*Genet. Sel. Evol.* (1989) 21, 341-357



**Original article**

# Comparison of four statistical methods for detection of a major gene in a progeny test design

P. Le Roy[1], J.M. Elsen[1] and S. Knott[2]

[1] *Institut National de la Recherche Agronomique, centre de recherches de Toulouse, station d'amélioration génétique des animaux, Auzeville 31326 Castanet-Tolosan Cedex, France*
[2] *AFRC, IAPGR, Edinburgh Research Station, Roslin, Midlothian, EH 25 9PS, UK*

(received 17 August 1988, accepted 23 January 1989)

**Summary** – In livestock improvement it is common to design a progeny test of sires in order to estimate their breeding values. The data recorded for these estimate are useful for the detection of major genes. They are the $n.m$ performances $Y_{ij}$ of $m$ progeny $j$ of $n$ sires $i$. These data need to be corrected for the polygenic influence of the sire on its progeny (sire $i$ effect $U_i$). Four statistical tests of the segregation of a major gene are compared. The first ($l_{SA}$ for "segregation analysis") is the classical ratio of the likelihoods under $H_0$ (no major gene) and $H_1$ (a major gene is segregating). The parameters describing the population (means and standard deviations within genotype) are estimated by maximizing the marginal likelihood of the $Y_{ij}$. The other statistics studied are approximations of this $l_{SA}$ statistic where the sire $i$ effect ($U_i$) is considered as a fixed effect ($l_{FE}$ statistic) or, following Elsen *et al.* (1988) and Höschele (1988), where the parameters, and $U_i$, are estimated maximizing the joint likelihood of $U_i$ and $Y_{ij}$ ($l_{ME1}$ and $l_{ME2}$ statistics). Simulation studies were done in order to describe the distribution of these statistics. It is shown that $l_{SA}$ and $l_{ME1}$ are the most powerful test, followed by $l_{ME2}$, whose relative loss of power ranged between 20 and 40%, depending on the $H_1$ case studied, when 400 progeny are measured ($n = m = 20$). The segregation analysis, based on direct maximization of the likelihood, required 30 times more computation time than the $l_{ME}$ test using an EM algorithm.

**major gene – segregation analysis – statistical test**

**Résumé – Comparaison de quatre méthodes statistiques pour la détection d'un gène majeur dans un test sur descendance.** *Il est fréquent, en sélection, de tester sur descendance, des mâles, afin d'estimer leur valeur génétique. Les données recueillies dans ce but peuvent être utilisées afin de mettre en évidence un gène majeur. Elles sont constituées des* n.m *performances* $Y_{ij}$ *de* m *descendants* j *de* n *mâles* i. *Ces données doivent être corrigées pour l'effet polygénique du père* ($U_i$) *sur ses descendants. Quatre tests statistiques de mise en évidence d'un tel gène majeur sont comparés. Le premier* ($l_{SA}$ *pour "segregation analysis") est le rapport classique des vraisemblances sous* $H_0$ *(pas de gène majeur) et sous* $H_1$ *(existence d'un gène majeur). Les paramètres caractéristiques de la population (moyennes et écarts types intragénotype) sont estimés en maximisant la vraisemblance marginale des* $Y_{ij}$. *Les autres statistiques de tests sont des approximations de* $l_{SA}$ *pour lesquelles, soit l'effet père* $U_i$ *est considéré comme un effet fixé (test* $l_{FE}$) *soit, comme proposé par Elsen* et al. *(1988) et Höschele (1988), les paramètres, et* $U_i$, *sont obtenus en maximisant la vraisemblance conjointe des* $Y_{ij}$ *et des* $U_i$ *(test* $l_{ME1}$

*et* $l_{ME2}$*). Nous avons réalisé des simulations afin de décrire les distributions de ces tests.* $l_{SA}$ *et* $l_{ME1}$ *sont les tests les plus puissants, suivi par* $l_{ME2}$*, dont la perte relative de puissance varie entre 20 et 40% selon l'hypothèse* $H_1$ *étudiées, quand 400 descendants sont mesurés (*n = m *=20). L'analyse de ségrégation, réalisée par maximisation directe de la vraisemblance, demande 30 fois plus de temps de calcul que les tests* $l_{ME}$ *réalisés l'aide d'un algorithme EM.*

***gène majeur – analyse de ségrégation – test statistique***

## INTRODUCTION

In recent years, several genes having major effects on commercial traits have been identified. The dwarf gene in poultry (Mérat & Ricard, 1974), the halothane sensitivity gene in pigs (Ollivier, 1980), the Booroola gene in sheep (Piper & Bindon, 1982), or the double muscling gene in cattle (Ménissier, 1982) are notable examples.

These discoveries, as well as improvement of transgenic techniques, have stimulated interest in new techniques for detection of single genes. Various tests have been described concerning livestock (Hanset, 1982). Their general principle is that the within family distribution of the trait depends on the parents' genotypes, and therefore varies from one family to another. These methods involve simple computations but are not powerful. Concurrently, segregation analysis in complex pedigrees was developed in human genetics (Elston & Stewart, 1971) by comparing the likelihoods of the data under different trait transmission models. These methods are much more powerful than the previous ones, but involve much computation. They require numerical simplification to deal with the population structure of farm animals. Additionally, the known properties of the test statistics, a likelihood ratio test, are only asymptotic, which raises the question of their validity when applied to samples of limited size.

In livestock improvement it is common to use progeny tests where males are mated to large numbers of females. Concentrating on this simple family structure the present paper tries to give some elements of a solution to the problems of simplification and validity. Four methods are compared on simulated data.

## METHODS

The four methods considered rely upon the same information structure and the same type of test statistics.

### *Experimental design*

The data are simulated according to a hierarchical and balanced family structure: one sample consists of $n$ sire families ($i = 1, ...n$) with $m$ mates per sire ($j = 1, ...m$) and one offspring per dam. Sires and dams are assumed to be unrelated. Only offspring are measured, with one $Y_{ij}$ datum per animal.

## *Models and notations*

### Models

The $Y_{ij}$ performances are considered under the two following models:

*General hypothesis ($H_1$): "mixed inheritance"*

In this model a monogenic component is added to the assumed polygenic variation.

When two alleles $A$ and $a$ are segregating at a major locus, three genotypes are possible ($AA$, $Aa$, $aa$) which we shall respectively denote 1, 2, 3. Sires are of genotype $s(s = 1, 2, 3)$ with probability $P_s$. Dams transmit to their offspring allele $A$ with a probability $q$ and allele $a$ with a probability $1 - q$. Conditional on its genotype $t(t = 1, 2, 3)$, the $ij$th progeny has the performance $Y_{ij}^t$. The following linear model can be formulated.

$$Y_{ij}^t = \mu_t + U_i + E_{ij}$$

Where $\mu_t$ is the mean value of the performances of genotype $t$ progeny.

$U_i$ is the sire $i$ random effect, assumed to be independent of the genotype $t$ and normally distributed with a mean 0 and a variance $\sigma_u^2$.

$E_{ij}$ is the residual random effect, assumed to be independent of the genotype $t$ and normally distribued with a mean 0 and a variance $\sigma_e^2$.

$U_i$ and $E_{ij}$ are assumed to be independent.

Concerning production traits of livestock, the proportion of variance explained by polygenic effects has been generally estimated in many populations. Thus, we shall assume known *a priori* the heritability of the trait, $h^2$, defined as:

$$h^2 = 4\sigma_u^2 \mid (\sigma_u^2 + \sigma_e^2)$$

so that sires are assumed to be unselected.

The model thus defined on seven parameters:

$$\mu_1, \mu_2, \mu_3, \sigma_e, q, p_1, p_2 \ (p_3 = 1 - p_1 - p_2)$$

*This hypothesis ($H_0$): "polygenic inheritance".*

Null subhypothesis, to be tested against the general model, is fixed by $\mu_1 = \mu_2 = \mu_3 = \mu_0$:

$$Y_{ij} = \mu_0 + U_i + E_{ij}$$

Where $\mu_0$ is the general mean of the performances. $U_i$ and $E_{ij}$ have the same definition as under $H_1$.

### Matrix notation

Let $\mathbf{S}$ be the vector of the genotypes of the $n$ males $\mathbf{S} = (S_1, \ldots, S_i, \ldots, S_n)$ and $\boldsymbol{s} = (s_1, \ldots\ s_i, \ldots\ s_n)$ one realization of $\mathbf{S}$.

$\mathbf{Y}_i$ be the vector of the $m$ performances of the $i$th sire's progeny: $\mathbf{Y}_i = (Y_{il}, \ldots T_{ij}, \ldots Y_{im})$, and $\boldsymbol{y}_i$ the vector of realizations of $\mathbf{Y}_i$.

$\mathbf{T}_i$ the vector of order $m$ of the genotypes at the major locus of the $i$th sire's progeny: $\mathbf{T}_i = (T_{i1}, \ldots T_{ij}, \ldots T_{im})$. Three realizations being possible for $T_{ij}$, $3^m$ different realizations $\boldsymbol{t}_i$ of $\mathbf{T}_i$ are possible. Prob $(\mathbf{T}_i = \boldsymbol{t}_i | s_i)$ is the probability of the realization of the genotypes vector $\boldsymbol{t}_i = (t_{il}, \ldots t_{ij}, \ldots t_{im})$ when sire $i$ is of genotype $s_i$.

$\boldsymbol{\mu}$ the vector of genotype means:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

Given $\mathbf{E}_i$, the vector of order $m$ of residuals, the vector $\mathbf{Y}_i$ can be written under $H_0$:

$$\mathbf{Y}_i = \mathbf{X}.\mu_0 + \mathbf{Z}.U_i + \mathbf{E}_i$$

where $\mathbf{X}$ and $\mathbf{Z}$ are two matrices of order $m \times 1$, whose elements all equal 1,

under $H_1$:

$$\mathbf{Y}_i = \mathbf{X}_{iti}.\boldsymbol{\mu} + \mathbf{Z}.U_i + \mathbf{E}_i$$

where $\mathbf{X}_{iti}$ is the $m \times 3$ incidence matrix for the fixed effects of the model, when the realization of the genotypes of the sire $i$ progeny is $\mathbf{t}_i$.

The $\mathbf{V}_i$ covariance matrix for the performances $Y_{ij}$ of the sire $i$ family is:

$$\mathbf{V}_i = \begin{pmatrix} \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & \ldots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \ldots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \ldots & \sigma_u^2 + \sigma_e^2 \end{pmatrix} = \mathbf{V}=\mathbf{Z}.\mathbf{D}.\mathbf{Z}'+\mathbf{R}$$

with $\mathbf{D} = \sigma_u^2$ and $\mathbf{R}$ the diagonal $m \times m$ matrix $\mathbf{R}= \sigma_e^2 \cdot \mathbf{I}_m$.

### *General expression of the likelihood ratio test (LR test)*

The test statistic is based on the ratio of the likelihoods under $H_0(M_0)$ and under $H_1(M_1)$, or an estimate of this ratio. In practice the test statistic considered is: $l = -2.\log\,(M_0/M_1)$. With our notation, and given the preceding hypothesis, $M_0$ is:

$$M_0 = \prod_{i=1}^{n} f_0(\boldsymbol{y}_i)$$

with

$$f_0(\boldsymbol{y}_i) = \frac{1}{\sqrt{2\pi^m |\mathbf{V}|}} \exp\,(-\frac{1}{2}\,(\boldsymbol{y}_i - \mathbf{X}_{\mu 0})'\,\mathbf{V}^{-1}\,(\boldsymbol{y}_i - \mathbf{X}_{\mu 0}))$$

and $M_1$ is:

$$M_1 = \prod_{i=1}^{n} f_1(\boldsymbol{y}_i)$$

with

$$f_1(y_i) = \sum_{s_i=1}^{3} p_{s_i} \sum_{t_i} Prob(\mathbf{T}_i = t_i | s_i) \frac{1}{\sqrt{2\pi^m |\mathbf{V}|}}$$
$$exp(-\frac{1}{2}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})'\mathbf{V}^{-1}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu}))$$

The four proposed methods are all based on the two following equalities:

$$\frac{1}{\sqrt{2\pi^m |\mathbf{V}|}} exp(-\frac{1}{2}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})'\mathbf{V}^{-1}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})) =$$
$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left(-\frac{1}{2}\left(\frac{u_i}{\sigma_u}\right)^2\right) \frac{1}{\sqrt{2\pi^m |\mathbf{R}|}}$$
$$exp(-\frac{1}{2}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}u_i)'\mathbf{R}^{-1}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}u_i))du_i \quad (1)$$

and:

$$exp(-\frac{1}{2}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})'\mathbf{V}^{-1}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu})) =$$
$$exp\left(-\frac{1}{2}\left(\frac{\hat{u}_i}{\sigma_u}\right)^2\right) exp(-\frac{1}{2}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}\hat{u}_i)'\mathbf{R}^{-1}(y_i - \mathbf{X}_{it_i}\boldsymbol{\mu} - \mathbf{Z}\hat{u}_i)) \quad (2)$$

Where $\hat{u}_i$ is the mode of the distribution of $U_i$ given $\mathbf{Y}_i$ and the genotypes $\mathbf{t}_i$. Formula (2) results from the equality of mode and expectation for symetrical distributions.

### *Definition and interests of the four proposed methods*

The differences between the four methods concern the sire effects.

### First method: SA

In the SA method ("segregation analysis", Elston 1980), we consider without simplification the model and the test statistic as they were defined above. The likelihoods under $H_1$ and $H_0$ are calculated using equality (1) and taking account of:

$$Prob(\mathbf{T}_i = t_i \mid s_i) = \prod_{j=1}^{m} Prob(T_{ij} = t_{ij} \mid s_i)$$

Then:

$$M_{1SA} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} \int_{-\infty}^{\infty} k(u_i) \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \cdot k_{t_{ij}}(y_{ij} \mid u_i) du_i$$

with:

$$k(u_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}}\, exp\left(-\frac{1}{2}\left(\frac{u_i}{\sigma_u}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma_u^2}}\, h(u_i)$$

$$k_{t_{ij}}(y_{ij} \mid u_i) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \mu_{t_{ij}} - u_i}{\sigma_e}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, h_{t_{ij}}(y_{ij} \mid u_i)$$

and:

$$M_{0\mathrm{SA}} = \prod_{i=1}^{n} \int_{-\infty}^{\infty} k(u_i) \prod_{j=1}^{m} k_0(y_{ij} \mid u_i)\, du_i$$

with:

$$k_0(y_{ij} \mid u_i) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \mu_0 - u_i}{\sigma_e}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}}\, h_0(y_{ij} \mid u_i)$$

The well known asymptotic properties of the LR test under $H_0$ are the main advantage of this method. If some regularity conditions hold, the test statistic $l$ is asymptotically distributed according to a central $\chi^2$ with $d$ degrees of freedom, $d$ being the number of parameters with fixed value under $H_0$ (Wilks, 1938). However, in the particular context of testing a number of components in a mixture, the regularity conditions are not satisfied since the mixing proportions $p_1$ and $p_2$ have the value zero under $H_0$, which defines the boundary of the parameter space.

Studying mixtures of $m$-normal distributions, Wolfe (1971) suggested that the distribution of the LR test is proportional to a $\chi^2$ distribution with $2d$ degrees of freedom. The proportionality coefficient $c$ should be $c = (n-1-m-1/2g_2)/n$ where $n$ represents the sample size, and $g_2$ the number of components in the mixture under $H_1$. If these results hold in our case, when the number or sires is very large, $l_{\mathrm{SA}}$ should have a $\chi^2$ distribution with 4 degrees of freedom.

The problem with this method is that it requires heavy computation: a complex function of the $Y_{ij}$ must be integrated $n$ times for each estimation of $l_{\mathrm{SA}}$.

## Second and third methods: ME

These methods ("modal estimation" of the sire effect $U_i$), use the equation (2).

Under $H_0$, the likelihood may be written as follows:

$$M_{0\mathrm{ME1}} = \frac{1}{\sqrt{2\pi^m \mid \mathbf{V} \mid}} \prod_{i=1}^{n} h(\hat{u}_i) \prod_{j=1}^{m} h_0(y_{ij} \mid \hat{u}_i) \tag{3}$$

Under $H_1$, the equality (2) leads to

$$M_{1\mathrm{ME1}} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} P_{s_i} \sum_{\boldsymbol{t}_i} Prob(\mathbf{T}_i = \boldsymbol{t}_i \mid s_i)\, .$$

$$\frac{1}{\sqrt{2\pi^m \mid \mathbf{V} \mid}}\, h(\hat{u}_i) \prod_{j=1}^{m} \cdot h_{t_{ij}}(y_{ij} \mid \hat{u}_{i\boldsymbol{t}_i})$$

However, the sums over the vectors $t_i$ for each sire make this computation practically impossible as soon as $m$ is larger than a few units ($3^5 = 243$, $3^{10} = 59\,049$).

Thus, following Elsen *et al.* (1988) we suggest the approximation

$$M_{1\mathrm{ME1}} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} \frac{1}{\sqrt{2\pi^m |\mathbf{V}|}} h(\hat{u}_i) \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \cdot h_{t_{ij}}(y_{ij} \mid \hat{u}_i) \tag{4}$$

Where $\hat{u}_i$ is the distribution mode of $U_i$ conditional on $Y_i$, whatever the genotypes $s_i$ and $t_i$ are. The statistic $l_{\mathrm{ME1}} = -2\log(M_{0\mathrm{ME1}}/M_{1\mathrm{ME1}})$ is no longer an LR test but an approximation lacking the asymptotic properties described above. However we hope that this statistic which requires much less computation will nonetheless retain the power of the first proposed.

An alternative to this second method is to estimate the likelihood $M_{0\mathrm{SA}}$ and $M_{1\mathrm{SA}}$ directly by:

$$M_{0\mathrm{ME2}} = \prod_{i=1}^{n} k(\hat{u}_i) \prod_{j=1}^{m} k_0(y_{ij} \mid \hat{u}_i) \tag{5}$$

$$M_{1\mathrm{ME2}} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} k(\hat{u}_i) \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \cdot k_{t_{ij}}(y_{ij} \mid \hat{u}_i) \tag{6}$$

where $\hat{u}_i$ is defined as above.

As stated by Höschele (1988) this "approximation will be close to $l_{\mathrm{SA}}$ only if the likelihood is very peaked ($m \to \infty$) with most of its probability mass concentrated over a small region about the ML estimates".

### Fourth method: FE

The method (fixed effect of the sires), does not consider the *a priori* information contained in the heritability of the trait. The $u_i$ sire effects are assumed to be fixed, and become supplementary parameters which need to be estimated. The likelihood ratio may be written:

$$I_{\mathrm{FE}} = -2\log\frac{M_{0\mathrm{FE}}}{M_{1\mathrm{FE}}}$$

with:

$$M_{0\mathrm{FE}} = \prod_{i=1}^{n} \prod_{j=1}^{m} k_0(y_{ij} \mid u_i)$$

and:

$$M_{1\mathrm{FE}} = \prod_{i=1}^{n} \sum_{s_i=1}^{3} p_{s_i} \prod_{j=1}^{m} \sum_{t_{ij}=1}^{3} Prob(T_{ij} = t_{ij} \mid s_i) \cdot k_{t_{ij}}(y_{ij} \mid u_i)$$

This method has the advantage of its computational simplicity, while retaining the well known asymptotic properties of the LR test. However, there may be an important loss of power, due to the loss of information on the polygenic variation.

### *The comparisons*

Three problems were studied:

#### Distributions of the statistics under $H_0$

We have just mentioned uncertainties concerning the asymptotic distributions ($\chi^2$ with 4 degrees of freedom for $l_{SA}$ and $l_{FE}$ if Wolfe's (1971) approximation is valid, no known property for $l_{ME}$). Furthermore these distributions are unknown in samples of limited size. In order to estimate these distributions, samples were simulated under $H_0$ (500 samples for SA, 1 000 for FE and ME) with different numbers of sires ($n = 5, 10, 20$) and of progeny per sire ($m = 5, 10, 20$). The test statistics $l_{SA}$, $l_{ME1}$, $l_{ME2}$ and $l_{FE}$ were calculated for each sample. The estimated distributions obtained were used to test the convergences to $\chi^2$ distributions. They also helped determine boundaries for critical regions in samples of a limited size. We used the Harrel and Davis (1982) method to estimate quantiles at 5 and 1% and their jackknife variance as defined by Miller (1974). These simulations were based on a heritability of 0.2.

#### Comparisons of the powers

By using the table of the critical regions thus obtained for each family structure, we have been able to compare the powers of the tests. These powers depend not only on the number and size of the families in the sample but also on the values of the parameters ($\mu$, $\sigma_e$, $p_1$, $p_2$, $q$) which characterize the major gene segregating in the population.

For each of the 9 family structures described above, three $H_1$ hypotheses were considered, each with a simulation of 100 samples. All these populations are assumed to follow the Hardy Weinberg law. The differences between the three $H_1$ hypotheses lie in the mean effects of the genotypes (expressed in standard deviation units) and the frequency of the allele $A$.

Case 1: complete dominance and equal allele frequencies

$$\mu_1 = \mu_2 = 0,\ \mu_3 = 2 \text{ and } q = 0.5.$$

Case 2: additivity, equal allele frequencies

$$\mu_1 = 0,\ \mu_2 = 1,\ \mu_3 = 2,\ \text{and } q = 0.5$$

Case 3: Complete dominance, recessive allele rare

$$\mu_1 = \mu_2 = 0,\ \mu_3 = 2 \text{ and } q = 0.9$$

The power of the tests was measured by the percentage of $H_0$ rejection.

#### Algorithms and cost of calculations

The methods must also be compared on the basis of how much computation they require. The calculations described above were made using the quadrature and

optimization subroutines of the NAG fortran library. In order to maximize the likelihoods of the sample we used a Quasi-Newton algorithm in which the derivatives are estimated by finite differences.

The same algorithm was used for the four methods, giving results of a similar degree of precision. However, various algorithms can be used to estimate the maximum likelihood of the parameters. In the ME and FE tests, the first derivatives have a simple algebraic form and the maximum likelihood solutions are reached by zeroing the first derivatives (with respect to each of the parameters) of the logarithm of the likelihood. Under $H_1$ the corresponding system of equations can be solved iteratively, but not directly, by using for instance the EM algorithm defined by Dempster *et al.* (1977): see appendix.

This is the algorithm we used for the ME2 test in order to obtain more extensive information on critical region: 5, 10, 20, and 40 sires, 5, 10, 20 and 40 progenies/sire, heritability of 0, 0.2, 0.4.

## RESULTS AND DISCUSSION

### *Comparison of the four methods*

Tables I to IV show the main characteristics of the distributions of the 4 test statistics: mean, standard deviation, 5% and 1% empirical quantiles and percentage of replicates beyond the 5% and 1% quantiles of a $\chi^2_4$. Table V shows their powers.

First, we can note that for the number of progeny increases, the mean distributions as the four test statistics decrease (except $l_{SA}$ between $m = 5$ and $m = 10$ for $n = 5$).

The fact that $l$ statistics distributions converge toward a $\chi^2$ with 4 degrees of freedom cannot be confirmed since all the distributions of $l$, but one (segregation analysis with 5 sires and 5 progenies/sire), are significantly different from a $\chi^2$ using a $\chi^2$ test of fit. Moreover, the scaled statistics $(2E(l)/\text{var}\ (l))$. $l$ are also significantly different from a $\chi^2$. It must be emphasized that the samples studied are far from the conditions of validity of Wolfe's approximation which requires that $n > 10.m$ (Everitt, 1981). The $l_{SA}$ statistics show a notable stability as the family size varies, whereas for $l_{FE}$ the statistics only reaches an asymptote as $m$, the number of progeny per sire increases. As regards the $l_{ME}$ statistics, the results are totally different.

The mean and standard deviation of the $l_{ME1}$ statistic decreases when the number of sires or progeny per sire increases. It appeared that the distribution of this $l_{ME1}$ statistic becomes very peaked near zero. It must be noticed that this pattern is close to the asymptotic distribution of the LR test of a mixture of 2 known distributions in unknown proportion studied by Titterington *et al.* (1985). These authors found that, under $H_0$ (only one component) the LR test "is 0 with a probability 0.5 and, with the same probability, is distributed as a $\chi^2$ with one degree of freedom". On the other hand, for a given number of progeny, the mean of the $l_{ME2}$ distribution increases with the number of sires. The fewer the progeny, the greater the increase.

The calculation of the power (Table V) shows some important facts: very low power of the four statistics for low number of sires and/or progeny, clear superiority of the segregation analysis and first of the modal estimation method whatever

**Table I.** Results of the simulations under $H_0$ for the $l_{SA}$ statistic: means ($\mu$), standard deviations ($\sigma$), 5% ($s_5$) and 1% ($s_1$) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% ($r_5$) and 1% ($r_1$) quantiles of a $\chi^2_4$.

| *Number of* *sires (n)* | *progeny (m)* | $\mu$ | $\sigma$ | $s_5$ | $s_1$ | $r_5$ | $r_1$ |
|---|---|---|---|---|---|---|---|
| | 5 | 4.42 | 2.92 | 9.74 (0.44) | 14.91 (1.87) | 5.71 | 0.98 |
| 5 | 10 | 4.48 | 3.00 | 9.99 (0.33) | 14.30 (0.69) | 6.04 | 1.59 |
| | 20 | 4.44 | 3.17 | 10.64 (0.45) | 14.36 (0.49) | 7.59 | 1.86 |
| | 5 | 4.71 | 3.16 | 10.86 (0.19) | 14.69 (0.27) | 8.36 | 1.91 |
| 10 | 10 | 4.47 | 3.20 | 10.50 (0.32) | 14.26 (0.61) | 7.62 | 1.66 |
| | 20 | 4.36 | 3.15 | 10.50 (0.46) | 14.31 (1.16) | 7.39 | 1.14 |
| | 5 | 4.74 | 3.36 | 11.10 (0.27) | 15.51 (0.74) | 8.87 | 1.94 |
| 20 | 10 | 4.42 | 3.25 | 10.75 (0.55) | 15.15 (0.83) | 7.38 | 1.75 |
| | 20 | 4.14 | 3.45 | 11.25 (0.51) | 15.65 (1.21) | 7.89 | 2.17 |

**Table II.** Results of the simulations under $H_0$ for the $l_{ME1}$ statistic: means ($\mu$), standard deviations ($\sigma$), 5% ($s_5$) and 1% ($s_1$) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% ($r_5$) and 1% ($r_1$) quantiles of a $\chi^2_4$.

| *Number of* *sires (n)* | *progeny (m)* | $\mu$ | $\sigma$ | $s_5$ | $s_1$ | $r_5$ | $r_1$ |
|---|---|---|---|---|---|---|---|
| | 5 | 4.61 | 3.52 | 11.01 (0.39) | 15.96 (0.83) | 8.8 | 2.4 |
| 5 | 10 | 3.65 | 3.09 | 9.51 (0.29) | 14.54 (0.74) | 4.8 | 1.4 |
| | 20 | 2.92 | 2.95 | 8.47 (0.39) | 12.86 (0.74) | 3.5 | 0.8 |
| | 5 | 3.83 | 3.20 | 10.31 (0.20) | 15.69 (1.20) | 7.6 | 1.5 |
| 10 | 10 | 2.77 | 2.94 | 8.71 (0.25) | 12.59 (0.58) | 3.6 | 0.7 |
| | 20 | 2.10 | 2.63 | 7.36 (0.32) | 12.80 (1.39) | 1.9 | 1.1 |
| | 5 | 2.75 | 3.04 | 8.55 (0.31) | 14.14 (1.09) | 3.2 | 1.2 |
| 20 | 10 | 1.81 | 2.52 | 6.97 (0.29) | 11.14 (0.80) | 1.7 | 0.4 |
| | 20 | 1.27 | 2.11 | 5.82 (0.34) | 9.46 (0.46) | 0.9 | 0.0 |

**Table III.** Results of the simulations under $H_0$ for the $l_{ME2}$ statistic: means ($\mu$), standard deviations ($\sigma$), 5% ($s_5$) and 1% ($s_1$) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% ($r_5$) and 1% ($r_1$) quantiles of a $\chi^2_4$.

| *Number of* *sires (n)* | *progeny (m)* | $\mu$ | $\sigma$ | $s_5$ | $s_1$ | $r_5$ | $r_1$ |
|---|---|---|---|---|---|---|---|
| | 5 | 12.28 | 4.41 | 20.27 (0.13) | 25.03 (0.32) | 71.4 | 35.5 |
| 5 | 10 | 9.71 | 4.09 | 16.90 (0.20) | 21.69 (0.59) | 48.5 | 17.5 |
| | 20 | 7.60 | 4.14 | 15.61 (0.30) | 19.79 (0.36) | 27.9 | 10.1 |
| | 5 | 17.28 | 5.27 | 26.81 (0.32) | 32.90 (0.54) | 95.4 | 77.2 |
| 10 | 10 | 13.52 | 5.11 | 22.94 (0.25) | 27.36 (0.51) | 78.0 | 48.0 |
| | 20 | 9.36 | 4.85 | 18.54 (0.31) | 23.51 (0.49) | 43.8 | 19.0 |
| | 5 | 26.47 | 6.66 | 38.24 (0.41) | 44.63 (0.78) | 99.9 | 99.1 |
| 20 | 10 | 19.56 | 6.49 | 31.15 (0.36) | 37.21 (0.86) | 96.7 | 82.9 |
| | 20 | 12.17 | 5.94 | 23.02 (0.41) | 29.30 (0.58) | 63.4 | 36.8 |

**Table IV.** Results of the simulations under $H_0$ for the $l_{FE}$ statistic: means ($\mu$), standard deviations ($\sigma$), 5% ($s_5$) and 1% ($s_1$) empirical quantiles (their standard deviations between brackets), and percentages of replicates beyond the 5% ($r_5$) and 1% ($r_1$) quantiles of a $\chi^2_4$.

| *Number of* *sires (n)* | *progeny (m)* | $\mu$ | $\sigma$ | $s_5$ | $s_1$ | $r_5$ | $r_1$ |
|---|---|---|---|---|---|---|---|
| | 5 | 9.62 | 5.13 | 18.93 (0.19) | 24.42 (0.47) | 46.0 | 21.3 |
| 5 | 10 | 6.26 | 4.28 | 14.30 (0.23) | 19.13 (0.54) | 20.5 | 6.87 |
| | 20 | 4.64 | 3.86 | 12.33 (0.28) | 16.72 (0.53) | 11.3 | 3.85 |
| | 5 | 12.27 | 6.68 | 24.03 (0.30) | 30.89 (1.27) | 63.2 | 39.6 |
| 10 | 10 | 7.19 | 5.40 | 17.31 (0.31) | 22.41 (0.59) | 30.3 | 13.9 |
| | 20 | 4.22 | 3.99 | 12.17 (0.30) | 17.68 (0.89) | 10.5 | 3.50 |
| | 5 | 16.20 | 9.61 | 32.79 (0.47) | 42.90 (1.22) | 73.2 | 59.8 |
| 20 | 10 | 7.86 | 6.82 | 20.72 (0.25) | 28.55 (1.00) | 34.8 | 20.6 |
| | 20 | 3.69 | 3.84 | 11.35 (0.36) | 16.29 (0.58) | 9.15 | 2.81 |

these numbers, with respectively a 90% and a 80% power in the best case (though involving only 400 animals), very poor performance of the $l_{FE}$ statistic, intermediate power for $l_{ME2}$.

Thus knowledge of heritability is a substantial advantage and gives a reason to prefer the $l_{ME}$ statistics against the $l_{FE}$, which requires similar amounts of computation.

**Table V.** Results of the simulations under $H_1$ powers of the 4 tests for a 5% first type error (percentage of $H_0$ rejection) and their 5% confidence intervals between brackets. Comparisons of different family structures and parameters values.

| *Number of sires* (*n*) | *progeny* (*m*) | *CASE 1* SA | ME1 | ME2 | FE | *CASE 2* SA | ME1 | ME2 | FE | *CASE 3* SA | ME1 | ME2 | FE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | 9 (6-14) | 5 (2-11) | 7 (4-11) | 2 (1-5) | 3 (1-8) | 6 (3-12) | 3 (1-8) | 3 (1-9) | 9 (4-18) | 5 (2-11) | 7 (3-15) | 4 (1-12) |
| 5 | 10 | 17 (12-24) | 14 (9-22) | 9 (5-16) | 8 (4-15) | 11 (6-19) | 5 (2-11) | 8 (4-15) | 3 (1-9) | 7 (3-15) | 5 (2-11) | 6 (3-13) | 6 (2-14) |
| | 20 | 24 (17-39) | 30 (22-40) | 18 (12-27) | 16 (10-25) | 9 (5-16) | 9 (5-16) | 8 (4-15) | 1 (0-6) | 19 (12-29) | 21 (14-30) | 14 (8-23) | 7 (3-15) |
| | 5 | 17 (12-24) | 14 (9-22) | 12 (7-20) | 7 (3-14) | 7 (3-14) | 6 (3-12) | 7 (3-14) | 2 (1-8) | 5 (2-12) | 8 (4-15) | 3 (1-9) | 2 (0-9) |
| 10 | 10 | 27 (21-23) | 41 (32-51) | 16 (11-21) | 9 (5-14) | 8 (4-16) | 11 (6-19) | 3 (1-9) | 1 (0-6) | 19 (12-28) | 18 (12-27) | 7 (3-14) | 1 (0-7) |
| | 20 | 54 (45-63) | 60 (50-69) | 38 (29-48) | 24 (17-33) | 15 (10-22) | 16 (10-24) | 5 (2-10) | 1 (0-5) | 34 (25-43) | 30 (22-40) | 16 (10-25) | 11 (6-20) |
| | 5 | 26 (19-34) | 27 (19-36) | 18 (12-26) | 7 (3-13) | 9 (5-16) | 11 (6-19) | 6 (3-13) | 2 (1-6) | 15 (10-22) | 18 (12-27) | 4 (2-9) | 1 (0-6) |
| 20 | 10 | 51 (42-61) | 48 (38-58) | 27 (19-36) | 7 (3-13) | 21 (14-30) | 18 (12-27) | 7 (4-14) | 7 (3-14) | 33 (25-42) | 37 (28-47) | 13 (8-20) | 8 (4-16) |
| | 20 | 90 (83-94) | 80 (71-87) | 72 (62-80) | 56 (46-65) | 25 (19-33) | 21 (14-30) | 15 (9-24) | 9 (5-16) | 48 (38-57) | 62 (52-71) | 34 (26-43) | 31 (23-41) |

The comparison of powers in hypothesis $H_1$ is also interesting: it is much more difficult to detect an additive major gene (case 2) than a dominant one (case 1) even with the segregation analysis which is 3 to 4 times less powerful in case 2 than in case 1. In comparison with the isofrequent case, the third case shows a 50% loss of power: with measurements made on a small population, very few individuals if any, belong to the high mean distribution.

The computation requirements have been estimated, on a 3083 IBM computer, by the CPU time needed for the evaluation of the statistics under $H_0$. Ten replicates of a sample of 10 sires and 10 progenies per sire used 640 s for the $l_{SA}$ statistic, 142 s for the $l_{FE}$ statistic and 48 s for the $l_{ME}$ statistics. Using the EM algorithm instead of the direct maximization of $l_{ME}$ with the NAG subroutines decreases the

time requirements to 20 s only. Thus, the proposed simplified tests $l_{ME}$ are 30 times as fast as the segregation analysis.

### *Tables of quantiles*

Although theoretical works are still needed in order to describe the asymptotic behaviour of the $l_{SA}$, $l_{ME1}$ and $l_{FE}$ tests, one can use, as a first approach, the quantiles given in our tables for larger populations since this will produce an overestimation of the first type error. On the contrary, some more calculations are needed for the $l_{ME2}$ test.

The 5 and 1% points for this statistic are given in figures 1 to 3 depending on the heritability (0.0, 0.2, 0.4). Each figure gives these points for varying numbers of sires and progeny per sire.

Note that when the heritability is 0., the sire effect is not defined and, thus, that the $u_i[a+1]$ terms disappear from the equations given in the appendix.

The results of Table III are confirmed: the quantile estimates increase with the number of sires $n$ (for a given number of progeny per sire, $m$) and decrease when the number of progeny per sire increases. Two other results must be noticed:
- given $n$ and $m$, the lower the heritability, the greater the quantiles.

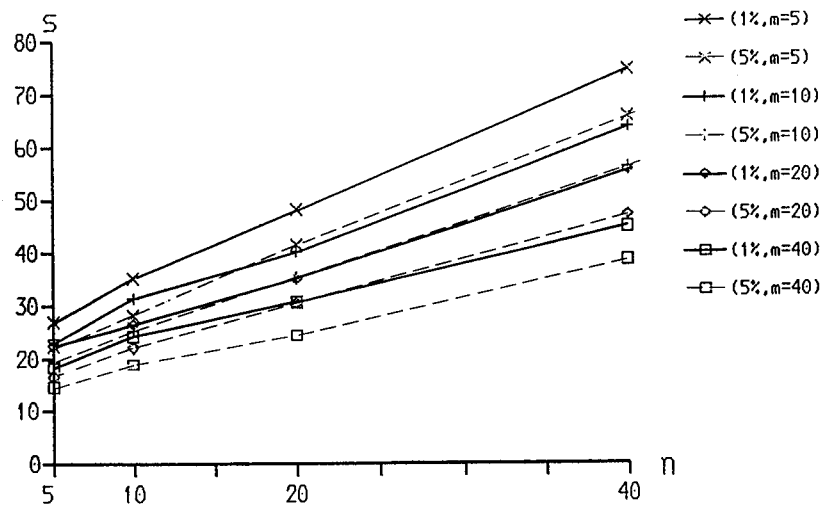


**Fig. 1.** 5% and 1% quantiles of the $l_{ME2}$ test statistic for varying family structures ($h^2 = 0$).

- on the variation range studied for $m$, the number of progeny per sire, the increase of the quantiles is nearly linear with $n$ (number of sires) allowing some extrapolations for higher values of this number.

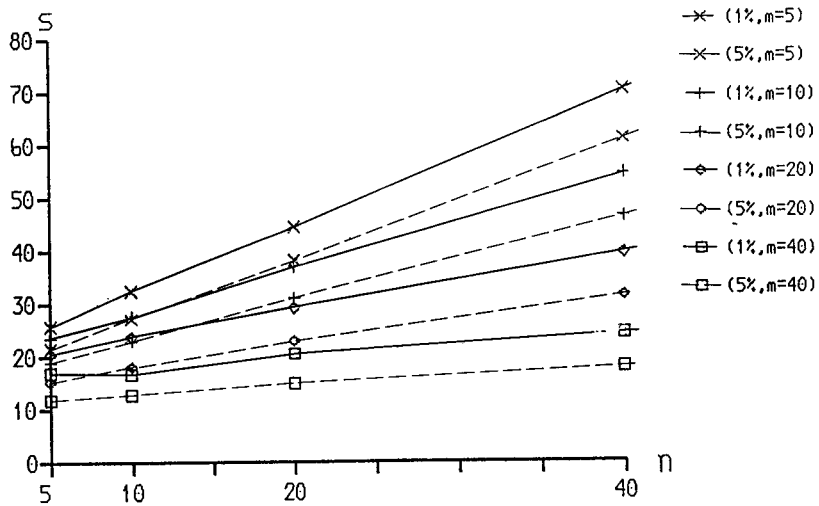


**Fig. 2.** 5% and 1% quantiles of the $l_{ME2}$ test statistic for varying family structures ($h^2 = 0.2$).

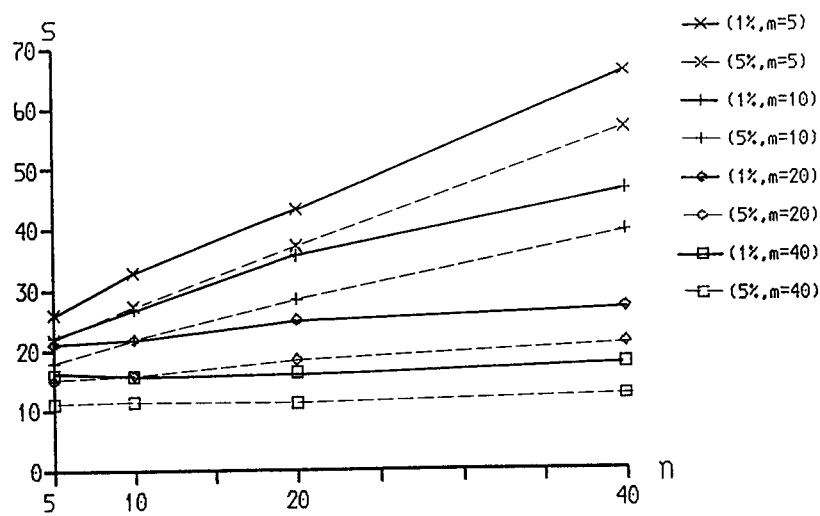


**Fig. 3.** 5% and 1% quantiles of the $l_{ME2}$ test statistic for varying family structures ($h^2 = 0.4$).

Finally, the jackknife standard deviation of the estimated quantile varies, for the 5% case, between 0.23 and 0.89, with a mean value of 0.52 and, for the 1% case, between 0.39 and 1.65 with a mean value of 0.92. These errors could explain the observed deviations of the plotted curves from smoothness.

## CONCLUSIONS

On the four statistical tests studied, the "segregation analysis" method is, as expected, the most powerful. Applied on a large scale, this test requires a great deal for computation. The "modal effect" method requires much less computation than the segregation analysis and shows practically no loss of power for the first version and a limited loss of power (diminishing as soon as the sample size is sufficient) for the second version. Unfortunately, the asymptotic distribution of this last statistic is unknown. The tables of quantiles we obtained by simulation permit the utilization of this test for typical sample sizes and for various heritability values.

## REFERENCES

Dempster A.P., Laird N.M. & Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc., Series B* 39, 1-38

Elsen J.M., Vu Tien Khang J. & Le Roy P. (1988) A statistical model for genotype determination at a major locus in a progeny test design. *Genet. Sel. Evol.* 20, 211-226

Elston R.C. (1980) Segregation analysis. *In: Current developments in anthropological genetics* (Mielke J.H. & Crawford M.H. eds), 1, Plenum Publishing Corporation, New York, 327-354

Elston R.C. & Stewart J. (1971) A general model for the genetic analysis of pedigree data. *Hum. Hered.* 21, 523-542

Everitt B.S. (1981) A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivar. Behav. Res.* 16, 171-180

Hanset R. (1982) Major genes in animal production, examples and perspectives: cattle and pigs. *2nd world congress on genetics applied to livestock production, Madrid, 4-8 oct., 1982*, 5, Editorial Garsi, Madrid, 439-453

Harrel F.E. & Davis C.E. (1982) A new distribution-free quantile estimator. *Biometrika* 69, 635-640

Höschele I. (1988) Statistical techniques for detection of major genes in animal breeding data. *Theor. Appl. Genet.* 76, 311-319

Ménissier F. (1982) Present state of knowledge about the genetic determination of muscular hypertrophy or the double-muscled trait in cattle. *In: Muscle hypertrophy of genetic origin and its use to improve beef production* (King J.W.B. & Ménissier F. eds), Martinus Nijhof, The Hague, 387-428

Mérat P. & Ricard F.H. (1974) Etude d'un gène de nanisme lié au sexe chez la poule: importance de l'état d'engraissement et gain de poids chez l'adulte. *Ann. Génét. Sél. Anim.* 6, 211-217

Miller R.G. (1974) The Jackknife. A review, *Biometrika* 61, 1-15

Ollivier L. (1980) Le déterminisme génétique de l'hypertrophie musculaire chez le porc. *Ann. Génét. Sél. Anim.* 12, 383-394

Piper L.R. & Bindon B.M. (1982) The *Booroola Merino* and the performance of medium *non-peppin* crosses at Armidale. *In: The* Booroola Marino, (Piper L.R., Bindon B.M. & Nethery R.D. eds), CSIRO, Melbourne, 9-20

Titterington D.M., Smith A.F.M. & Makow U.E. (1985) *Statistical analysis of finite mixture distributions.* Wiley, New York

Wilks S.S. (1938) The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* 9, 60-62

Wolfe J.H. (1971) A Monte Carlo study of the sampling distribution of the likelihood ratio for mixture of multinormal distributions. *Tech. Bull.*, STB 72-2, Naval Personnel and Training Research Laboratory, San Diego

## APPENDIX

### *Application of the EM algorithm to the estimation of the test statistic $l_{ME}$ under $H_1$*

The EM algorithm is an iterative procedure. Each of its iterations consists of two steps E (Expectation) and M (Maximization). In our calculations we have considered that convergence is obtained when, a being the iteration number, the following inequality is satisfied:

$$|l_{ME}[a+1] - l_{ME}[a]| < 10^{-6}|l_{ME}[a]|$$

**Step E of the ath iteration consists of estimating posterior probabilities of the observations**

$$q_i(s_i)[a+1] = Prob(S_i = s_i \mid \mathbf{Y}_i,\ u_i[a])$$
$$q_{ij}(t_{ij} \mid s_i)[a+1] = Prob(T_{ij} = t_{ij} \mid S_i = s_i,\ \mathbf{Y}_i,\ u_i[a])$$
$$q_{ij}(t_{ij})[a+1] = Prob(T_{ij} = t_{ij} \mid \mathbf{Y}_i,\ u_i[a])$$

These probabilities are estimated using the $a$th iteration values of $\sigma_e[a]$, $q[a]$, $u_i[a]$ $(i = 1, ..., n)$, $\mu_t[a]$ $(t = 1, 2, 3)$ and $p_s[a]$ $(s = 1, 2, 3)$. The following quantities are calculated successively:

$$k_{t_{ij}}(y_{ij} \mid u_i)[a+1] = \frac{1}{\sqrt{2\pi}\ \sigma_e[a]}\ exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \mu_{t_{ij}}[a] - u_i[a]}{\sigma_2[a]}\right)^2\right)$$

$$q_{ij}(t_{ij} \mid s_i)[a+1] = \frac{Prob(T_{ij} = t_{ij} \mid s_i)k_{t_{ij}}(y_{ij} \mid u_i[a+1])}{\Sigma_{t'_{ij}} Prob(T_{ij} = t'_{ij} \mid s_i)k_{t'_{ij}}(y_{ij} \mid u_i[a+1])}$$

$$q_i(s_i)[a+1] = \frac{p_{s_i}[a]\Pi_j(\Sigma_{t_{ij}} Prob(T_{ij} = t_{ij} \mid s_i) k_{t_{ij}}(y_{ij} \mid u_i[a+1]))}{\Sigma_{s_i'} p_{s_i'}[a]\Pi_j(\Sigma_{t_{ij}} Prob(T_{ij} = t_{ij} \mid s_i') k_{t_{ij}}(y_{ij} \mid u_i[a+1]))}$$

$$q_{ij}(t_{ij}[a+1] = \sum_{s_i} q_i(s_i)[a+1] \,.\, q_{ij}(t_{ij} \mid s_i)[a+1]$$

$l_{\mathrm{ME1}}[a+1]$ is calculated as in (3) and (4), and $l_{\mathrm{ME2}}[a+1]$ is calculated as in (5) and (6).

***Step M of the ath iteration***

Given the previous posterior probabilities, the distribution parameters are obtained by annulling the derivatives of $l_{\mathrm{ME}}[a+1]$ with respect to these parameters. We then get:

for $t = 1, 2, 3$

$$\mu_t[a+1] = \frac{\Sigma_i \, \Sigma_j \, q_{ij}(t)[a+1] \,.\, (y_{ij} - u_i[a])}{\Sigma_i \, \Sigma_j \, q_{ij}(t)[a+1]}$$

for $i = 1, \ldots, n$

$$u_i[a+1] = \frac{\Sigma_j \, \Sigma_t \, q_{ij}(t)[a+1] \,.\, (y_{ij} - \mu_t[a+1])}{\sigma_e^2 \, \sigma_u^{-2} + \Sigma_j \, \Sigma_t \, q_{ij}(t)[a+1]}$$

$$\sigma_e^2[a+1] = \frac{\sigma_e^2 \, \sigma_u^{-2} \, \Sigma_i \, u_i^2[a+1] + \Sigma_i \Sigma_j \Sigma_t \, q_{ij}(t)[a+1] \,.\, (y_{ij} - u_i[a+1] - \mu_t[a+1])^2}{nm}$$

the denominator being $n(m+1)$ for the $l_{\mathrm{ME2}}$ test.

$$p_{s_i}[a+1] = \frac{\Sigma_i q_i(s_i)[a+1]}{n}$$

$$q[a+1] = \frac{\Sigma_i \Sigma_j q_{ij}(1)[a+1] + \Sigma_i (q_i(3)[a+1] \,.\, \Sigma_j q_{ij}(2|3)[a+1])}{nm - \Sigma_i (q_i(2)[a+1] \,.\, \Sigma_j q_{ij}(2|2)[a+1])}$$