A moment estimation of the haplotypes' distribution using phenotypes' data[1]

F. Fève and J.-P. Florens

June 2003

[1]We thank the European Consortium MADO (for Marrow Donors for initially raising the question, France Greffe de Moelle for providing data. We are grateful to Anne-Cambon and Pierre-Antoine Gourraud for helpful discussions and comments. We are grateful to Fabrice Collard for his computational assistance.

Abstract

The polymorphism at several loci on the same chromosome is often a problem for interpreting individual phenotypes because the phase of heterozygous can't be determined. The phase of haplotypes is usually unknown when diploids individuals are heterozygous at more than one locus. Haplotypes can be obtained, at considerable cost, experimentally through genetyping of additional family members. Alternatively, a statistical method can be used to estimate the haplotypes distribution. We present a new statistical method, base on Hardy Weinberg Equilibrium to infer the phase of phenotypes. To contend with some weakness of the existing algorithms, we propose a new statistical approach, which is neither an EM algorithm nor a bayesian MCMC approach. The intuition of this method may be presented in the following way. Empirical distributions deduced from phenotypes' data are biased estimates of the true estimation of haplotypes. However if we consider Q loci the bias depends on the distribution of subsets of $Q-1$ loci. As the marginal distributions may be estimated consistently we construct a bias correction recursion which provides unbias asymptotically normal estimations.

Keywords: haplotype distribution, estimation of mixtures, phenotypes latent variables, moments estimation.

EMS:

1 Introduction

Haplotype information is essential for many analyses of genetics data, for example, in disease mapping (Risch and Merikangas 1996) or in DNA pooling (Wang, Kidd, Zhao 2003). Haplotype estimation is an important issue, both in population genetics (Single, Merger et alii 2002) and in the identification of complex diesease genes (Niu, Qin et alii 2002). For example, associations between markers and disease loci that are not evident with a single marker locus may be identified in multi-locus market analyses using estimated haplotype frequencies. Current genotyping methods do not provide phase information. This can be obtained, partially through genotyping of additional family members (Duldbridge, Kolleman et alii 2000).

If no information is available from family members, a statistical method may be used. From phenotypes observations, the joint distribution of haplotypes is estimated and the knowledge of this distribution and of the phenotype of an individual may be used to infer the phase.

This paper addresses the question of estimation of the joint distribution of the haplotypes using phenotypes data and proposes a new simple estimation method. The use of this distribution to infer the phase for an individual does not depend on the estimation procedure and is not explicitely considered in this paper. The two most popular existing methods are maximum likelihood, implemented via the EM Algorithm (Excoffier and Slatkin, 1995), and a parsimony method created by Clark (1990)). Thus, a third method is proposed by Stephens, Smith and Donnely (2001). Their phase reconstruction method (a bayesian one) uses Gibb's sampling, a type of MCMC algorithm. We present a new statistical method, based on Hardy-Weinberg Equilibrium to infer the phase of phenotype, which is neither an EM Algorithm nor a bayesian MCMC approach. This method is a moment estimation, based on the fitting between theoretical and empirical moments and provides estimation of the haplotype's distribution using phenotypes data. Even if moment estimation does not reach asymptotic efficiency bound like maximum likelihood, its has similar advantages. This estimator is more easy to compute, it does not depend on a stopping rule and simulations show that is can perform better then maximum likelihood in small sample analysis.

In section II we describe the latent model and the observable. In section III we define both likelihood of latent and observable data. In section IV we propose an introductory case, for our moment estimation in order to illustrate our general theory (see section V). The properties of the asymptotic distribution are examined in section VI. We demonstrate our methodology with simulation and examples set in section VII.

2 The latent model and the observables

The specification of the statistical model starts by assumptions on a set of latent observations. These assumptions will be completed by the description of the observation scheme and the distribution of the observables is deduced from these two parts of the specification.

The latent variables are the observations at the haplotype level. They are constituted by a sequence :

$$(\xi_1^q(i), \xi_2^q(i)) \quad q = 1, ..., Q \quad i = 1, ..., n \tag{2.1}$$

where i denotes the individual and q the locus. For each individual and each locus, $(\xi_1^q(i), \xi_2^q(i))$ represents the values of the alleles on the paternal chromosom (1 in our notation) and on the maternal chromosom (indexed by 2).

The number of possible alleles for locus q is r^q and the set of these alleles is $J^q = \{1, ..., r^q\}$.

The parameter of interest is the joint distribution of the alleles on different locus on each phase, namely the numbers :

$$\begin{aligned} p(j^1, ..., j^Q) &= P(\xi_1^1(i) = j^1, ..., \xi_1^Q(i) = j^Q) \\ &= P(\xi_2^1(i) = j^1, ..., \xi_2^Q(i) = j^Q) \end{aligned} \tag{2.2}$$

This notation implicitely assumes that the distribution of the alleles is identical for each individual and each chromosom. We assume more over that individuals and chromosoms are independent. These assumptions are implicitely based on the Hardy Weinberg equilibrium of the population. In other words the latent model describes a sample of $2n$ observations of a Q varied discrete random vector.

Unfortunately, the location of alleles on the two chromosoms of a given individual is not observable. Different ways exist for modeling this partial observability.

The more common way is to transform each pair $(\xi_1^q(i), \xi_2^q(i))$ into the rank statistic and the order statistic and to analyse the case where the order statistic only is observable. The difficulty of this method is the possibility of ties (homozygous individual) which requires a specific analysis. We prefer to adopt an other strategy based on the idea of a random (non observable) permutation of the data.

Let us extend the latent model by considering a vector

$$(\delta^q(i)) \quad q = 1, ..., Q \quad i = 1, ..., n$$

where $\delta^q(i) \in \{0, 1\}$. The new latent observations are now

$$(\xi_1^q(i), \xi_2^q(i), \delta^q(i)) \quad q = 1, ..., Q \quad i = 1, ..., n$$

and we define

$$\begin{aligned} X_1^q(i) &= \delta^q(i)\ \xi_1^q(1) + (1 - \delta^q(i))\ \xi_2^q(i) \\ \\ X_2^q(i) &= (1 - \delta^q(i)\ \xi_1^q(i) + \delta^q(i))\ \xi_2^q(i) \end{aligned}$$

Equivalently $X_1^q(i)$ is equal to $\xi_1^q(i)$ if $\delta^q(i) = 1$ and equal to $\xi_2^q(i)$ else. The variable $\delta^q(i)$ may be interpreted as the indicator of the allele which is observed first.

The probabilistic specification of the model is completed by the distribution of the $\delta^q(i)$. They are assumed to be *i.i.d* between individuals and between locus and

$$P(\delta^q(i) = 1) = P(\delta^q(i) = 0) = \frac{1}{2}.$$

Moreover the $\delta^q(i)$ are independent of the $(\xi_1^q(i), \xi_2^q(i))_{q,i}$. Let us underline that the distribution of the $\delta^q(i)$ is the marginal*i.i.d.* distribution. As we will see later, this distribution of the $\delta^q(i)$ given the $(X_1^q(i), X_2^q(i))$ is different.

There obviously exists a one to one transformation between the $(\xi_1^q, (i), \xi_2^q(i), \delta^q(i))$ and $(X_1^q(i), X_2^q(i), \delta^q(i))$. Indeed

$$\begin{aligned} \xi_1^q(i) &= \delta^q(i)\ X_1^q(i) + (1 - \delta^q(i))\ X_2^q(i) \\ \\ \xi_2^q(i) &= (1 - \delta^q(i))\ X_1^q(i) + \delta^q(i)\ X_2^q(i) \\ &= (\delta'(i), ..., \delta^Q(i)) \end{aligned}$$

The vector $\delta(i)$ may be called the phase configuration for an individual i.

Finally the observed sample is characterized by the

$$(X_1^q(i), X_2^q(i)) \quad q = 1, ..., Q \quad i = 1, ..., n.$$

or equivalently the $(\delta^q(i)) \quad q = 1, ..., Q \quad i = 1, ..., n$ are not observed. The actual observations are called the phenotypes.

Remark: It should be stressed that the informational contained of our observed sample is equivalent to the order statistic. Actually the knowledge of the $X_1^q(i), X_2^q(i)$ implies the knowledge of the order statistic

$$min(X_1^q(i), X_2^q(i)) = min(\xi_1^q(i), \xi_2^q(i))$$

and

$$max(X_1^q(i), X_2^q(i)) = max(\xi_1^q(i), \xi_2^q(i))$$

Reciprocally given the order statistic, one can draw a vector of $(\delta^q(i))_{q,i}$ and construct the $X_1^q(i), X_2^q(i)$ from the order statistic. In that case the likelihood of the $X_1^q(i), X_2^q(i)$ construct from the latent observations or from the order are identical.

3 Likelihood

The sampling distribution of the latent variables $(\xi_1^q(i), \xi_2^q(i), \delta^q(i))$ is equal to:

$$L^* = \prod_{i=1}^{n} \left\{ \frac{1}{2^Q} p(\xi_1^1(i), ..., \xi_1^Q(i)) p(\xi_2^1(i), ..., \xi_2^Q(i)) \right\} \tag{3.1}$$

or using the one to one transformation between the $(\xi_1^q(i), \xi_2^q(i), \delta^q(i))$ and $(X_1^q(i), X_2^q(i), \delta^q(i))$ this likelihood is equal to:

$$\begin{aligned} L^* = \prod_{i=1}^{n} \Big\{ & \frac{1}{2^Q} p(\delta^1(i)X_1^1(i) + (1-\delta^1(i))X_2^1(i), ..., \ \delta^Q(i)X_1^Q(i) + (1-\delta^Q(i))X_2^Q(i)) \times \\ & p(1-\delta^1(i))X_1^1(i) + \delta^1(i)X_1^1(i), ..., (1-\delta^Q(i))X_1^Q(i) + \delta^Q(i)X_2^Q(i)) \Big\} \end{aligned} \tag{3.2}$$

and the marginal likelihood of the observable data is obtained by summing up the $(\delta^q(i))_q = \delta(i)$ in the set of all possible $\delta(i)$, namely $\{0,1\}^Q$. Then the likelihood of the phenotype is:

$$\begin{aligned} L^* = \prod_{i=1}^{n} \Bigg\{ & \frac{1}{2^Q} \sum_{\delta(i) \in \{0,1\}^Q} \\ & p(\delta^1(i)X_1^1(i) + (1-\delta^1(i))X_2^1(i), ..., \delta^Q(i)X_1^Q(i) + (1-\delta^Q(i))X_2^Q(i)) \\ & \times p((1-\delta^1(i))X_1^1(i) + \delta^1(i)X_2^1(i), ..., (1-\delta^Q(i))X_1^Q(i) + \delta^Q(i)X_2^Q(i)) \Bigg\} \end{aligned} \tag{3.3}$$

This marginal likelihood involves a sum over 2^Q terms and is untractable but the EM algorithm provides an efficient way to compute numerically the value of $p(j^1, ..., j^Q)$ which maximises this likelihood. This method is based on two features of this model :

- Given the parameters $p(j^1, ..., j^Q)$ the probability of $\delta(i)$ given

$(X_1(i), X_2(i)) = (X_1^q(i), X_2^q(i))_{q=1,...,Q}$ is easily deduced from the Bayes rule:

$$P(\delta(i)|X_1(i), X_2(i), p) = \frac{P(\delta(i)|p)P(X_1(i), X_2(i)|\delta(i), p)}{\sum_{\delta(i)\in\{0,1\}^Q} P(\delta(i)|p)P(X_1(i), X_2(i)|\delta(i), p)} \tag{3.4}$$

The term $P(\delta(i)|p) = \frac{1}{2^Q}$ can be simplified and using an elementary vectoriel notation:

$$P(\delta(i)|X_1(i), X_2(i), p) =$$

$$\frac{p(\delta(i)X_1(i) + (1-\delta(i))X_2(i))p((1-\delta(i))X_1(i) + \delta(i)X_2(i))}{\sum_{\delta(i)\in\{0,1\}^Q} p(\delta(i)X_1(i) + (1-\delta(i))X_2(i))p((1-\delta(i))X_1(i) + \delta(i)X_2(i))} \tag{3.5}$$

- Given the $\delta(i)$ the loglikelihood of $(X_1(i), X_2(i))$ may be rewritten on the form :

$$\sum_{\substack{\bar{j} = (j^1, ..., j^Q) \\ \bar{k} = (k^1, ..., k^Q)}} \alpha(\bar{j}, \bar{k})\{ln\ p(j^1, ..., j^Q) + ln\ p(k^1, ..., k^Q)\} \tag{3.6}$$

where $\alpha(\bar{j}, \bar{k}) = P(\delta(i)|X_1(i), X_2(i), p)$

Then given α the $\alpha(\bar{j}, \bar{k})$ this likelihood may be easily maximized (M step) and given p the $\alpha(\bar{j}, \bar{k})$ may easily computed (E step). The EM algorithm is based on a recursive application of these two steps under the convergence.

A bayesian analogous of the EM algorithm is provided by an MCMC treatment of the posterior distribution of the vector p. If the prior probability on p is a Dirichlet distribution, its posterior given $(X_1(i), X_2(i)), \delta(i))i = 1, ..., n$ is also a Dirichlet distribution. Then, samples from p given $(X_1(i), X_2(i), \delta(i))$ are easily generated. Using the previous argument, $(\delta(i))i = 1, ..., n$ given p and $(X_1(i), X_2(i))$ are easily generated and a recursive use of the Gibb sampling algorithm will provided drawns, after convergence from $(p, (\delta(i))i = 1, ..., n)$ given the actual data.

4 A moment estimation: an introductory case

Let us consider a simple case where Q, the number of loci, is equal to 2. The goal of the procedure is to estimate the joint distribution of the alleles on this two loci, described by the

$$p(j^1, j^2) \quad j^1 = 1, ..., r^1 \quad j^2 = 1, ..., r^2$$

where $p(j^1, j^2) \geq 0$ and $\sum_{j^1, j^2} p(j^1, j^2) = 1$. We denote by $p(j^1, .)$ and $p(., j^2)$ the marginal probabilities, ie:

$$\begin{aligned} p(j^1, .) = \sum_{j^2=1}^{r^2} p(j^1, j^2) &= P(\xi_1^1(i) = j^1) \\ &= P(\xi_2^1(i) = j^1) \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} p(., j^2) = \sum_{j^1=1}^{r^1} p(j^1, j^2) &= P(\xi_1^2(i) = j^2) \\ &= P(\xi_2^2(i) = j^2) \end{aligned} \tag{4.2}$$

It is well known that the lack of observation of the phase configuration does not raise any problem for the estimation of these marginal probabilities.

Indeed:

$$\hat{p}(j^1,.) = \frac{1}{2n}\sum_{i=1}^{n}\left\{ \mathbb{1}(X_1^1(i) = j^1) + \mathbb{1}(X_2^1(i) = j^1)\right\} \tag{4.3}$$

and

$$\hat{p}(.,j^2) = \frac{1}{2n}\sum_{i=1}^{n}\left\{ \mathbb{1}(X_1^2(i) = j^2) + \mathbb{1}(X_2^2(i) = j^2\right\} \tag{4.4}$$

where $\mathbb{1}(X_1^1(i) = j^1)$ equal 1 if $X_1^1(i) = j^1$ and 0 else provide consistent estimators if $p(j^1,.)$ and $p(.,j^2)$. This consistency follows from the strong law of large number.

Consider now the statistic:

$$\begin{array}{rl} \hat{A}(j^1,j^2) = \frac{1}{4n}\sum_{i=1}^{n} & \mathbb{1}(X_1^1,(i) = j^1, X_1^2(i) = j^2) \\ & + \mathbb{1}(X_1^1(i) = j^1, X_2^2(i) = j^2) \\ & + \mathbb{1}(X_2^1(i) = j^1, X_1^2(i) = j^2) \\ & + \mathbb{1}(X_2^1(i) = j^1, X_2^2(i) = j^2) \end{array} \tag{4.5}$$

which count all the possible pairs of the alleles on the two loci equal to (j^1, j^2).

It will easily show that the expectation of each terms of the sum is equal to

$$\frac{1}{2}(p(j^1,j^2) + p(j^1,.)p(.,j^2)) \tag{4.6}$$

Let us take for example the expectation of the first term:

$$
\begin{aligned}
E \quad & (\mathbb{1}(X_1^1(i) = j^1, X_1^2(i) = j^2) \\
= \quad & \sum_{\delta(i)\in\{0,1\}^2} P(\delta(i))P(X_1^1(i) = j^1, X_1^2(i) = j^2|\delta(i)) \\
= \quad & \frac{1}{4}\big\{P(X_1^1(i) = j^1, X_1^2(i) = j^2)|\delta^1(i) = 0, \delta^2(i) = 0) \\
& \quad +P(X_1^1(i) = j^1, X_1^2(i) = j^2|\delta^1(i) = 1, \delta^2(i) = 0) \\
& \quad +P(X_1^1(i) = j^1, X_1^2(i) = j^2|\delta^1(i) = 0, \delta^2(i) = 1) \\
& \quad +P(X_1^1(i) = j^1, X_1^2(i) = j^2|\delta^1(i) = 1, \delta^2(i) = 1)\big\} \\
= \quad & \frac{1}{4}\big\{P(\xi_1^1(i) = j^1, \xi_1^2(i) = j^2) \\
& \quad +P(\xi_1^1(i) = j^1, \xi_2^2(i) = j^2) \\
& \quad +P(\xi_2^1(i) = j^1, \xi_1^2(i) = j^2) \\
& \quad +P(\xi_2^1(i) = j^1, \xi_2^2(i) = j^2)\big\} \\
= \quad & \frac{1}{4}\big\{p(j^1, j^2) + p(j^1, .)p(., j^2) + p(j^1, .)p(., j^2) + p(j^1, j^2)\big\} \qquad (4.7)
\end{aligned}
$$

Then, using the strong law of large number

$$\hat{A}(j^1, j^2) \underset{a.s}{\to} \frac{1}{2}(p(j^1, j^2) + p(j^1, .)p(., j^2)) \qquad (4.8)$$

The intuition behind this result is that a pair of two observed alleles on two loci has a (marginal) probability $\frac{1}{2}$ to be on the same locus and then to be generated with a probability $p(j^1, j^2)$ and a (marginal) probability $\frac{1}{2}$ to be on different chromosoms and then to be independently generated.

Remark: A more tedious computation would show that the behavior of $\hat{A}(j^1, j^2)$ is identical if the observed data are constructed using the order statistics on each locus. Actually one can remark that $\hat{A}$ is chosen such that its value is invariant by permutation of the data between the two phases.

Following (4.3) and (4.8), a consistent estimation of $p(j^1, j^2)$ for any value of j^1, j^2 is given by:

$$\hat{p}(j^1, j^2) = 2\hat{A}(j^1, j^2) - p(j^1, .)\hat{p}(., j^2) \qquad (4.9)$$

This argument may be extended to three loci. Let us now consider $\hat{A}(j^1, j^2, j^3)$ equal to the total number of possible triplets of alleles j^1, j^2

and j^3 observed for each individual, divided by $8n$. Using an equivalent argument to the two loci case (a general presentation will be given in the next section) we can check that:

$$\begin{aligned}\hat{A}(j^1,j^2,j^3) \to \tfrac{1}{4}\{&j^1,j^2,j^3) + p(j^1,j^2,.)p(.,.,j^3) \\ &+p(j^1,.,j^3)p(.,j^2,.) + p(j^1,.,.)p(.,j^2,j^3)\}\end{aligned} \tag{4.10}$$

where e.g. $p(j^1,j^2,.)$ is the marginal distribution on the two first loci.

Using (4.3),(4.4) and (4.9) the marginal probabilities on a single locus or on two loci can be estimated and we obtain a consistent estimator of $p(j^1,j^2,j^3)$ by:

$$\begin{aligned}\hat{p}(j^1,j^2,j^3) &= 4\hat{A}(j^1,j^2,j^3) - \hat{p}(j^1,j^2,.)p(.,.,j^3) \\ &- \hat{p}(j^1,.,j^3)\hat{p}(.,j^2,.) - \hat{p}(j^1,.,.)\hat{p}(.,j^2,j^3)\end{aligned} \tag{4.11}$$

5 Moment estimation : the general case

For any individual we consider the following random element:

$$Z_{j_1,...,j_Q}(i) = \frac{1}{2^Q}\sum_{\tau\in T} I\!\!I(X^1_{\tau(1)}(i) = j^1,...,X^Q_{\tau(Q)}(i) = j^Q)$$

where T is the set of all functions from $\{1,...,Q\}$ into $\{1,2\}$ whose cardinality is 2^Q. Intuitively $Zj_1,...,j_Q(i)$ counts the number of Q-uple equal to $j^1,...,j^Q$ obtained by all the possible selections of one element in each pair of alleles for an individual i. Thanks to the strong law of large number, the empirical mean converges to the theoretical mean, i.e. :

$$\hat{A}(j_1,...,j_Q) = \frac{1}{n}\sum_{i=1}^{n} Z_{j^1,...,j^Q}(i) \underset{a.s}{\to} E(Z_{j^1,...,j^Q}(i)) \tag{5.1}$$

We now compute the theoretical mean, which does not depend on the individual and we drop out for simplicity the index i.

$$\begin{aligned}E(Zj^1,...,j^Q) &= \frac{1}{2^Q}\sum_{\tau\in T} E(I\!\!I(X^1_{\tau(1)} = j^1,...,X^Q_{\tau(Q)} = j^Q)) \\ &= \frac{1}{(2^Q)}\sum_{\tau\in T}\sum_{\delta\in D} P(X^1_{\tau(1)} = j^1,...,X^Q_{\tau(Q)} = j^Q|\delta)\end{aligned} \tag{5.2}$$

where $\delta = (\delta^1, ..., \delta^Q)$ may be any element of the set D of all the function from $\{1, ..., Q\}$ to $\{0, 1\} (= \{0, 1\}^Q)$.

Then:

$$E(Zj^1, ...j^Q) = \frac{1}{(2^Q)} \sum_{\tau \in T} \sum_{\delta \in D} \left(\sum_{\substack{j^1,...,j^Q \\ s.t.\ \tau(q)-1 \neq \delta^q}} p(j^1, ..., j^Q) \right) \left(\sum_{\substack{j^1,...,j^Q \\ s.t.\ \tau(q)-1 \neq \delta^q}} p(j^1, ..., j^Q) \right) \tag{5.3}$$

In the first parenthesis the sum of $p(j^1, ..., j^Q)$ is compute with the index j^q only if $\tau(q) - 1 \neq \delta^q$.

By regrouping equal terms in the sum, we get

$$E(Z_{j^1,...,j^Q}) = \frac{1}{2}$$

This expression will be denote by $\lambda_{j^1,...,j^Q}(p)$ where p is the vector of probabilities to be estimated.

$$\begin{aligned} &= \frac{1}{2^Q} \sum_{\delta \in D} \left(\sum_{\substack{j^1,...,j^Q \\ \delta^q = 0}} p(j^1, ..., j^Q) \right) \left(\sum_{\substack{j^1,...,j^Q \\ \delta^q = 1}} p(j^1, ..., j^Q) \right) \\ &= \lambda j^1, ..., j^Q(p) \end{aligned}$$

where p is the vector of probabilities to be estimated.

Formulae (4.3), (4.6) and (4.10) give particular cases of this results for $Q = 1, 2$ and 3. For the case $Q = 4$ we obtain:

$$\begin{aligned} E(Zj^1, j^2, j^3, j^4) = \frac{1}{16} \Big\{ 2 \Big(& p(j^1, j^2, j^3, j^4) + p(j^1, j^2, j^3, .) p(., ., ., j^4) \\ & + p(j^1, j^2, ., j^4) p(., ., j^3, .) + p(j^1, ., j^3, j^4) p(., j^2, ., .) \\ & + p(., j^2, j^3, j^4) p(j^1, ., ., .) + p(j^1, j^2, ., .) p(., ., j^3, j^4) \\ & + p(j^1, ., ., j^4) p(., j^2, j^3, .) + p(j^1, ., j^3, .) p(., j^2, ., j^4) \Big) \Big\} \end{aligned}$$

Then a recursive estimation method may be easily implemented to solve the set of moment conditions (5.3).

6 Asymptotic distribution

This approach belongs to the family of estimation using estimating function or moment estimation. The estimation of p is constructed in order to match the empirical moments $\hat{A}(j^1, ..., j^Q)$ and the theoretical moments $\lambda j^1, ...j^Q(p)$ depending on the parameters of interest. The number of conditions is $r^1 \times ... \times r^Q$ (the number of possible combinations of alleles on the Q locus on a chromosom), but one condition may be dropped out because the $\lambda j^1, ..., j^Q(p)$ are themselves probabilities and sum to one. In other words, if we stack the $Zj^1, ..., j^q(i)$ in a $(r^1 \times ... \times r^Q)_{-1}$ vector $Z(i)$ and the $\lambda j^1, ..., j^Q(p)$ in a $(r^1 \times ... \times r^Q)_{-1}$ vector $\lambda(p)$, the estimation of p we propose is obtained by solving:

$$\frac{1}{n}\sum_{i=1}^{n} Z(i) = \lambda(p) \tag{6.1}$$

Let us underline that the recursive method we have proposed is only a resolution method of this system and the equations corresponding to subsets of the Q locus are contained in the set (6.1). In a very compact way our estimator may be summarized by

$$\hat{p} = \lambda^{-1}(\frac{1}{n}\sum Z(i)) \tag{6.2}$$

and the existence of the inverse follows from the recursive solution we have introduced. The functions λ and λ^{-1} are continuously differentiable. Then:

i) $\hat{p} \underset{a.s}{\rightarrow} p$ solution of $E(Z(i)) = \lambda(p)$

ii) $\sqrt{n}(\hat{p} - p) \Longrightarrow N(0, \Sigma)$

where $\Sigma = \left(\frac{\partial \lambda}{\partial p'}\right)^{-1} VarZ(i) \left(\frac{\partial \lambda'}{\partial p}\right)^{-1}$

(see Serfling (1980)).

The variance Σ can be easily estimated : $VarZ(i)$ may be estimated by the empirical variance of the $Z(i)$ and $\frac{\partial \lambda}{\partial p}$ the matrix of partial derivatives of λ is a matrix of functions of p. The general computation of this matrix is very tedious (but may be immediately realized by computer software as Mapple or Mathematica) and we just illustrate this asymptotic distribution by an example.

Consider an example with two locus and two alleles on each locus. The probability of interest are defined by:

locus 1 \ locus 2	allele 1	allele 2
allele 1	p_{11}	p_{12}
allele 2	p_{21}	p_{22}

Actually this case only involves three parameters because $p_{11}+p_{12}+p_{21}+p_{22}=1$ and this constraint is introduced by replacing p_{22} by $1-(p_{11}+p_{12}+p_{21})$. In that case we have $\lambda(p)=(\lambda_{11}(p),\lambda_{12}(p),\lambda_{21}(p))$ where:

$$\lambda_{11}(p) = \frac{1}{2}(p_{11}+(p_{11}+p_{12})(p_{11}+p_{21})) \quad (6.3)$$

$$\lambda_{12}(p) = \frac{1}{2}(p_{12}+(p_{11}+p_{12})(p_{12}+p_{22})) \quad (6.4)$$

$$\lambda_{21}(p) = \frac{1}{2}(p_{21}+(p_{11}+p_{21})(p_{21}+p_{22})) \quad (6.5)$$

and the matrix of partial derivatives is:

$$\frac{\partial\lambda}{\partial p}=\frac{1}{2}\begin{pmatrix} 1+(p_{11}+p_{12})+(p_{11}+p_{21}) & p_{11}+p_{21} & p_{11}+p_{12} \\ p_{22}-p_{11} & 1+(p_{21}+p_{22}) & -(p_{11}+p_{12}) \\ p_{22}-p_{11} & -(p_{11}+p_{21}) & 1+(p_{21}+p_{22}) \end{pmatrix}$$

This matrix is estimated by replacing the unknown values of p by their estimated values.

In practice, bootstrap confidence interval may be used in order to improve asymptotic distribution. This approach will be discussed in section 9.

7 Correction for negative probabilities

The asymptotic analysis done in section 6 is implicitely based on the assumption that the true value p of the parameter is an interior point of the set of all possible values of this parameter, namely the simplex of probabilities of $(r^1\times ...\times r^Q)-1$ dimensions. If this assumption is satisfied (i.e. $p(j^1,...,j^Q)>0\ \forall j^1,...,j^Q$) the estimator $\hat{p}(j^1,...,j^Q)$ is necessarily positive for n sufficiently large because this estimator is consistent.

In practice our estimation method constructs estimations $\hat{p}(j^1, ..., j^Q)$ which verify

$$\sum_{j^1,...,j^Q} \hat{p}(j^1, ..., j^Q) = 1$$

but which may fail to be positive. Remember that, in the two loci cases we have (see 4.8)

$$\hat{p}(j^1, j^2) = 2A(j^1, j^2) - \hat{p}(j^1, .)\hat{p}(., j^2)$$

and the $2A(j^1, j^2)$ may be equal to zero (if the pair (j^1, j^2) is never observed) or smaller to the product of the estimated marginal probabilities. In that case, we suggest to transform our estimator by the following rule :

i) put equal to 0 any probability estimated by a negative number

ii) renormalized the positive probabilities by dividing their sum.

As noted below, this modification does not affect the asymptotic behavior of the estimator if the true probabilities are positive. In that case, the correction is only a small sample improvement of the estimator. However, if the some of the true probabilities are zero, the asymptotic distribution of our estimator, as well as the distribution of the maximum likelihood estimator, is definitly more complex and will not be consider in this paper (see Andrews 1999).

8 A Monte Carlo simulation

In order to evaluate the small sample performance if our estimator and to compare with other approaches we have done a Monte Carlo simulation using the following design.

1) We consider two loci and two alleles for each locus and the simulation are generated using the values:

locus 1 \ locus 2	allele 1	allele 2
allele 1	$0,1$	$0,3$
allele 2	$0,2$	$0,4$

2) For different sample sizes (100, 500, 1000 and 10000) a sample of pairs of chromosom is generated and four estimations are performed

- $\hat{q}(j^1, j^2)$ is the maximum likelihood estimation using haplotypes which represents the "best" estimation (inacessible in practice but accessible using simulation)
- $\hat{p}(j^1, j^2)$ is the estimation consider in the paper (see section 4). Negative probability are never obtained by the simulation.
- $\hat{p}_{EM}(j^1, j^2)$ and $\hat{p}_{Max}(j^1, j^2)$ denotes two evaluations of the maximum likelihood estimation : the first one is based on the EM algorithm (using as as shopping role the variation of the four parameter is lower than 10^{-6}). The second is a direct resolution of the first order condition of the likekihood maximization by the procedure "solve" of matlab. This two methods only differs by the numerical computation of the maximum likelihood estimator.

3) This experiment has been reproduced 100 times and results are summarized by the root meansquare errors of each parameter for each sample size.

The results are summarized in table I. For almost all the cases (except p_{11} for a sample size of 100) our estimator is superior to the two numerical evaluations of the maximum likelihood, even for a sample size as large as 10000. For large sample size the difference between our estimator and the "best" possible estimator $\hat{q}$ is very low.

In order to check the sensitivity of the different estimation to low values of true probabilities.

9 Application to the relation between the microsatellite MOGC and gene A

The original motivation of this research was to analyse the capacity of a set of microsatellites to predict groups of HLA types. In this paper we just present a preliminary step of this study concentrated on a single microsatellite MOGC and the A element of the HLA system. We consider a sample (of size 2117)[1] of phenotypes used for the estimation of the joint distribution of size 2117 of MOGC and A on a single chromosom.

[1]This sample was randomly extracted from the France Greffe de Moelle Registry. In this data set missing data are reconstructed by answering homozygoty

The result of our estimation is given in table II where "0" denotes pairs of MOGC/A never observed. Probability values are nounded off.

The precision of this estimation result is analysed by a non parametric bootstrap. From the original sample we have contracted 1000 samples by random drawing with replacement. Each sample is used for a new estimation of the joint probability (see Hall (1999)).

We just illustrate the power of this analysis by two examples. We have constructed the bootstrap distribution of two measures of the linkage desequilibrium. The first one is the entropy measure defined by

$$I = \sum_{j,k} p_{ik} \ln \frac{p_{jk}}{p_{j.} p_{k}} \tag{9.1}$$

where j is the index of possible alleles of MOGC, k is the index of possible alleles of A, p_{jk} is the joint probability and p_j and p_k the marginal probabilities.

The estimated value of I (9.1) is 0,9701. The bootstrap is 0,9676 and a confidence interval of I at 95 % is [0,9215; 1,0173]. The distribution of I is given by the histogram in table III.

It is well know that entropy has some undurable features and a better association measure is provided by Hellinger distance between the joint distribution and the product of its marginals, i.e.

$$H = \frac{1}{\sqrt{2}} \left[\sum_{j,k} \left(\sqrt{p_{ik}} - \sqrt{p_{j.p.k}} \right)^2 \right]^{\frac{1}{2}}$$

In particular, by construction, it is normalized in order to be between 0 and 1 where 0 is equivalent to independence. The actual estimated value of H is 0,4270. The bootstrap mean is 0,4271 and a confidence interval is [-0,4033 ; 0,4520] different histograms of bootstrap distribution of this linkage desequilibrium measure is given in table IV.

10 Conclusion

This paper presents a moment estimation of the joint distribution to the alleles on several loci on a chromosom using phenotypes data. This estimator is not constructed as the limit of a recursive algorithm (dependent on starting point and on stopping rule) but is immediately computable. This estimator is strongly consistant and asymptotically normal and then its does not reach the efficiency bound as maximum likelihood, Monte-Carlo simulations shows that

it performs better is some small sample cases. Moreover a bootstrap analysis of the distribution of the estimator is possible thank to its efficiency in term of computation time. We have illustrate the power of our methodology by an empirical analysis of linkage desiquilibrium between MOGC and gene A. Two extensions of this approach are in project. First the computation of the estimator in case of numerous loci may be improve by an optimisation of the several countings required for the estimation and secondly asymptotic properties of Maximum Likelihood Estimation and of our estimator should be studied in case where the true joint probability has elements exactly equal to 0. Then a major hypothesis of MLE is not satisfied (namely the true parameter is an interior point of the parametric space) and optimality of MLE is not a large warrants.

Table I

	Taille échantillon	1 000	10 000
	RMSE		
^q	**q11**	**0,0024**	**0,00066**
^P	**pe11**	**0,0034**	**0,00096**
^Pem	**pe11**	**0,0042**	**0,0026**
^Pmax	**pe11**	**0,003**	**0,0013**
	q12	**0,0102**	**0,0031**
	pe12	**0,0106**	**0,0032**
	pe12	**0,0116**	**0,0033**
	pe12	**0,0117**	**0,0035**
	q21	**0,0048**	**0,0013**
	pe21	**0,0053**	**0,0015**
	pe21	**0,008**	**0,0049**
	pe21	**0,0085**	**0,0062**
	q22	**0,0106**	**0,0032**
	pe22	**0,0109**	**0,0033**
	pe22	**0,0131**	**0,0075**
	pe22	**0,0137**	**0,0088**
	2 caractères et 2 modalités		
	0,01	**0,35**	
	0,04	**0,6**	

Table II

Joint Distribution of MOGC and HLA-A on a single chromosom

MOGC\HLA-A	1	2	3	9	10	11	23	24	25	26	28	29	30	31	32	33	34	36	43	66	68	69	74	80
121	0	0,076	0,003	0,000	0,000	0,001	0,003	0,015	0,016	0,035	0,001	0,052	0,002	0,004	0,002	0,010	0,000	0	0,000	0,002	0,014	0	0,000	0
123	0	0,000	0	0	0	0	0,001	0	0	0,000	0	0	0	0	0,000	0	0	0	0	0	0	0	0	0
125	0,000	0,000	0	0,000	0	0,000	0	0,000	0	0	0	0	0,000	0	0,001	0	0	0	0	0	0,000	0	0	0
127	0	0	0,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,000	0	0	0
129	0,005	0,016	0,117	0,000	0	0,002	0	0,022	0	0	0,000	0,005	0,000	0,001	0,014	0,003	0,001	0,000	0	0,001	0	0,000	0	0,001
131	0,005	0,169	0,003	0	0	0	0,001	0,004	0,001	0,001	0,001	0,001	0	0,001	0,013	0,001	0	0	0	0	0,012	0,000	0,000	0
133	0,002	0,002	0,000	0	0	0,000	0	0	0	0	0	0,001	0,011	0,016	0,001	0	0	0,000	0,000	0	0,000	0	0	0
135	0	0	0,002	0,000	0	0,051	0,020	0,001	0,001	0	0	0	0,001	0,000	0,000	0,001	0	0	0	0,000	0,000	0	0	0,000
137	0	0,003	0,000	0,000	0	0,002	0,001	0,024	0,000	0,001	0,000	0	0,016	0,004	0,001	0	0,000	0	0	0	0,010	0,000	0	0,000
139	0,001	0,001	0,000	0	0	0,000	0	0	0	0,000	0	0,000	0,002	0,000	0	0	0	0	0	0	0,000	0	0	0
141	0,000	0,000	0	0	0	0	0	0,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
143	0,052	0,003	0	0	0	0	0	0,002	0	0	0,000	0,001	0,001	0	0,000	0,001	0,000	0,000	0	0	0	0,000	0	0
145	0,001	0,001	0	0	0	0	0	0,003	0	0,000	0	0,000	0,000	0,000	0,000	0	0	0	0	0	0	0	0	0
147	0,061	0,006	0,002	0	0	0,001	0,000	0,030	0,000	0	0,000	0,000	0	0,001	0	0	0	0,000	0	0	0,001	0,000	0,000	0,000
149	0,000	0	0	0	0	0	0	0,001	0	0	0	0	0,000	0	0	0	0	0	0	0	0,001	0	0	0
151	0,000	0,000	0,000	0	0	0	0	0,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
153	0,000	0	0	0	0	0	0	0,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0,13	**0,28**	**0,13**	**0,00**	**0,00**	**0,06**	**0,03**	**0,10**	**0,02**	**0,04**	**0,00**	**0,06**	**0,04**	**0,03**	**0,03**	**0,02**	**0,00**	**0,00**	**0,00**	**0,00**	**0,04**	**0,00**	**0,00**	**0,00**
	0,129	0,28	0,13	0,00	0,00	0,06	0,03	0,10	0,02	0,04	0,00	0,06	0,04	0,03	0,03	0,02	0,00	0,00	0,00	0,00	0,04	0,00	0,00	0,00

Table III

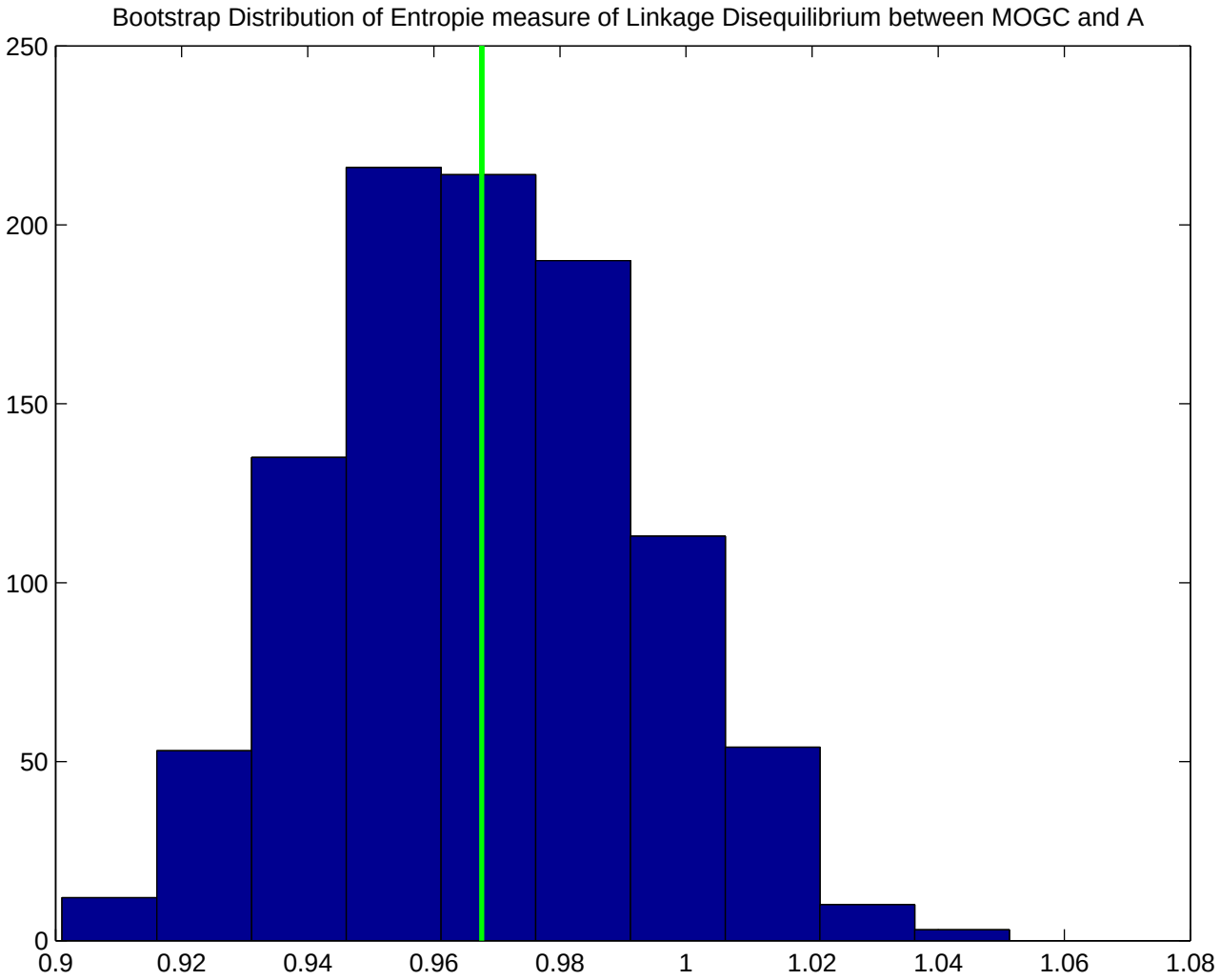
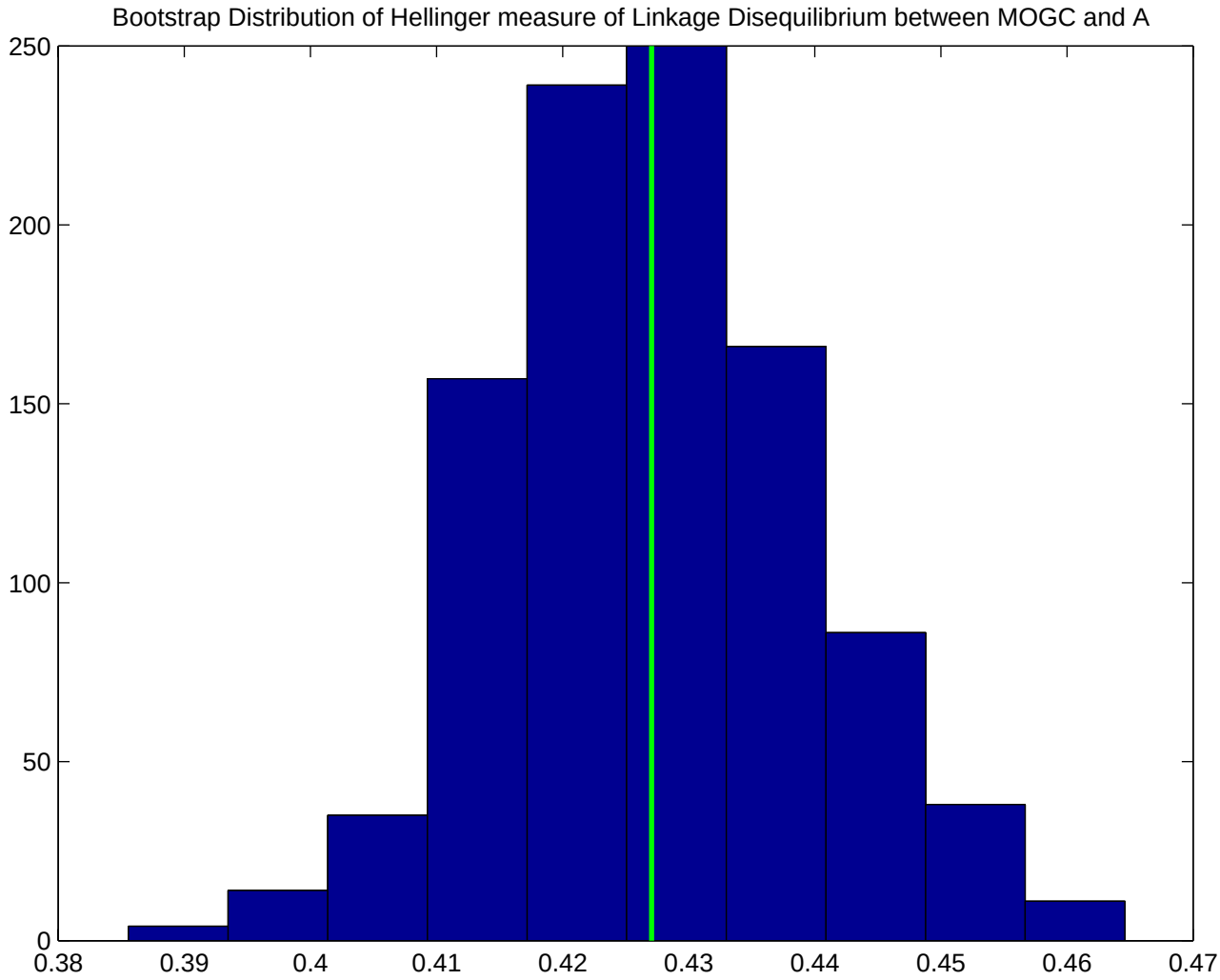

Bootstrap Distribution of Entropie measure of Linkage Disequilibrium between MOGC and A
250
200
150
100
50
0
0.9
0.92
0.94
0.96
0.98
1
1.02
1.04
1.06
1.08

Table IV


Bootstrap Distribution of Hellinger measure of Linkage Disequilibrium between MOGC and A
250
200
150
100
50
0
0.38
0.39
0.4
0.41
0.42
0.43
0.44
0.45
0.46
0.47

References

Andrews, D. (1999), Estimation when a Parameter is on the Boundary, Econometrica, Vol 67, 1341-1383.

Clark, A. (1990), Inference of Haplotypes from PCR-amplified Samples of Diplaid Populations, Md. Bid Eva. ((2): 111-112.

Dudbridge, F. B. Koeleman, J. Todd, D. Clayton (2000) Unbiased Applications of the Transmission/Disequilibrium Test to Multilocus Haplotypes, Am. J. Hum. Genet; 66: 2009-2012.

Excoffier, L. M. Slatkin (1995), Maximum Likelihood Estimation of Molecules Haplotype Frequencies in a Diploid Population, Mal. Bid. Eval. 12(5): 921-927.

Hall, P. (1999), The Bootstrap and Edgeworth Expression, Verlag, New York.

Niu, T. Z. Qin, X. Xu, J. Liu (2002), Bayesian Haplotype Inference for Multiple Linked Single. Nucleodide Polymorphims, Am.J. Hum. Genet., 70: 157-169.

Risch, N., K. Merikangas (1966), Science, New series, Volume 273, Issue 5281, 1516-1517.

Serfling, R. (1980), Approximation Theorems of Mathematical Statistics, Wiley.

Single, R.M., D. Meyer, J. Hollenback, M.P. Nelson, J. Noble, H. Erlich, G. Thomson (2002), Haplotype Frequency Estimation in Patient Populations: The effect of Departures from Hardy-Weinberg Proportions and Collasping over a Locus in the HLA Region, Genetic Epidemiology 22: 186-195.

Stephens, M. N. Smith, P. Donnelly (2001), A New Statistical Method for Haplotype Reconstruction from Population Data, Am. J. Hum. Genet. 68:978-989.

Wang, S. K.K. Kidd, H. Zhao (2003), On the use of DNA Pooling to Estimate Haplotype Frequencies, Genetic Epidemiology, 24: 74-82.