4 x_{ij}\right)^r = \sum_{r_1+r_2+r_3+r_4=r}\binom{r}{r_1\ r_2\ r_3\ r_4}\left[\sum_{j=1}^4\prod_{i=1}^4 x_{ij}^{r_i}\right]$$

的展开式中不同项 $\sum_{j=1}^4\prod_{i=1}^4 x_{ij}^{r_i}$ 的数目, 即

$$f(r)=\begin{cases}\dfrac{(r+1)(r+2)(r+3)}{24}, & \text{如果 } r \text{ 是奇数},\\ \dfrac{(r+2)[(r+1)(r+3)+9]}{24}, & \text{如果 } r \text{ 是偶数}.\end{cases}$$

为证明此定理, 我们首先给出组合数学中一个著名的引理.

引理 (Polya 计数定理, 参见文献 [24, p. 123]) 令 D 和 A 为两个 (有限) 集合, G 为作用在 D 上的一个置换群. 若对任意的 $D\to A$ 的函数 ϕ_1, ϕ_2, 存在一个 $g\in G$, 使得 $\phi_1(g)=\phi_2$, 则称函数 ϕ_1, ϕ_2 是 G 等价的. 令每一个 $a\in A$ 都有一个权, 即 $w(a)=\alpha_a$. G 等价的函数具有相同的权, 在等价关系下 ϕ 所在的等价类记为 Φ, 且 $w(\Phi)=w(\phi)$, 则有

$$\sum_{\Phi} w(\Phi) = P\left(G; \sum_{a\in A} w(a), \sum_{a\in A} w^2(a), \dots\right),$$

如果用 $\lambda_i(g)$ 表示 g 中长为 i 的轮换个数, 则多项式 $P(G; y_1, \dots, y_n) = \frac{1}{|G|}\sum_{g\in G} y_1^{\lambda_1(g)}y_2^{\lambda_2(g)}\cdots y_n^{\lambda_n(g)}$.

引理的证明可参见文献 [24], 下面我们证明定理.

定理的证明 在上面的引理中取 $D=\{1,2,3,4\}$, $A=\{0,1,2,\dots,r\}$. 对任意的 $a\in A$, 令 $w(a)=z^a$. 令 M_2 表示由对称置换 $\{x_{ij}\}$ 的下标构成的表, 并令 G 表示由 M_2 的行构成的集合, 这里

$$M_2=\begin{vmatrix}1&2&3&4\\2&1&4&3\\3&4&1&2\\4&3&2&1\end{vmatrix}.$$

群 G 中包含了两种置换, 一种是恒等置换, 即 M_2 的第一行:

$$(1)(2)(3)(4);$$

另一种是由 2 元轮换构成的置换

$$(12)(34)(13)(24)(14)(23).$$

由上面引理我们知道,

$$P(G)=\frac{1}{4}[(1+z+z^2+\cdots+z^r)^4+3(1+z^2+z^4+\cdots+z^{2r})^2]$$

中 z^r 的系数等于 $\sum_{j=1}^4(\sum_{i=1}^4 x_{ij})^r$ 的展开式中不同项 $\sum_{j=1}^4\prod_{i=1}^4 x_{ij}^{r_i}$ 的个数.

因此,

$$\begin{aligned}f(r)&=[z^r]\left(\frac{1}{4}[(1+z+z^2+\cdots+z^r)^4+3(1+z^2+z^4+\cdots+z^{2r})^2]\right)\\&=[z^r]\left(\frac{1}{4}\left[\left(\frac{1-z^{r+1}}{1-z}\right)^4+3\left(\frac{1-z^{2r+2}}{1-z^2}\right)^2\right]\right)\\&=[z^r]\left(\frac{1}{4}\left[\left(\frac{1}{1-z}\right)^4+3\left(\frac{1}{1-z^2}\right)^2\right]\right)\\&=[z^r]\left(\frac{1}{4}\left[\sum_{k\geqslant 0}\binom{4+k-1}{k}z^k+3\sum_{k\geqslant 0}\binom{2+k-1}{k}z^{2k}\right]\right)\\&=\begin{cases}\dfrac{(r+1)(r+2)(r+3)}{24}, & \text{如果 } r \text{ 是奇数},\\[2mm]\dfrac{(r+2)[(r+1)(r+3)+9]}{24}, & \text{如果 } r \text{ 是偶数},\end{cases}\end{aligned}$$

其中 $[z^r](\cdot)$ 表示括号里表达式中 z^r 的系数. □

评论 南开大学陈永川教授等人曾指出一个更一般的结果, 即在 $\sum_{j=1}^{2^m}(\sum_{i=1}^{2^m}x_{ij})^r$ 的展开式中不同项 $\sum_{j=1}^{2^m}\prod_{i=1}^{2^m}x_{ij}^{r_i}$ 的数目情况: 若 r 是奇数, 该数目等于 $\frac{1}{2^m}\binom{2^m+r-1}{r}$; 否则, 等于 $\frac{1}{2^m}\binom{2^m+r-1}{r}+\frac{2^m-1}{2^m}\binom{2^m+\frac{r}{2}-1}{\frac{r}{2}}$.

3.2 针对多子女家系数据的 REM 算法

这里我们仅对三后代家庭 ($r=3$) 的情形来演示 REM 算法 (更多后代的情形完全类似). 对于这种情形, 我们把观测家庭分成 $f(3)=5$ 类, 具体见表 2.

令此时观测家庭的总数为 m, 分入第 k 类的家庭数为 $m_k\,(k=1,2,3,4,5)$, 则 $\sum_{k=1}^5 m_k=m$, 并且 $(m_1,m_2,m_3,m_4,m_5)\sim\text{Multi}(m,q_1,q_2,q_3,q_4,q_5)$. 我们不妨给出在这种情况下类似文献 [5] 中的重组率的无约束估计. 由于 $q_1+q_2=(g_{11}+g_{10})^3+(g_{01}+g_{00})^3=\theta_{AB}^3+(1-\theta_{AB})^3=3\theta_{AB}^2-3\theta_{AB}+1$, 解这个方程就可以得到 θ_{AB} 的表达式，从而获得无约束的 MLE: 如果 $\hat{q}_1+\hat{q}_2>1/4$, 则 $\hat{\theta}_{AB}^U=1/2-1/6\sqrt{12(\hat{q}_1+\hat{q}_2)-3}$, 否则 $\hat{\theta}_{AB}^U=1/2$. 类似的, $\hat{\theta}_{BC}^U$ 可以由 $\hat{q}_1+\hat{q}_3$ 决定, $\hat{\theta}_{AC}^U$ 可以由 $\hat{q}_1+\hat{q}_4$ 决定 (称其为无约束方法).

然而正如前面所提到的, 无约束的 MLE 有时候可能会不合理, 为了克服这个局限, 我们还是采用约束的最大似然方法 (即 REM 算法). 经过数据扩充之后, 我们得到完全数据 $\{m_{kj},\ k=1,2,3,4,5, j=1,2,3,4\}$. 进一步，完全数据对数似然的条件期望为

$$Q(\mathbf{g}\mid\mathbf{g}^{(s)},\{m_k\})=3[b_1^{(s+1)}\ln(1-g_{01}-g_{10}-g_{11})+b_2^{(s+1)}\ln(g_{01})+b_3^{(s+1)}\ln(g_{10})+b_4^{(s+1)}\ln(g_{11})],$$

这里 $\mathbf{g}^{(s)}$ 表示 $\mathbf{g}$ 的当前估计, $\mathbf{b}^{(s+1)}=(b_1^{(s+1)},b_2^{(s+1)},b_3^{(s+1)},b_4^{(s+1)})$ 的具体表达式列于附录中. 类似于文献 [17] 中用于两后代情形的算法, 下面我们给出针对三后代家庭数据估计重组率的约束 EM 算法:

取 $\mathbf{g}^{(0)}=(g_{10}^{(0)},g_{01}^{(0)},g_{11}^{(0)})$ 为算法的初始值;

表 2　相型未知的三回交家庭每家有三个后代时的表型分类

k	ijl^a	q_k
1	111,222,333,444,555,666,777,888,118,881,227,772,336,663,445,554	$g_{11}^3+g_{10}^3+g_{01}^3+g_{00}^3$
2	112,117,882,887,122,177,822,877, 334,335,664,665,344,355,644,655	$3(g_{00}^2g_{01}+g_{01}^2g_{00}$
	182,187,271,278,364,365,453,456	$+g_{11}^2g_{10}+g_{10}^2g_{11})$
3	223,226,773,776,233,266,733,766,114,115,884,885,144,155,844,855	$3(g_{01}^2g_{11}+g_{00}^2g_{10}$
	273,276,362,367,184,185,451,458	$+g_{11}^2g_{01}+g_{10}^2g_{00})$
4	113,116,883,886,133,166,833,866,224,225,774,775,244,255,744,755	$3(g_{00}^2g_{11}+g_{01}^2g_{10}$
	183,186,361,368,274,275,452,457	$+g_{11}^2g_{00}+g_{10}^2g_{01})$
5	123,126,124,125,173,176,174,175, 823,826,824,825,873,876,874,875	$6(g_{00}g_{01}g_{11}+g_{01}g_{00}g_{10}$
	134,135,164,165,834,835,864,865, 234,235,264,265,734,735,764,765	$+g_{11}g_{10}g_{00}+g_{10}g_{11}g_{01})$
合计		1

[a] ijl: i, j 和 l 是指表 1 中的单倍型代码, 这里可解释为表型.

E 步: 在第 s 步迭代时, 由 $\mathbf{g}^{(s)}$ 计算期望的重组事件的数目 $\mathbf{b}^{(s+1)}$;

M 步: 由 $\mathbf{b}^{(s+1)}$ 计算 $\mathbf{g}^{(s+1)}$. 首先计算 $\tilde{\mathbf{g}}^{(s+1)}=(\tilde{g}_{10}^{(s+1)},\ \tilde{g}_{01}^{(s+1)},\ \tilde{g}_{11}^{(s+1)})=(b_3^{(s+1)}/m,\ b_2^{(s+1)}/m,\ b_4^{(s+1)}/m)$. 如果 $\tilde{\mathbf{g}}^{(s+1)}$ 满足约束 (3), 则 $\mathbf{g}^{(s+1)}=\tilde{\mathbf{g}}^{(s+1)}$; 否则 $\mathbf{g}^{(s+1)}$ 必然是下列情形之一 (即只有一种情形满足):

情形 1　$g_{00}^{(s+1)}=g_{01}^{(s+1)}=g_{10}^{(s+1)}=g_{11}^{(s+1)}=1/4$, 如果下面的不等式同时成立:

$$\begin{cases} b_3^{(s+1)}+b_4^{(s+1)}>b_1^{(s+1)}+b_2^{(s+1)},\\ b_2^{(s+1)}+b_4^{(s+1)}>b_1^{(s+1)}+b_3^{(s+1)},\\ b_2^{(s+1)}+b_3^{(s+1)}+b_4^{(s+1)}>3b_1^{(s+1)}; \end{cases}$$

情形 2　$g_{01}^{(s+1)}=g_{11}^{(s+1)}=g_{10}^{(s+1)}=\frac{b_2^{(s+1)}+b_3^{(s+1)}+b_4^{(s+1)}}{3m}, g_{00}^{(s+1)}=\frac{b_1^{(s+1)}}{m}$, 如果

$$\begin{cases} b_3^{(s+1)}+b_4^{(s+1)}>2b_2^{(s+1)},\\ b_2^{(s+1)}+b_4^{(s+1)}>2b_3^{(s+1)},\\ 3m/4\geqslant b_2^{(s+1)}+b_3^{(s+1)}+b_4^{(s+1)}>0; \end{cases}$$

情形 3　$g_{01}^{(s+1)}=g_{11}^{(s+1)}=\frac{b_2^{(s+1)}+b_4^{(s+1)}}{2m}, g_{10}^{(s+1)}=g_{00}^{(s+1)}=\frac{b_1^{(s+1)}+b_3^{(s+1)}}{2m}$, 如果

$$\begin{cases} b_3^{(s+1)}b_4^{(s+1)}>b_1^{(s+1)}b_2^{(s+1)},\\ b_3^{(s+1)}>b_1^{(s+1)},\\ b_1^{(s+1)}+b_3^{(s+1)}\geqslant b_2^{(s+1)}+b_4^{(s+1)}>0; \end{cases}$$

情形 4　$g_{10}^{(s+1)}=g_{11}^{(s+1)}=\frac{b_3^{(s+1)}+b_4^{(s+1)}}{2m}, g_{01}^{(s+1)}=g_{00}^{(s+1)}=\frac{b_1^{(s+1)}+b_2^{(s+1)}}{2m}$, 如果

$$\begin{cases} b_2^{(s+1)}b_4^{(s+1)}>b_1^{(s+1)}b_3^{(s+1)},\\ b_2^{(s+1)}>b_1^{(s+1)},\\ b_1^{(s+1)}+b_2^{(s+1)}\geqslant b_3^{(s+1)}+b_4^{(s+1)}>0; \end{cases}$$

情形 5 $g_{01}^{(s+1)}=g_{11}^{(s+1)}=\frac{b_2^{(s+1)}+b_4^{(s+1)}}{2m}, g_{10}^{(s+1)}=\frac{b_3^{(s+1)}}{m}, g_{00}^{(s+1)}=\frac{b_1^{(s+1)}}{m}$, 如果

$$\begin{cases} b_4^{(s+1)}>b_2^{(s+1)}, \\ 2b_3^{(s+1)}\geqslant b_2^{(s+1)}+b_4^{(s+1)}>0, \\ b_2^{(s+1)}+2b_3^{(s+1)}+b_4^{(s+1)}\leqslant m; \end{cases}$$

情形 6 $g_{10}^{(s+1)}=g_{11}^{(s+1)}=\frac{b_3^{(s+1)}+b_4^{(s+1)}}{2m}, g_{01}^{(s+1)}=\frac{b_2^{(s+1)}}{m}, g_{00}^{(s+1)}=\frac{b_1^{(s+1)}}{m}$, 如果

$$\begin{cases} b_4^{(s+1)}>b_3^{(s+1)}, \\ 2b_2^{(s+1)}\geqslant b_3^{(s+1)}+b_4^{(s+1)}>0, \\ 2b_2^{(s+1)}+b_3^{(s+1)}+b_4^{(s+1)}\leqslant m; \end{cases}$$

情形 7 $g_{01}^{(s+1)}=\frac{b_2^{(s+1)}}{2(b_2^{(s+1)}+b_3^{(s+1)})}, g_{10}^{(s+1)}=0.5-g_{01}^{(s+1)}, g_{11}^{(s+1)}=\frac{b_4^{(s+1)}}{2(b_1^{(s+1)}+b_4^{(s+1)})}, g_{00}^{(s+1)}=0.5-g_{11}^{(s+1)}$, 如果

$$\begin{cases} b_2^{(s+1)}+b_3^{(s+1)}>b_1^{(s+1)}+b_4^{(s+1)}, \\ b_4^{(s+1)}>0, \\ b_1^{(s+1)}b_2^{(s+1)}>b_3^{(s+1)}b_4^{(s+1)}, \\ b_1^{(s+1)}b_3^{(s+1)}>b_2^{(s+1)}b_4^{(s+1)}. \end{cases}$$

把上面的过程进行迭代直到收敛，我们就可以得到 $\mathbf{g}$ 的约束的 MLE, 进而通过等式 (1) 便可得到 θ 的约束的 MLE $\hat{\theta}^R$.

4 后代数目不同的情形

在实际中, 我们也可能获得不同后代数目的家庭数据, 这时只要把我们的方法 (REM) 进行适当修改, 同样可以得到重组率的约束的 MLE $\hat{\theta}^R$.

为了演示的目的，我们考虑下面的例子. 假定获得的观测中有 n 个两后代的家庭，又有 m 个三后代的家庭. 这时，我们需要把所有家庭分成 9 个表型类, 其中 4 类相应于两后代的家庭，分到每一类的观测家庭数用 n_k 来表示, 对应的概率分别为 p_k, $k=1,2,3,4$; 另外 5 类相应于三后代的家庭, 分到每一类的观测家庭数用 m_t 来表示, 对应的概率分别为 q_t, $t=1,2,3,4$. 首先还是把观测数据 $\{n_k,\ k=1,2,3,4,m_t,t=1,2,3,4,5\}$ 扩充成完全数据 $\{n_{kj},k,j=1,2,3,4,m_{tl},t=1,2,3,4,5,l=1,2,3,4\}$, 注意这里 (n_{kj},m_{tl}) 服从乘积的多项分布. 其次, 在 REM 算法的第 s 步迭代中，需要计算的期望重组数目 $c_i^{(s+1)}=2a_i^{(s+1)}+3b_i^{(s+1)}, i=1,2,3,4$，其中 $a_i^{(s+1)}$ 和 $b_i^{(s+1)}$ 分别是前面提到的两后代情形 [17] 和三后代情形 (见附录) REM 算法中第 s 步迭代的期望重组数目，算法的其它步骤和前面针对三后代情形的算法步骤相同, 只需要把 $b_i^{(s+1)}$ 换成 $c_i^{(s+1)}$, m 换成 $2n+\ 3m$ 即可. 从而可以看出用于两后代情形和三后代情形的 REM 算法也是此时的两种特例.

5 交换干扰的估计及检验问题

交换干扰是指交换在相邻的染色体区间内不独立出现的现象. 许多研究已经证实在动物和人类的染色体中均存在一定程度的正干扰, 而且随着更多遗传标记的发现, 证据也随之增加 [25,26]. 在得到了

重组率的估计后, 可以在三位点分析中进一步考虑关于干扰的统计推断, 因为干扰的存在在实际中可能会影响到基因的定位, 以及对更复杂家系数据的推断.

对于前面所考虑的三个有序位点, 干扰的量可由一致性系数 c 来衡量 (参见文献 [5, 27]), 其中 $c=\frac{g_{11}}{\theta_{AB}\theta_{BC}}$. 事实上, 干扰值

$$I=1-c=1-\frac{g_{11}}{\theta_{AB}\theta_{BC}}. \tag{4}$$

当 $I=0$ 时, 区间 AB 和 BC 内的重组状况是独立的; 当 $I>0$ 时表示正干扰; $I<0$ 时表示负干扰. 若数据与 $I=0$ 情形有显著的偏离, 则意味着有干扰的存在. 在统计学中, 相关系数 ρ 用来衡量与独立的偏离程度. 如果我们用取值分别为 1 和 0 的随机变量 X 表示在区间 AB 内重组的发生与否, 类似地, 用 Y 表示区间 BC 内重组的发生与否, 则我们可得到下面的关系式:

$$\rho=-I\sqrt{\frac{\theta_{AB}\theta_{BC}}{(1-\theta_{AB})(1-\theta_{BC})}}.$$

这个等式正是从统计学的角度展示了干扰 I 在生物学中的地位.

根据等式 (4) 的函数关系以及 MLE 的性质, 我们可直接得到 I 的估计, 即

$$\hat{I}=1-\frac{\hat{g}_{11}}{\hat{\theta}_{AB}\hat{\theta}_{BC}}.$$

下面我们考虑对干扰的检验问题, 即基于所收集的数据检验是否存在与 $I=0$ 偏离的情形. 检验问题为

$$\mathrm{H}_0:\ I=0,\quad \mathrm{H}_1:\ I\neq 0.$$

对于该假设, 我们采用经典的似然比检验方法, 其统计量为

$$\mathrm{LRT}=-2\log\frac{L(\tilde{\theta}\mid \mathrm{Data})}{L(\hat{\theta}\mid \mathrm{Data})},$$

其中 $\tilde{\theta}$ 与 $\hat{\theta}$ 分别表示 H_0 和 H_1 下参数向量 θ 的 MLE. 在 H_0 下, LRT 渐近服从自由度为 1 的 χ^2 分布. 如果其它情形所需的正则条件不能满足, 则可以用到置换检验 [28], 该检验不依赖 LRT 的分布, 可以确定检验的临界值, 从而去判断检验是否显著. 在第 7 节中我们将给出一个实例演示对干扰的推断.

6 模拟与结果

6.1 模拟设计

我们将通过模拟研究来评价所提出的 REM 算法的优良性和稳健性. 令 $\theta_0=(\theta_{AB},\theta_{BC},\theta_{AC})$ 表示重组率的真值. 为了充分的比较 REM 算法与无约束算法的优缺点，我们考虑区间 AB 和 BC 连锁状态的 6 种不同的组合情形: CC, CM, CL, MM, ML 和 LL, 其中 C, M 和 L 分别表示近、中度和松弛连锁. 在每种情形中, θ_{AB} 和 θ_{BC} 分别取 0.05, 0.15 和 0.35 作为每种连锁程度的代表值. 在每种组合情形中, θ_{AC} 取三个等间隔的值, 使之能够保证 $(\theta_{AB},\theta_{BC},\theta_{AC})$ 满足不等式约束 (2), 并且较小的值和较大的值分别位于约束区域的边界附近, 中间的那个值位于约束区域的内部. 由于在回交家庭中三纯合型的亲本只能传递单倍型 abc 给下一代, 因此我们只需要考虑从杂合型的亲本产生后代单倍型的随机抽样. 根据表 1 中的条件概率, 以及四种相型的等概率先验, 基于每组 θ_0 的真值我们都随机地产生了 300 个三后代家庭数据.

表 3 300 个三后代家庭经 1000 次重复的估计结果

	参数			REM			无约束方法		
Scenario[a]	θ_{AB}	θ_{BC}	θ_{AC}	$\hat{\theta}^R_{AB}$	$\hat{\theta}^R_{BC}$	$\hat{\theta}^R_{AC}$	$\hat{\theta}^U_{AB}$	$\hat{\theta}^U_{BC}$	$\hat{\theta}^U_{AC}$
CC	0.05	0.05	0.06	0.0497	0.0499	0.0601	0.0499	0.0501	0.0598
			0.075	0.0501	0.0502	0.0749	0.0502	0.0503	0.0750
			0.09	0.0498	0.0502	0.0901	0.0498	0.0502	0.0902
CM	0.05	0.15	0.16	0.0502	0.1495	0.1606	0.0502	0.1498	0.1605
			0.175	0.0499	0.1505	0.1758	0.0499	0.1505	0.1757
			0.19	0.0498	0.1503	0.1898	0.0498	0.1504	0.1901
CL	0.05	0.35	0.36	0.0500	0.3500	0.3628	0.0500	0.3524	0.3638
			0.375	0.0501	0.3526	0.3773	0.0501	0.3542	0.3814
			0.39	0.0498	0.3500	0.3894	0.0498	0.3519	0.3596
MM	0.15	0.15	0.16	0.1494	0.1492	0.1622	0.1505	0.1502	0.1602
			0.225	0.1501	0.1508	0.2255	0.1501	0.1508	0.2255
			0.29	0.1499	0.1496	0.2880	0.1501	0.1498	0.2896
ML	0.15	0.35	0.36	0.1501	0.3499	0.3664	0.1501	0.3544	0.3641
			0.425	0.1505	0.3499	0.4287	0.1506	0.3520	0.4328
			0.49	0.1498	0.3480	0.4706	0.1499	0.3527	0.4620
LL	0.35	0.35	0.36	0.3473	0.3482	0.3723	0.3531	0.3536	0.3654
			0.425	0.3521	0.3520	0.4335	0.3537	0.3534	0.4337
			0.49	0.3509	0.3518	0.4676	0.3530	0.3539	0.4599

[a]Sccnario: 区间 AB 和 BC 连锁状态的 6 种组合 (C: 近连锁; M: 中度连锁; L: 松弛连锁).

对于每组模拟数据, 我们都用无约束方法和 REM 方法分别去计算 $\hat{\theta}^U$ 和 $\hat{\theta}^R$, 把整个过程重复 $M = 1000$ 次, 我们计算在 1000 次模拟当中 $\hat{\theta}^U$ 和 $\hat{\theta}^R$ 的平均值 (见表 3). 同时, 为了更好的比较两种方法的估计效果, 我们采用了下面三种评价估计精确性的测度 (见表 4):

(1) KK, 在 1000 次模拟当中, 由无约束方法产生的不满足约束 (2) 的估计的个数;

(2) $\hat{\theta}^R_i$ 的标准差 (SD), 以及两种估计 SD 的比值 $\mathrm{rSD} = \mathrm{SD}(\hat{\theta}^U_i)/\mathrm{SD}(\hat{\theta}^R_i)$;

(3) 估计 $\hat{\theta}^R$ 的平均绝对误差 (MAE), 其中

$$\mathrm{MAE} = \sum_{l=1}^{1000}(|\hat{\theta}^R_{ABl} - \theta_{AB}| + |\hat{\theta}^R_{BCl} - \theta_{BC}| + |\hat{\theta}^R_{ACl} - \theta_{AC}|)/3M,$$

以及两种估计 MAE 的比值 $\mathrm{rMAE} = \mathrm{MAE}(\hat{\theta}^U)/\mathrm{MAE}(\hat{\theta}^R)$.

此外, 基于 θ_0 的真值, 我们也随机地产生了 300 个包含 150 个两后代家庭和 150 个三后代家庭的数据. 在每种情形, 我们采用 REM 算法计算 $\hat{\theta}^R$, 并重复整个过程 M 次, 我们得到 M 次模拟中 $\hat{\theta}^R$ 的平均值, 相应的 SD 和 MAE 也列于表 5 中.

6.2 结果

首先, 在表 3 中我们可以发现在 1000 次模拟中由 REM 算法所得到的 $\hat{\theta}^R$ 的平均值更接近于参数真值, 尤其是在区间 AB 和 BC 中至少有一个是松弛连锁的时候, 结果更为明显. 主要原因就是无约束方法在此时更容易给出不合理的估计.

表 4 针对 300 个三后代家庭数据两种估计方法的比较结果

	参数			SD			rSD[b]					
Scenario[a]	θ_{AB}	θ_{BC}	θ_{AC}	$\hat{\theta}^R_{AB}$	$\hat{\theta}^R_{BC}$	$\hat{\theta}^R_{AC}$	$\hat{\theta}^U_{AB}$	$\hat{\theta}^U_{BC}$	$\hat{\theta}^U_{AC}$	MAE	rMAE[c]	KK[d]
CC	0.05	0.05	0.06	0.0070	0.0075	0.0075	1.0564	1.0696	1.1023	0.0058	1.0396	164
			0.075	0.0074	0.0077	0.0093	1.0038	1.0044	1.0093	0.0065	1.0045	0
			0.09	0.0075	0.0073	0.0096	1.0016	1.0030	1.0463	0.0064	1.0089	25
CM	0.05	0.15	0.16	0.0073	0.0133	0.0133	1.0008	1.0375	1.0461	0.0090	1.0171	108
			0.175	0.0074	0.0135	0.0144	1.0022	1.0098	1.0501	0.0094	1.0158	3
			0.19	0.0074	0.0133	0.0148	1.0005	1.0093	1.0887	0.0095	1.0199	87
CL	0.05	0.35	0.36	0.0076	0.0288	0.0292	1.0003	1.2293	1.4110	0.0173	1.0940	279
			0.375	0.0073	0.0303	0.0321	1.0005	1.1652	1.5752	0.0181	1.1113	229
			0.39	0.0075	0.0287	0.0312	1.0008	1.1252	1.9167	0.0177	1.1761	365
MM	0.15	0.15	0.16	0.0120	0.0120	0.0126	1.1633	1.1736	1.2040	0.0096	1.0935	371
			0.225	0.0135	0.0133	0.0171	1.0083	1.0054	1.0601	0.0116	1.0146	0
			0.29	0.0132	0.0131	0.0198	1.0372	1.0018	1.3363	0.0124	1.0707	243
ML	0.15	0.35	0.36	0.0125	0.0269	0.0311	1.0002	1.3862	1.3094	0.0182	1.1120	342
			0.425	0.0127	0.0289	0.0390	1.0041	1.2254	1.5560	0.0217	1.1612	88
			0.49	0.0129	0.0255	0.0351	1.0120	1.6825	2.1959	0.0184	1.3305	203
LL	0.35	0.35	0.36	0.0247	0.0255	0.0311	1.6445	1.6292	1.4160	0.0209	1.2154	525
			0.425	0.0307	0.0304	0.0446	1.2376	1.1861	1.3048	0.0283	1.1019	123
			0.49	0.0311	0.0298	0.0311	1.2002	1.2191	1.6059	0.0257	1.1773	58

[a]Scenario, 见表 3 中的解释; [b]rSD = SD($\hat{\theta}^U_i$)/SD($\hat{\theta}^R_i$), $i = AB, BC, AC$; [c]rMAE = MAE($\hat{\theta}^U$)/MAE($\hat{\theta}^R$); [d]KK: 在 1000 次模拟中由无约束方法产生的不满足约束 (2) 的估计的个数.

我们下面考虑针对三后代家庭模拟数据两种估计方法的比较情形 (见表 4). 在多数情形, 无约束方法都产生了许多不合理的估计, 它们都不满足约束 (2). 但相反, 由 REM 算法计算得到的估计均满足约束条件. 在所有情形中, KK 的值呈现出某些规律性. 读者可参见文献 [17].

不难发现, 我们的 REM 算法在每种情形中都要比无约束方法做得好. 从精度的角度看, 由 REM 算法计算的估计其 SD 均小于由无约束方法所得到的相应的值, 尤其是在区间 AB 和 BC 中至少有一个是松弛连锁时, 结果更为明显. 与 $\hat{\theta}^U$ 相比, $\hat{\theta}^R$ 更加接近真实值 θ_0 (表 4 中 rMAE 均大于 1). 所有这些结果显示利用 REM 算法计算重组率的估计可得到更高的精度. 另外, 在估计中一定要考虑到约束 (2), 否则将对实际推断造成显著的影响. 综上, 我们得出结论: 使用 REM 算法会得到更好的估计结果, 而且它也是一个稳健的算法.

表 5 中给出了 300 个包含 150 个两后代家庭和 150 个三后代家庭数据的模拟结果. 通过比较表 4 和表 5 中 REM 算法所给出的结果, 我们发现对于同一个真值 θ_0, 表 4 中每个 SD (或 MAE) 均小于表 5 中相应的值; 再与文献 [17] 的结果比较, 我们发现, 表 4 和表 5 中所有的 SD 或 MAE 的值均小于两后代家庭数据时的相应值; 另外, 随着后代数目的增加, 无约束方法产生的不合理估计的数目也相应减少. 这说明每个家庭中后代越多确实能够提供越多的连锁信息, 并给出更精确的估计. 若每个家庭中后代的数目再增加, 上面的比较结果仍会保持.

我们也做了其它的模拟来评价 REM 算法的收敛速度, 以及评价样本量对估计的影响. REM 算法本质上还是 EM 算法, 因此其收敛速度不快. 然而, 在三位点分析中仅有三个参数, 并且在算法的 M

表 5 包含 150 个两后代家庭和 150 个三后代家庭数据的模拟结果

	参数			MLE			SD			
Scenario	θ_{AB}	θ_{BC}	θ_{AC}	$\hat{\theta}^R_{AB}$	$\hat{\theta}^R_{BC}$	$\hat{\theta}^R_{AC}$	$\hat{\theta}^R_{AB}$	$\hat{\theta}^R_{BC}$	$\hat{\theta}^R_{AC}$	MAE
CC	0.05	0.05	0.06	0.0495	0.0493	0.0604	0.0082	0.0078	0.0085	0.0066
			0.075	0.0499	0.0499	0.0750	0.0083	0.0081	0.0103	0.0071
			0.09	0.0500	0.0495	0.0896	0.0080	0.0082	0.0109	0.0072
CM	0.05	0.15	0.16	0.0500	0.1498	0.1602	0.0080	0.0150	0.0154	0.0103
			0.175	0.0500	0.1496	0.1750	0.0077	0.0153	0.0167	0.0105
			0.19	0.0497	0.1503	0.1892	0.0081	0.0150	0.0165	0.0104
CL	0.05	0.35	0.36	0.0499	0.3499	0.3629	0.0082	0.0337	0.0343	0.0199
			0.375	0.0498	0.3501	0.3752	0.0083	0.0344	0.0362	0.0206
			0.39	0.0501	0.3519	0.3904	0.0082	0.0350	0.0377	0.0212
MM	0.15	0.15	0.16	0.1493	0.1491	0.1621	0.0137	0.0142	0.0142	0.0112
			0.225	0.1506	0.1505	0.2258	0.0144	0.0153	0.0193	0.0130
			0.29	0.1505	0.1497	0.2903	0.0143	0.0150	0.0213	0.0135
ML	0.15	0.35	0.36	0.1509	0.3488	0.3691	0.0154	0.0323	0.0377	0.0220
			0.425	0.1450	0.3526	0.4292	0.0153	0.0350	0.0436	0.0255
			0.49	0.1508	0.3465	0.4642	0.0150	0.0301	0.0427	0.0222
LL	0.35	0.35	0.36	0.3465	0.3473	0.3744	0.0306	0.0300	0.0355	0.0246
			0.425	0.3533	0.3537	0.4360	0.0376	0.0367	0.0472	0.0325
			0.49	0.3539	0.3520	0.4679	0.0374	0.0364	0.0437	0.0289

步中参数 $\mathbf{g}$ 的估计有显式表达, 因此它的收敛情况还是令人满意的. 对于上面我们在每种情形中所模拟产生的 300 个三后代家庭数据, 我们计算了 REM 算法在 1000 次模拟中的平均迭代步数. 结果显示最大值仅为 19, 而最小值是 4, 并且每一步的计算时间也较短. 根据我们模拟和实际经验, 若使用无约束估计作为 REM 算法的初值还会提高其计算速度.

模拟中也考虑了不同的样本量 (家庭数) 情形. 我们发现随着样本量的增加, 估计会更为精确 (结果未列出). 根据我们的经验, 当样本量 m 超过 50 时, REM 算法给出的结果就可以接受; 当 m 达到 100 时, 结果已经令人满意; 并且在任何时候与无约束方法的比较结果都能保持.

7 一个例子

我们用提出的方法来分析文献 [29] 中的一个真实的数据集. 该数据集包含了 203 个来自回交家庭的老鼠. 我们考虑 2 号染色体连锁图上 3 个有序的位点 D2Mit120, D2Mit182 和 D2Mit133, 为方便起见, 仍用 A, B 和 C 表示它们. 根据数据集中所给的基因型, 我们记录下来自杂合亲本的每个个体的单倍型. 然后把三个个体随机的分入一个家庭, 并且认为它们确实来源于同一个家庭. 这样做不会影响到连锁信息, 因为对于这组数据, 在给定所有亲本的基因型的条件下所有后代个体的基因型是独立的. 于是按照表 2 中的分类, 我们得到了 $m = 67$ 个三后代家庭, 且 $(m_1, m_2, m_3, m_4, m_5) = (6, 15, 12, 15, 19)$, 及 1 个两后代家庭.

首先, 基于该数据集我们利用 REM 和无约束两种方法去估计重组率. 经计算我们分别得到: $\hat{\theta}^R_{AB} =$

0.3482, $\hat{\theta}^R_{BC} = 0.3797$, $\hat{\theta}^R_{AC} = 0.3797$; 以及 $\hat{\theta}^U_{AB} = 0.3546$, $\hat{\theta}^U_{BC} = 0.4211$, $\hat{\theta}^U_{AC} = 0.3546$. 很明显, 所得的无约束估计不满足约束 (2) 中第二个条件, 因而这个估计值与 2 号染色体连锁图上真实的位点顺序相矛盾 [29]. 对于人类数据而言, 使用多点分析更为实际. 然而, 三位点分析的方法也能够推广至多点分析中去.

其次, 根据第 5 节中的讨论, 我们得到了干扰值的估计, 即 $\hat{I} = -0.317$. 我们也计算了似然比检验统计量 LTR 的值是 0.819, 当检验的水平取为 0.05 时, 该检验是不显著的. 实际上, 这组数据的样本量对于检验干扰来说并不充分. Ott[5] 指出, 若检验水平为 0.05, 功效要达到 80%, 则需要 800 多次全信息的减数分裂才能探测出干扰. 这里我们主要是为了演示对交换干扰的推断过程.

8 结论与讨论

本文中我们考虑了针对具有任意后代数目的相型未知的三回交家庭数据的重组率估计问题. 无约束方法会经常给出不满足约束 (2) 的估计, 因为此方法没有把参数的约束考虑到估计中去. 我们提出的 REM 算法能够解决该问题, 而且作为一种通用的方法, 它可以被推广到多种应用情形. 即使是后代数目较大时, 表型分类以及分类数也可以通过我们提出的分类方法解决.

我们采用了真实和模拟两种数据演示了 REM 算法的优良性. 对于三后代家庭数据, 模拟显示我们提出的算法要优于无约束方法, 并且新方法是稳健和有效的. 同时, 新方法也成功地用于分析老鼠的数据集. 对于人类遗传数据, 注重参数 (重组率) 约束的思想也应纳入到分析当中. 我们的研究也证实了每个家庭中的后代个数越多, 则给连锁分析提供的信息也越多这一观点, 并且我们的方法能给出更精确的结果. 当后代数超过 3 时, 本文给出的分类方法可以用于处理表型的分类.

模拟结果也显示出样本量对重组率的估计具有一定的影响. 样本量越大, 估计也越精确. 当样本量超过 50 时, 估计结果便可以接受. 尽管 EM 算法的收敛速度通常很慢, 然而因参数少、及算法的 M 步中参数 $\mathbf{g}$ 的估计有显式表达, 使得 REM 的收敛速度仍是令人满意的. 在 1000 次模拟中, REM 算法达到收敛时, 平均迭代次数最多为 19, 而且每步中计算时间都很短, 因而我们的算法并不费时.

在更为一般的情形, REM 算法也可以处理观测家庭中后代数目不同的数据. 此外, 受诸多因素的影响 (如连锁不平衡 [2]), 杂合型亲本的连锁相可能并不以等概率出现, 但是该算法同样可以作为一种通用的方法应用于该情形 (见文献 [17]).

本文的分析集中于三个双等位基因位点, 即三点分析. 当位点数超过 3 个, 或者某一位点有更多的等位基因时, 例如远交群体, 上面的约束参数问题将变得复杂. 当位点数超过 3 个时, Zhou 等 [17] 建议了一种分析方法. 该方法把多位点分析转化成多个三位点分析, 而后对三位点分析的结果进行整合再得到最终的结果.

在统计遗传学中, 最新研究显示对两位点重组率的统计推断确实能够为构建和分析标记位点与遗传疾病位点间的连锁图提供一种有效的手段. 合理的重组率的估计在基因定位中是很关键的, 尤其是数量性状位点 (QTL) 的区间定位 [30,31]. 重组率和 QTL 效应是 QLT 定位中的重要参数, QTL 位置的确定直接取决于对重组率的精确估计. 只有合理的估计才能在基因组上定位到控制某种性状的真实的基因 [31]. 然而, 交换干扰是影响基因定位的一个主要因素 [32], 因此有必要在基因定位前对干扰的存在与否进行检验. 在第 5 节中我们所考虑的方法可以用于处理该问题. 但要注意的是, 检验要用到的样本应当是充足的 [5]. REM 算法适用于实验性的 (回交) 群体, 它可推广至多位点情形. 因此, 我们认为把该算法嵌入到区间定位中以提高其效率是可行的. 当然在这一领域还需进一步的研究.

致谢 特别地感谢陈永川教授、付梅博士和朱文圣博士对文章中某些问题的评论, 同时也衷心地感谢编辑和审稿人对本文提出的宝贵意见.

参考文献

1 Elston R C, Stewart J. A general model for the analysis of pedigree data. Hum Hered, 1971, 21: 523–542
2 Risch N. Linkage strategies for genetically complex traits. Amer J Hum Genet, 1990, 46: 222–253
3 Weir B S. Genetic Data Analysis II. Sunderland, MA: Sinauer Associates, 1996
4 Kruglyak L, Daly M J, Reeve-Daly M P, et al. Paremeteric and nonparameteric linkage analysis: a unified multipoint approach. Amer J Hum Genet, 1996, 58: 1347–1363
5 Ott J. Phase-unkown triple backcross with two offspring. In: Analysis of Human Genetic Linkage, 3rd ed. Baltimore: The Johns Hopkins University Press, 1999, 122–124
6 Thompson E A. Statistical Inference from Genetic Data on Pedigree. Ohio: Institute of Mathematical Statistics Beachwood, 2000
7 Haldane J B S. The recombination of linkage values and the calculation of distances between the loci of linked factors. J Genet, 1919, 8: 299–309
8 Morgan T H. The Theory of Genes. New Haven: Yale University Press, 1928
9 Felsenstein J. A mathematically tractable family of genetic mapping functions with different amounts of interference. Gentics, 1979, 91: 769–775
10 Fisher R A. The experimental study of multiple crossover-over. Caryologia, 1954, 6: 227–231
11 Thompson E A. Information gain in joint linkage analysis. IMA J Math Appl Med Biol, 1984, 1: 31–49
12 Ridout M S, Tong S, Vowden C J, et al. Three-point linkage analysis in crosses of allogamous plant species. Genet Res, 1998, 72: 111–121
13 Wu R L, Ma C M, Painter I, et al. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing populations. Theo Pop Biol, 2002, 61: 349–363
14 Lu Q, Cui Y H, Wu R L. A multilocus likelihood approach to joint modelling of linkage, parnet diplotype and gene order in a full-sib family. BMC Genet, 2004, 5: 20
15 Lathrop G M, Lalouel J M, Julier C, et al. Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA, 1984, 81: 3443–3446
16 Lathrop G M. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Amer J Hum Genet, 1985, 37: 482–498
17 Zhou Y, Shi N Z, Fung W K, et al. Maximum likelihood estimates of two-locus recombination fractions under some natural inequality restrictions. BMC Genet, 2008, 9: 1
18 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B, 1977, 39: 1–38
19 Dykstra R L. An algorithm for restricted least squares regression. J Amer Statist Assoc 1983, 78: 837–842
20 Agresti A. Analysis of Ordinal Categorical Data. New York: Wiley, 1984
21 Robertson T, Wright F T, Dykstra R. Order Restricted Statistical Inference. New York: Wiley, 1988
22 Liu C. Estimation of discrete distribution with a class of simplex constraints. J Amer Stat Assoc, 2000, 95: 109–120
23 Shi N Z, Zheng S R, Guo J H. The restricted EM algorithm under inequality restrictions on the parameters. J Multivariate Anal, 2005, 92: 53–76
24 Krishnamurthy V. Combinatorics: Theory and Applications. New York: John Wiley and Sons, 1986
25 Weeks D E, Ott J, Lathrop G M. Detection of genetic interference: simulation studies and mouse data. Genetics, 1994, 136: 1217–1226
26 Broman K W, Weber J L. Characterization of human crossover interference. Amer J Hum Genet, 2000, 66: 1911–1926
27 Ott J. Testing for interference in human genetic maps. J Mol Med, 1997, 75: 414–419
28 Churchill G A, Doerge R W. Empirical threshold values for quantitative trait mapping. Genetics, 1994, 138: 963–971
29 Reifsnyder P C, Churchill G, Leiter E H. Maternal environment and genotype interact to establish diabesity in mice. Genome Res, 2000, 10: 1568–1578
30 Kao C H, Zeng Z B. General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics, 1997, 53: 653–665
31 Chen Z. The full EM algorithm for the MLEs of QTL effects and positions and their estimated variance in multiple-interval mapping. Biometrics, 2005, 61: 474–480

32 Augera D L, Sheridanb W F. Negative crossover interference in maize translocation heterozygotes. Genetics, 2001, 159: 1717–1726

附录

第 3.2 节中出现在 $Q(\mathbf{g} \mid \mathbf{g}^{(s)}, \{m_k\})$ 中的 $b_i^{(s+1)}$ 的表达式:

$$
\begin{aligned}
b_1^{(s+1)} &= m_1\frac{(g_{00}^{(s)})^3}{p_1^{(s)}} + 2m_2\frac{(g_{00}^{(s)})^2 g_{01}^{(s)}}{p_2^{(s)}} + m_2\frac{(g_{01}^{(s)})^2 g_{00}^{(s)}}{p_2^{(s)}} + m_3\frac{(g_{10}^{(s)})^2 g_{00}^{(s)}}{p_3^{(s)}} + 2m_3\frac{(g_{00}^{(s)})^2 g_{10}^{(s)}}{p_3^{(s)}} + m_4\frac{(g_{11}^{(s)})^2 g_{00}^{(s)}}{p_4^{(s)}} \\
&\quad + 2m_4\frac{(g_{00}^{(s)})^2 g_{11}^{(s)}}{p_4^{(s)}} + 2m_5\frac{g_{00}^{(s)} g_{01}^{(s)} g_{11}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{01}^{(s)} g_{00}^{(s)} g_{10}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{11}^{(s)} g_{10}^{(s)} g_{00}^{(s)}}{p_5^{(s)}}, \\
b_2^{(s+1)} &= m_1\frac{(g_{01}^{(s)})^3}{p_1^{(s)}} + 2m_2\frac{(g_{01}^{(s)})^2 g_{00}^{(s)}}{p_2^{(s)}} + m_2\frac{(g_{00}^{(s)})^2 g_{01}^{(s)}}{p_2^{(s)}} + m_3\frac{(g_{11}^{(s)})^2 g_{01}^{(s)}}{p_3^{(s)}} + 2m_3\frac{(g_{01}^{(s)})^2 g_{11}^{(s)}}{p_3^{(s)}} + m_4\frac{(g_{10}^{(s)})^2 g_{01}^{(s)}}{p_4^{(s)}} \\
&\quad + 2m_4\frac{(g_{01}^{(s)})^2 g_{10}^{(s)}}{p_4^{(s)}} + 2m_5\frac{g_{00}^{(s)} g_{01}^{(s)} g_{11}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{01}^{(s)} g_{00}^{(s)} g_{10}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{10}^{(s)} g_{11}^{(s)} g_{01}^{(s)}}{p_5^{(s)}}, \\
b_3^{(s+1)} &= m_1\frac{(g_{10}^{(s)})^3}{p_1^{(s)}} + 2m_2\frac{(g_{10}^{(s)})^2 g_{11}^{(s)}}{p_2^{(s)}} + m_2\frac{(g_{11}^{(s)})^2 g_{10}^{(s)}}{p_2^{(s)}} + m_3\frac{(g_{00}^{(s)})^2 g_{10}^{(s)}}{p_3^{(s)}} + 2m_3\frac{(g_{10}^{(s)})^2 g_{00}^{(s)}}{p_3^{(s)}} + m_4\frac{(g_{01}^{(s)})^2 g_{10}^{(s)}}{p_4^{(s)}} \\
&\quad + 2m_4\frac{(g_{10}^{(s)})^2 g_{01}^{(s)}}{p_4^{(s)}} + 2m_5\frac{g_{01}^{(s)} g_{00}^{(s)} g_{10}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{10}^{(s)} g_{11}^{(s)} g_{01}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{11}^{(s)} g_{10}^{(s)} g_{00}^{(s)}}{p_5^{(s)}}, \\
b_4^{(s+1)} &= m_1\frac{(g_{11}^{(s)})^3}{p_1^{(s)}} + 2m_2\frac{(g_{11}^{(s)})^2 g_{10}^{(s)}}{p_2^{(s)}} + m_2\frac{(g_{10}^{(s)})^2 g_{11}^{(s)}}{p_2^{(s)}} + m_3\frac{(g_{01}^{(s)})^2 g_{11}^{(s)}}{p_3^{(s)}} + 2m_3\frac{(g_{11}^{(s)})^2 g_{01}^{(s)}}{p_3^{(s)}} + m_4\frac{(g_{00}^{(s)})^2 g_{11}^{(s)}}{p_4^{(s)}} \\
&\quad + 2m_4\frac{(g_{11}^{(s)})^2 g_{00}^{(s)}}{p_4^{(s)}} + 2m_5\frac{g_{00}^{(s)} g_{01}^{(s)} g_{11}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{11}^{(s)} g_{10}^{(s)} g_{00}^{(s)}}{p_5^{(s)}} + 2m_5\frac{g_{10}^{(s)} g_{11}^{(s)} g_{01}^{(s)}}{p_5^{(s)}}.
\end{aligned}
$$

Maximum likelihood estimates of recombination fractions under restrictions for family data with multiple offspring

ZHOU Ying, HAN GuoNiu, SHI NingZhong, FUNG Wing-Kam & GUO JianHua

Abstract This paper discusses the estimation problem of recombination fractions under some natural inequality restrictions in phase-unknown triple backcross population. We consider the offspring phenotype classification of multiple offspring family, and present an explicit formula of the number of the offspring phenotype classification. We then adopt the restricted expectation-maximization (REM) algorithm to estimate the two-locus recombination fractions based on the data of phenotype classification. Simulation studies show that the REM algorithm outperforms unrestricted method, and validate that family with more offspring can provide more linkage information.

Keywords: constrained parameter problems, linkage analysis, recombination fraction, restricted EM algorithm

MSC(2000): 97K80, 92D99