Computing approximate monogenic model likelihoods in large pedigrees with loops

Llg Janss, Jam Van Arendonk, Jhj Van Der Werf

Genet Sel Evol (1995) 27, 567-579

Original article

Computing approximate monogenic model likelihoods in large pedigrees with loops

LLG Janss, JAM Van Arendonk, JHJ Van der Werf

Wageningen Agricultural University, Department of Animal Breeding, PO Box 338, 6700 AH Wageningen, The Netherlands

(Received 30 January 1995; accepted 11 September 1995)

Summary – In this study 'iterative peeling' is introduced, a method equivalent to the traditional recursive peeling method for computing exact likelihoods in nonlooped pedigrees, but which can also be used to obtain approximate likelihoods in looped pedigrees. Iterative peeling is an interesting tool for animal breeding, where exact recursive peeling is generally unfeasible due to the abundant number of loops in animal pedigrees. In simulations, hypothesis testing and parameter estimation were compared based on approximate likelihoods in looped pedigrees and exact likelihoods in nonlooped pedigrees, showing no biases introduced by the approximation in looped pedigrees.

likelihood / pedigree peeling / major gene / looped pedigree

Résumé – Calcul approximatif de vraisemblance pour un modèle monogénique dans de grands pedigrees à boucles. *Dans cette étude on introduit une procédure itérative de condensation de l'information contenue dans un pedigree, appelée « épluchage », qui est équivalente à l'épluchage récursif pour le calcul des vraisemblances exactes dans des pedigrees sans boucles, mais qui est également utilisable pour le calcul de vraisemblances approximatives dans les pedigrees à boucles. L'épluchage itératif est une méthode intéressante en génétique animale où la méthode récursive exacte est généralement inapplicable à cause du grand nombre de boucles dans les pedigrees animaux. À l'aide de simulations, on a comparé des tests d'hypothèse et l'estimation de paramètres basés sur des vraisemblances approximatives dans des pedigrees à boucles et des vraisemblances exactes dans des pedigrees sans boucles, montrant qu'il n'y a pas de biais introduit par le calcul approximatif dans des pedigrees à boucles.*

vraisemblance / condensation d'information de pedigree / gène majeur / pedigree à boucles

INTRODUCTION

Research into the use of major gene models in animal breeding has been aimed mainly at approximations to a mixed inheritance model, including polygenes, in one generation half-sib structures (Hoeschele, 1988; Le Roy *et al*, 1989; Knott *et al*, 1992). Because of the pedigree loops that arise in animal breeding situations, extension to multigeneration pedigrees is difficult. A pedigree loop arises when 2 individuals are connected by more than one path of descent or marriage relationships. Lange and Elston (1975) described various types of loops, among which are inbreeding loops, marriage rings and marriage loops. In animal breeding pedigrees these kinds of loops are very common. In particular, multiple matings which are generally applied to males and often to females, result in many marriage loops and marriage rings.

For genotype probability and likelihood computation, loops can only be dealt with in an exact manner in pedigrees with a few simple non-overlapping loops using the traditional recursive peeling method (Elston and Stewart, 1971; Cannings *et al*, 1976; Cannings *et al*, 1978). However, in highly looped pedigrees, common in animal breeding, exact recursive peeling is too demanding computationally and recursive peeling is not flexible enough to allow for approximate computations.

In this study we introduce 'iterative peeling'. Iterative peeling is developed as an exact method for application in nonlooped pedigrees, equivalent to recursive peeling, but which, unlike the original recursive variant, can be used without modifications in looped pedigrees to obtain approximate likelihoods. The main objective of this paper is to introduce iterative peeling for such approximations in looped pedigrees, allowing for a more general application of major gene models in animal breeding. Using simulations, the usefulness of the approximation for likelihood-based hypothesis testing and parameter estimation in looped pedigrees is investigated. A monogenic model will be considered, which can be extended to a mixed inheritance model, as will be discussed.

RECURSIVE AND ITERATIVE PEELING

In the first section, recursive peeling is described for obtaining monogenic model likelihoods in nonlooped pedigrees. In the second section, 'iterative peeling' is introduced as an equivalent method for exact computations in nonlooped pedigrees. The equivalent exact method in nonlooped pedigrees can be used as an approximate method in looped pedigrees.

Recursive peeling

Probability and likelihood computations in nonlooped pedigrees can be done by recursive peeling (Elston and Stewart, 1971; Cannings *et al*, 1976; Cannings *et al*, 1978) using 2 basic peeling operations of 'peeling up' and 'peeling down'. Roughly, considering a single family, a peel-up operation represents the information in a family in probabilities for the genotype G_i of a parent i, and a peel-down operation represents this information in probabilities for the genotype G_k for an offspring k. Here, a notation based on Van Arendonk *et al* (1989) is used, where the result

of the peel-up operation is denoted by $prog(G_i)$ and the result of the peel-down operation is denoted by $prior(G_k)$. The corresponding notation in Cannings *et al* (1976, 1978) is the $R^*(..;G_i)$ function for peeling up and the $R^+(..;G_k)$ function for peeling down.

Peeling operations are used recursively, *eg*, the computation of a *prog* term for a parent based on progeny data may include previously computed *prog* terms of those progeny, representing information from grand-progeny. The aim of peeling is to condense all information from a pedigree into a *prior* and *prog* term for a single individual l, obtaining the likelihood L for all data in the pedigree as:

$$L = \Sigma_{Gl}\, prior(G_l)\, f(y_l|G_l)\, prog(G_l) \qquad [1]$$

where $f(y_l|G_l)$ is the penetrance function, which is the probability for the observed data y_l on individual l, given that it has genotype G_l. The individual l may be an individual from the base population, in which case the base-population genotype frequency $P(G_l)$ is used in place of $prior(G_l)$. Individual l may also have no own data or no progeny, in which case the corresponding penetrance term or *prog* term is removed. Computationally this is implemented using a penetrance or *prog* term containing 1's.

Peeling equations

A peeling equation for an individual is obtained by considering the collection of possible base-population genotype frequencies, genotype transmission probabilities, penetrance probabilities and other peeling terms pertaining to the individuals in its family and summing over all possible genotypes of the family members. The terms thus entering a peeling equation are difficult to give in general. Here, equations will be given to use peeling in a pedigree structure with dams nested within sires. In this structure a family is a half-sib family of one sire with several mates, containing groups of full sibs which are, across groups, paternal half-subs. Three different peeling equations are considered: 2 for peeling up, dependent on whether this is done for a sire or a dam, and 1 for peeling down. In the peeling equations, *prior*, *prog* and penetrance functions on family members are specified in all places where they can enter. When these are not relevant, *eg*, when a progeny does not have progeny of its own, these are removed or, computationally, terms containing 1's are used. *Prior* terms for individuals in the base populations are substituted with base-population genotype frequencies.

To condense all information in a *prog* term for a sire i the following expression is used:

$$prog(G_i) = \Pi_j \Sigma_{Gj}\, prior(G_j)\, f(y_j|G_j) \Pi_k \Sigma_{Gk}\, P(G_k|G_i, G_j)\, f(y_k|G_k)\, prog(G_k) \qquad [2]$$

where $j = 1$ to n_i are mates of i, each mate having $k = 1$ to n_{ij} progeny, and $P(G_k|G_i, G_j)$ is the genotype transmission probability of sire i and a dam j to offspring k. To condense all the information from a half-sib family into a *prog* term for 1 particular dam j^* of the family, the following expression is used:

$$prog(G_{j^*}) = \Sigma_{Gi}\, prior(G_i)\, f(y_i|G_i)\, prog_{-j^*}(G_i) \Pi_k \Sigma_{Gk}\, P(G_k|G_i, G_{j^*})\, f(y_k|G_k)\, prog(G_k) \qquad [3]$$

where i is the sire of the family, $prog_{-j^*}(G_i)$ is like in equation [2], but excluding dam j^* and $k = 1$, n_{ij^*} are progeny of dam j^*. To condense all the information in a *prior* term for 1 particular progeny k^* with dam j^*, the following expression is used:

$$prior(G_{k^*}) = \Sigma_{Gi}\, prior(G_i)\, f(y_i|G_i)\, phs(G_i)$$
$$\Sigma_{Gj^*}\, prior(G_{j^*})\, f(y_{j^*}|G_{j^*})\, fs(G_i, G_{j^*})\, P(G_{k^*}|G_i, G_{j^*}) \qquad [4]$$

where i is the sire of the family, $phs(G_i)$ is a term that includes information on the paternal half-sibs of k^*, which is a function of the genotype of its sire i and is computed as:

$$phs(G_i) = \Pi_{j,j\neq j^*}\Sigma_{Gj}\, prior(G_j)\, f(y_j|G_j)\Pi_k\Sigma_{Gk}P(G_k|G_i, G_j)\, f(y_k|G_k)\, prog(G_k)$$

In [4] $fs(G_i, G_{j^*})$ is a term that includes information on the full-subs of k^*, which is a function of the genotypes of its sire i and dam j^*, and is computed as:

$$fs(G_i, G_{j^*}) = \Pi_{k,k\neq k^*}\Sigma_{Gk}\, P(G_k|G_{i^*}, G_{j^*})\, f(y_k|G_k) prog(G_k)$$

Iterative peeling

Iterative peeling is equivalent to recursive peeling used in nonlooped pedigrees. Iterative peeling is based on algebraic partitioning of the likelihood and on repeated computation of peeling equations, based on the idea of iterative computation of genotype probabilities (Van Arendonk *et al*, 1989).

Partitioning of likelihood

The aim of obtaining the likelihood of all data using equation [1] requires families to be handled in a certain order and requires peeling, within each family, to be in a certain direction. Peeling operations can be used to partition the likelihood pertaining to parts of the pedigree. This partitioning is continued until parts are obtained pertaining to single families. This allows a family-wise evaluation of the likelihood, and the requirement of peeling to have a direction within each family becomes obsolete.

Consider the pedigree with 5 individuals in figure 1. In this pedigree 2 families are present, one family with individuals 1, 2 and 3, and a second with individuals 3, 4 and 5. Here, one partitioning above and below individual 3 divides the pedigree in 2 families, with individual 3 being in both families. Individual 3 is called a linking individual. The likelihood for a monogenic model, assuming data is available on all 5 individuals, is computed as:

$$L = \Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\Sigma_{G4}\Sigma_{G5}\, P(G_1)P(G_2)\, P(G_3|G_1, G_2)\, P(G_4)\, P(G_5|G_3, G_4)$$
$$f(y_1|G_1)\, f(y_2|G_2)\, f(y_3|G_3)\, f(y_4|G_4)\, f(y_5|G_5)$$

Now, L is multiplied and divided by $L_1 = \Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\, P(G_1)\, P(G_2)\, P(G_3|G_1, G_2)$ $f(y_1|G_1)\, f(y_2|G_2)$, which is the likelihood of family 1, ignoring data on progeny 3. Some reordering yields:

$$L = L_1 * \Sigma_{G3}\Sigma_{G4}\Sigma_{G5}\{\Sigma_{G1}\Sigma_{G2}\, P(G_1)\, P(G_2)\, P(G_3|G_1, G_2)\, f(y_1|G_1)\, f(y_2|G_2)/L_1\}$$
$$*P(G_4)\, P(G_5|G_3, G_4)\, f(y_3|G_3)\, f(y_4|G_4)\, f(y_5|G_5)$$

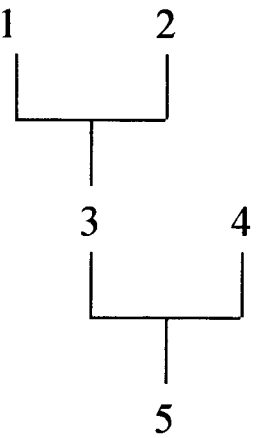

Fig 1. Example pedigree to demonstrate partitioned computation of the likelihood.

where the part $\Sigma_{G1}\Sigma_{G2}\, P(G_1)\, P(G_2)\, P(G_3|G_1, G_2)\, f(y_1|G_1)\, f(y_2|G_2)$ has been isolated. This part is $prior(G_3)$. The term defined as L_1 can be rewritten as $\Sigma_{G3}\, \Sigma_{G1}\Sigma_{G2}\, P(G_1)\, P(G_2)\, P(G_3|G_1, G_2)\, f(y_1|G_1)\, f(y_2|G_2)$, which is $\Sigma_{G3}\, prior(G_3)$. This simplifies L to:

$$L = L_1\{\Sigma_{G3}\Sigma_{G4}\Sigma_{G5}\, prior^{sc}(G_3)\, P(G_4)\, P(G_5|G_3, G_4)\, f(y_3|G_3) f(y_4|G_4)\, f(y_5|G_5)\}$$

where $prior^{sc}(G_3)$ stands for a scaled, or normalised, prior term. Now the likelihood can be written as $L = L_1 L_2$, or $\ln(L) = \ln(L_1) + \ln(L_2)$, with one likelihood term per family. This is a partitioning using a *prior* term for the linking individual. It shows that for this type partitioning (i) in the family where the linking individual is a progeny, after the partitioning, information on the linking individual, *ie* own data and progeny data, is ignored; and (ii) in the family where the linking individual is a parent, a scaled *prior* term is used for the linking individual. This term is used in a manner like a base-population genotype frequency for base individuals. The scaled *prior* term for a linking individual l, is computed in general as:

$$prior^{sc}(G_l) = prior(G_l)/\Sigma_{Gl}\, prior(G_l)$$

Although the partioning is shown only for 1 example, the partitioning is very general. The term L_1 above is in general the sum of the *prior* term for a linking individual l, which is the collection of all probability terms pertaining to anterior individuals of l and the transmission probability to l, summed over all possible genotypes of l and of its anterior individuals. At the same time this term represents the likelihood of the entire anterior part of the pedigree and l, excluding data on l. The remaining part after the partitioning, L_2 in the example, is the likelihood of the posterior part of the pedigree of l, including l with a scaled *prior* term. In larger pedigrees this partitioning is repeated to yield parts corresponding to single families. When repeating the partitionings, results of earlier partitionings must be taken into account, *eg*, the result that after a partitioning information on a linking individual is ignored in the family where the linking individual was a progeny.

The likelihood of a pedigree can be partitioned entirely using *prior* terms. However, the iterative computation, as will be introduced hereafter, can be speeded up by also using a partitioning of the likelihood using a *prog* term. Showing this based on the example, the likelihood L is multiplied and divided by a term representing the likelihood of family 2, ignoring data on individual 3, $L_2^* = \Sigma_{G3} \Sigma_{G4}\Sigma_{G5}\, P(G_4)\, P(G_5|G_3, G_4)\, f(y_3|G_3)\, f(y_4|G_4)$, which leads to:

$$L = \Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\, P(G_1)\, P(G_2)\, P(G_3|G_1, G_2)\, f(y_1|G_1)\, f(y_2|G_2)\, f(y_3|G_3) \\ *\{\Sigma_{G4}\Sigma_{G5}\, P(G_4)\, P(G_5|G_3, G_4)\, f(y_4|G_4)\, f(y_5|G_5)/L_2^*\}L_2^*$$

Here a term $\Sigma_{G4}\Sigma_{G5}\, P(G_4)\, P(G_5|G_3, G_4)\, f(y_4|G_4)\, f(y_5|G_5)$ has been isolated, which is $prog(G_3)$. The division by L_2^* scales this term, L_2^* being $\Sigma_{G3}\, prog(G_3)$. Hence, L is written as:

$$L = \{\Sigma_{G1}\Sigma_{G2}\Sigma_{G3}\, P(G_1)\, P(G_2)\, P(G_3|G_1, G_2) \\ f(y_1|G_1)\, f(y_2|G_2)\, f(y_3|G_3)\, prog^{sc}(G_3)\}L_2^*$$

where $prog^{sc}(G_3)$ denotes the scaled or normalised *prog* term. For a partitioning using a *prog* term it is seen that (i) in the family where the linking individual is a progeny, a $prog^{sc}$ term is added as information for the individual; and (ii) in the family where the linking individual is a parent, all information from observations and from prior terms is ignored. The scaled *prog* term for a linking individual l is generally computed as:

$$prog^{sc}(G_l) = prog(G_l)/\Sigma_{Gl}\, prog(G_l)$$

Partitioning in a nested design

In a nested design, partitionings are carried through until parts are obtained corresponding to sire families. In such families, several female parents can be present. The linking individuals are all the sires and dams of the families, except when they are in the base population. In this design we consider a partitioning using a *prog* term for each male and a *prior* term for each female that is a linking individual. When all parents of a family are in the base population, the part of the likelihood pertaining to such a family is computed as:

$$\begin{aligned} L_s = \{&\Sigma_{Gi}\, P(G_i)\, f(y_i|G_i) \\ &\Pi_j \Sigma_{Gj}\, P(G_j)\, f(y_j|G_j) \\ &\Pi_k \Sigma_{Gk}\, P(G_k|G_i, G_j)\, f(y_k|G_k)\, prog^{sc}(G_k) \\ &\Pi_l \Sigma_{Gl}\, P(G_l|G_i, G_j) \\ &\Pi_m \Sigma_{Gm}\, P(G_m|G_i, G_j)\, f(y_m|G_m)\} \end{aligned} \qquad [5]$$

where i indicates the sire of family s, j sums over the dams of the family, k indicates male progeny that are linking individuals, l indicates female progeny that are linking individuals and m indicates all other progeny. When the sire of the family is not in the base population, the term $P(G_i)\, f(y_i|G_i)$ on the first line of [5] is removed and for each dam that is not in the base population the term $P(G_j)$ on the second line

of [5] is replaced with $prior^{sc}(G_j)$. The considered partitionings using *prog* terms for all male linking individuals lead to this removal of information from sires on the first line of [5] when sires are not in the base population and lead to the inclusion of the $prog^{sc}$ for males on the third line of equation [5]. The considered partitionings using *prior* terms for all female linking individuals, lead to the inclusion of a $prior^{sc}$ term on the second line of [5] when dams are not in the base population and the removal of all information of females on the fourth line of equation [5]. Based on the results from the previous paragraph, the likelihood of the entire pedigree after the partitionings is:

$$\ln(L) = \Sigma_s \ln(L_s) \qquad [6]$$

Repeated computation of peeling equations

Iterative peeling uses repeated computation of peeling equations. The repeated computation is a method to establish the order in which equations should be handled. Therefore, iterative peeling does not require knowledge of such an order beforehand, as is required for recursive peeling.

For each individual a *prior* and a *prog* term is computed and remains stored because results of peeling terms can be required as input for the computation of other peeling terms. Iterative peeling computes a series of solutions $prior^{[0]}$, $prior^{[1]}$, *etc*, for these terms. Starting values are taken for individual i as $prior^{[0]}(G_i) = P(G_i)$, the genotype frequencies in the base population and $prog^{[0]}(G_i)$ equals 1 for all G_i. Iterative computation starts by computing $prior^{[1]}(G_i)$ for each individual i, in order of descending age. Evaluation of these *prior* terms is based on $prior^{[1]}$ terms of parents, which are available because older individuals are updated before younger individuals, and on $prog^{[0]}$ terms of sibs. Subsequently, $prog^{[1]}(G_i)$ is computed for each individual i, in order of ascending age. Evaluation of these *prog* terms is based on $prior^{[1]}$ terms of mates, on $prog^{[1]}$ terms of progeny, which are available because now younger individuals are updated before older individuals, and for female parents, on a $prog^{[0]}$ or $prog^{[1]}$ term of their male mate. Whether this last term is already updated as $prog^{[1]}$ depends on the order in which *prog* terms are computed. After computation of all $prior^{[1]}$ and $prog^{[1]}$ terms is completed, a new iteration starts computing $prior^{[2]}$ and $prog^{[2]}$, *etc.*

Starting values are such that $prior^{[0]}$ terms are correct for all individuals in the base populations, and $prog^{[0]}$ terms are correct for all individuals without progeny. Terms that can be correct after the first cycle of computations are, for instance, $prior^{[1]}$ terms of individuals descending from 2 base individuals and $prog^{[1]}$ terms of parents without grandprogeny. Correct computation of a term is shown when in the next cycle recomputed terms are equal to old terms. Once it is found that a term is correctly computed, recomputation can be omitted in following iterations of the algorithm. The order in which terms are found correct gives information on the order in which recursive peeling could be used. Generally, in each iteration, reasonably large groups of terms appear correct, keeping the number of cycles required to compute all terms correctly reasonably small, typically about the number of generations in the data set. When all terms are found correctly computed, likelihood of the data can be obtained using [5] and [6].

Application in looped pedigrees

The series of solutions $prior^{[0]}$, $prior^{[1]}$, *etc*, obtained with iterative peeling can be considered as temporary solutions for the required terms, corresponding to solutions based on a not yet fully determined peeling order. 'Temporary' likelihoods can also be computed using [5] and [6] based on a not yet fully determined order. In nonlooped pedigrees, a peeling order can eventually be found and temporary solutions become exact. In looped pedigrees, a peeling order for recursive peeling cannot be determined. In the iterative peeling algorithm the impossibility of finding a peeling order in looped pedigrees is shown by continuing changes in peeling terms. In looped pedigrees, these changes were found to decrease in size quickly and temporary likelihoods were found to stabilise, supplying an approximation. Because in iterative peeling every following update of terms includes information from 50% less related individuals, a geometric rate of convergence is plausible. As a stopping rule to use the approximation in looped pedigrees, we used the average absolute difference between subsequent normalised heterozygote probabilities, based on computed peeling terms. For convenience, only the heterozygote probability, which changed the most, was monitored.

SIMULATION STUDY

Application of iterative peeling to obtain approximate likelihoods in looped pedigrees was the aim of this study. Simulations were therefore performed to investigate the usefulness of this approximation. Because exact computations are unfeasible in large looped pedigrees, approximate likelihoods could not be compared with exact ones. Hence, an indirect way to study the approximation was found by studying the distribution of test statistics and of parameter estimates over a number of replicated analyses in looped as well as in nonlooped pedigrees. In nonlooped pedigrees exact likelihoods could be computed, serving as a reference. Simulations and analysis are based on a biallelic autosomal locus and a normal penetrance function.

Simulated data

Data sets had a nested structure each generation, with full-sibs nested within paternal half-sibs. Three different data structures were used (table I), 1 structure without loops and 2 structures with loops. The data structures were designed to contain approximately the same number of observations, the same number of base individuals (structure 1 *vs* 2) and the same family sizes (1 *vs* 3). In structures 2 and 3, the third generation was produced by taking 1 son from each sire and 1 daughter from each dam, maintaining the same breeding structure across generations. No directional selection was practised, and breeding females for a male were each taken from a different sire-family. Half- and full-sib matings were avoided, so that inbreeding was absent within the 3 generations considered. The additional third generation in structures 2 and 3 caused many pedigree loops in the form of marriage loops. All individuals used for breeding the last generation, *ie* 120 for structure 2 and 60 individuals for structure 3, were involved in 1 or more such loops, often overlapping.

Table I. Possible structures of simulated data sets.

Structure	*Generations (including parents)*	*Sires, dams and progeny (per generation)*	*Total observations*
1	2	20, 5, 10	1 120
2	3	20, 5, 5	1 120
3	3	10, 5, 10	1 060

Genotype G_i of an individual equals 1, 2 or 3 corresponding to genotypes A_1A_1, A_1A_2 and A_2A_2 at an autosomal locus. Genotypes for individuals in the base population were randomly sampled using genotype frequencies according to Hardy–Weinberg proportions, after which genotypes of other individuals were randomly sampled based on realised parental genotypes assuming Mendelian transmission probabilities. For each individual a random normally distributed environmental component was sampled and added to a pre-determined effect of each genotype to obtain a phenotypic observation. Random numbers were generated using GGUBFS and GGNQF (IMSL, 1984). Details on the parameters used for these simulations are given in the following sections.

Model and model fitting

The statistical model can be specified by the probability terms in [2], [3] and [4] which are $P(G_i)$, the genotype frequency in the base population for individual i, $P(G_i|G_s, G_d)$, the transmission probability for individual i given the genotypes of its sire s and dam d, and the penetrance function $f(y_i|G_i)$, the probability for the data y_i on individual i given the genotype G_i of individual i. From these 3 terms, transmission probabilities are assumed known to be Mendelian. Genotype frequencies in the base population depend on the unknown frequency f of the A_1 allele, assuming Hardy–Weinberg proportions of genotypes. The penetrance function for an individual i is taken as:

$$f(y_i|G_i) = (2\pi\sigma)^{-1/2}\exp\{-1/2(y - \mu_{Gi})^2/\sigma^2\}$$

This penetrance function is a normal probability density function with variance σ^2 around the mean μ_{Gi} for genotype G_i. No dominance is assumed. For analysis, means attributed to the genotypes are expressed as $\mu_1 = \mu - 1/2t$, $\mu_2 = \mu$ and $\mu_3 = \mu + 1/2t$, where t is the difference between homozygotes, referred to as the gene effect. The unknown parameters in the model are then f, μ, t, and σ^2.

Likelihoods were computed using iterative peeling. For structure 1, without loops, computations were done exactly by repeating the computations until no further changes occurred, having found the order for recursive computation. For the looped pedigrees of structures 2 and 3, iterative peeling was used to obtain approximate likelihoods. The stopping rule was a change less than 10^{-8} for the average absolute heterozygote probabilities of all individuals. The maximum of the likelihood was sought using the downhill simplex algorithm (Nelder and Mead, 1965), using as convergence criteria the variance of likelihood values of points in the simplex to be less than 10^{-12}.

Comparisons

Looped and nonlooped pedigrees were compared in hypothesis tests and parameter estimation. In hypothesis testing, a null hypothesis postulating the absence of a major gene is used, described by a model with parameters μ and σ^2, and an alternative hypothesis postulating the presence of a major gene is used, described by a model with parameters f, μ, t and σ^2. Tests are based on the likelihood ratio (LR) test statistic, which is twice the natural logarithm of the ratio of maximum likelihoods under each hypothesis. Type I error and power, the complement of type II error, were investigated at their nominal level, *ie* assuming the expected classical asymptotic χ^2 distribution for the LR test statistic under the null hypothesis (Wilks, 1938). Using the classical rules, rejection thresholds were obtained from a χ^2 distribution with 2 degrees of freedom, *ie* the difference in number of parameters between the null and alternative hypothesis. It should be noted that for testing mixtures, these classical rules do not lead exactly to the nominal type I errors (Titterington *et al*, 1985), but this is not of importance for the comparisons between looped and nonlooped pedigrees to be made here. The likelihood L_0 for the null hypothesis is computed as:

$$L_0 = \Pi_i (2\pi\sigma^2)^{-1/2} \exp\{-1/2(y_i - \mu)^2/\sigma^2\}$$

where y_i are observations with $i = 1, \ldots, N$, the total number of observations, assumed normally and independently distributed. Under the null hypothesis, the maximum likelihood estimate for the mean is $\widehat{\mu} = \Sigma y_i/N$ and for the variance is $\widehat{\sigma}^2 = \Sigma(y_i - \widehat{\mu}_0)^2/N$.

Type I error of the test for a major gene was investigated by simulating 1 000 data sets of each structure (table I), generating for each individual only a randomly distributed error term with $\sigma^2 = 100$ as phenotype. Likelihoods for the null hypothesis and the alternative hypothesis were computed in each of these replicated data sets, and the likelihood ratio test statistic was obtained. The number of significant tests in these 1 000 data sets was counted using rejection thresholds of 4.605 and 5.991, corresponding to nominal type I errors of 10 and 5%. Power to detect a major gene was investigated by simulating 100 data sets of each structure (table I) for 3 different gene effects $t = 5$, $t = 7.5$ and $t = 10$ and using allele frequency $f = 0.5$ and residual variance $\sigma^2 = 100$. Hence, relative gene effects t/σ were 0.5, 0.75 and 1. Power was based on a nominal type I error of 5%, using a rejection threshold of 5.991. Parameter estimates were compared using the 100 data sets of each structure (table I) used to investigate power with $t = 10$.

RESULTS

Type I errors were significantly lower than their nominal, *ie* asymptotically expected, level, but comparison of type I errors between looped and nonlooped structures did not show significant differences (table II). This indicates that absolute values of approximate likelihoods obtained are on average close to expected and that the distribution of the test statistic over a number of replicates is not significantly altered when loops are present. Similar conclusions can be drawn by comparing power of the test under the alternative hypothesis (table III). Parameter

estimates for gene effect under the alternative hypothesis are biased in general, but estimates for gene effects as well as allele frequency do not differ between looped and nonlooped structures (table IV). This indicates that location of the maximum is, on average over replicates, not altered for approximate likelihoods.

Table II. Estimated type I errors (%) under the null hypothesis of no major gene, given for nonlooped structures (1) and for looped structures (2, 3) based on 1 000 simulated data sets for each structure.

Structure	*Nominal level*	
	10%	*5%*
1	2.8	1.4
2	3.2	1.4
3	2.5	1.7

Table III. Estimated power (%) for a major gene test under the alternative hypothesis of presence of a major gene, given for nonlooped structures (1) and for looped structures (2, 3) based on 100 simulated data sets for each structure and for each of 3 different genetic effects.

Structure	*Genetic effect* t/σ		
	0.5	*0.75*	*1*
1	20	66	96
2	13	58	94
3	15	72	92

Table IV. Average parameter estimates for genetic effect (t) and allele frequency (f) with empirical standard errors of the mean ($\pm$ SEM) under the alternative hypothesis of presence of a major gene, given for nonlooped structures (1) and for looped structures (2, 3), based on 100 simulated data sets for each structure.

Structure	$\widehat{t} \pm SEM$	$\widehat{f} \pm SEM$
1	10.95 $\pm$ 0.30	0.479 $\pm$ 0.021
2	11.33 $\pm$ 0.23	0.499 $\pm$ 0.021
3	10.87 $\pm$ 0.25	0.501 $\pm$ 0.021

Simulated parameters: $t = 10$ and $f = 0.5$.

DISCUSSION AND CONCLUSIONS

An alternative peeling algorithm, called iterative peeling, has been presented. The iterative peeling algorithm includes an algorithm to find an order for evaluating peeling equations. When an order cannot be found, as in looped pedigrees, an approximate likelihood is supplied. In this case, use of a partitioned computation of the likelihood is also crucial. Traditional recursive peeling does not involve such approximations, because this method only computes the exact likelihood once a peeling order is found and computes the likelihood by representing all pedigree information in terms for a single individual. Usefulness of iterative peeling as an approximate method in looped pedigrees was investigated by simulations. At an aggregate level, *ie* compared on average over a number of replicated data sets, no differences were found between looped and nonlooped pedigrees. Exact computations were unfeasible due to the large number of loops in the typical animal breeding pedigrees we considered, and properties of iterative peeling could not be studied comparing exact and approximated likelihoods in individual data sets.

The iterative peeling method may be of interest for application in animal breeding. In human populations, pedigrees are generally small and loops are not abundant so that exact computations can be considered using more complicated forms of peeling (see Cannings *et al*, 1978). These more complicated forms of peeling consider genotypes on sets of individuals jointly. Larger pedigrees and more abundant looping in animal breeding, however, makes the sets of genotypes considered jointly too large for exact computations to be feasible. Therefore, approximate methods are required for application in animal breeding. Iterative peeling seems very suited, being exact without loops, and automatically supplying approximate likelihoods when loops are present. Note that, due to the partitioned computation of likelihood, iterative peeling also automatically handles pedigrees consisting of independent families, *ie* data traditionally handled with sire or sire-and-dam models. The equations and partitionings given here could be extended to allow for more general pedigrees. In particular, allowance could be made for females being mated with several males. In this case, partitionings should accommodate for 'linking individuals' being parents in several families, rather than just one. The monogenic model used could also be extended to a mixed inheritance model, the model usually required for analysis of animal breeding data. In iterative peeling only uni- and bivariate functions of genotypes are considered on single families. This can be combined with for instance a Hermitian integration (Le Roy *et al*, 1989; Knott *et al*, 1992) to include a polygenic component.

ACKNOWLEDGMENTS

This research was supported financially by the Dutch Product Board for Livestock and Meat, the Dutch Pig Herdbook Society, Bovar BV, VOC Nieuw-Dalland BV, Euribrid BV and Fomeva BV.

REFERENCES

Cannings C, Thompson EA, Skolnick MH (1976) The recursive derivation of likelihoods on complex pedigrees. *Adv Appl Prob* 8, 622-625

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. *Adv Appl Prob* 10, 26-61

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

Hoeschele I (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theor Appl Genet* 76, 81-92

IMSL (1984) *Library Reference Manual*, Edition 9.2, International and Statistical Libraries, Houston, TX

Knott SA, Haley CS, Thompson R (1992) Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity* 68, 299-311

Lange K, Elston RC (1975) Extensions to pedigree analysis I. Likelihood calculations for simple and complex pedigrees. *Hum Hered* 25, 95-105

Le Roy P, Elsen JM, Knott SA (1989) Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genet Sel Evol* 21, 341-357

Nelder JA, Mead R (1965) A simplex method for function minimization. *Comp J* 7, 147-151

Titterington DM, Smith AFM, Makov EU (1985) *Statistical Analysis of Finite Mixture Distributions*. Wiley and Sons, New York, NY

Van Arendonk JAM, Smith C, Kennedy BW (1989) Method to estimate genotype probabilities at individual loci in farm livestock. *Theor Appl Genet* 78, 735-740

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