Adjusting the Parameter Estimation of the Parentage Analysis Software *MasterBayes* to the Presence of Siblings

A thesis presented in partial fulfillment of the requirements for the degree of
Master of Applied Statistics
at Massey University, Albany,
New Zealand

Florian Heller
2009

Abstract

Parentage analysis is concerned with the estimation of a sample's pedigree structure, which is often essential knowledge for estimating population parameters of animal species, such as reproductive success. While it is often easy to relate one parent to an offspring simply by observation, the second parent remains frequently unknown. Parentage analysis uses genotypic data to estimate the pedigree, which then allows inferring the desired parameters. There are several software applications available for parentage analysis, one of which is *MasterBayes*, an extension to the statistical software package R. *MasterBayes* makes use of behavioural, phenotypic, spatial and genetic data, providing a Bayesian approach to simultaneously estimate pedigree and population parameters of interest, allowing for a range of covariate models. *MasterBayes* however assumes the sample to be a randomly collected from the population of interest. Often however, collected data will come from nests or otherwise from groups that are likely to contain siblings. If siblings are present, the assumption of a random population sample is not met anymore and as a result, the parameter variance will be underestimated. This thesis presents four methods to adjust *MasterBayes'* parameter estimate to the presence of siblings, all of which are based on the pedigree structure, as estimated by *MasterBayes*. One approach, denoted as DEP, provides a Bayesian estimate, similar to *MasterBayes'* approach, but incorporating the presence of siblings. Three further approaches, denoted as W1, W2 and W3, apply importance sampling to re-weight parameter estimates obtained from *MasterBayes* and DEP. Though fully satisfying adjustment of the estimate's variance is only achieved at nearly perfect pedigree assignment, the presented methods do improve *MasterBayes*' parameter estimation in the presence of siblings considerably, when the pedigree is uncertain. DEP and W3 show to be the most successful adjustment methods, providing comparatively accurate,

though yet underestimated variances for small family sizes. W3 is the superior approach when the pedigree is highly uncertain, whereas DEP becomes superior when about half of all parental assignments are correct. Large family sizes introduce to all approaches a tendency to underestimate the parameter variance, the degree of underestimation depending on the certainty of pedigree. Additionally, the importance sampling schemes provide at large uncertainty of pedigree comparatively good estimates of the parameter's expected values, where the non importance sampling approaches severely fail.

Acknowledgements

I would like to acknowledge a number of people for the support provided during the course of this research project.

First and foremost, my special thanks go to my supervisor, Dr Beatrix Jones. Her deep knowledge, insightful criticism, patience and continued encouragement guided me through the process of this project.

I would further like to thank Dr Howard Edwards for providing me with the helpful course material of the Bayesian Statistics paper.

A big thanks goes also to Tim Napier, who went through the hassle of proof-reading the thesis.

Lastly, I would like to thank my family, my friends and especially my partner Chen Geng for their continued encouragement and support.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1. Introduction

1.1 Introduction to Parentage Analysis

Determining the pedigree structure in a population is often of interest in biology and may be essential in drawing further inferences. Knowledge of the pedigree allows inferring population parameters such as fertility of certain population groups. While it is mostly possible to tell one parent of an offspring (in mammal species usually the mother) simply by observation, it is generally impossible to determine both parents in polygamous species. Parentage analysis approaches the problem of finding the pedigree structure by means of genetic analyses. Typically, the genotype of offspring and candidate parents is extracted at highly polymorphic microsatellite loci and then examined to estimate relational probabilities.

A likelihood-based approach to parental analysis is presented by Marshall et al. (1998) and realized in the software package CERVUS. Based on offspring's and candidate parent's genotypes (if available including the genotype of known parents), CERVUS tests the hypothesis of the proposed parent being the true parent against the hypothesis of not being the true parent. The likelihoods are the product of the likelihood at each examined locus, leading to the likelihood ratio between an offspring's two most likely parents. The ratio is presented in the natural logarithm and hence a score of zero indicates that the proposed parents are equally likely to be the true parent, whereas a positive score increases the chance of the proposed parent being the true one, or decreases respectively. Parentage for tested offspring is then assigned for the most likely parent if a pre-determined level of confidence is reached and otherwise is left unassigned. For further parameter analysis, the most likely mother and father are used. Those parameter inferences however will be separate from the pedigree estimation, using classical statistical methods, and therefore will not incorporate the uncertainty about the parental assignments but treat the assigned parents as true parents. This is, treating the assigned parents as true parents and base all following analysis on this assumption.

In a further approach, Emery et al. (2000) provide a Bayesian method to infer the number of parents and relationships within nest-structured samples. From knowledge of offspring's genotypes in a nest, it reconstructs parental genotypes participating in the nest, specifically from egg strings of veined squid, and so provides the pedigree structure, but also is able to include known parents' genotypes, if applicable, into the analysis. The approach samples from the posterior of pedigree structure by means of Monte Carlo Markov Chains (MCMC) and is realized in a more flexible form for general parentage problems than the paper specifies in the package PARENTAGE [Wilson (2001)], which provides a tool in analysing the problem of inferring the number of parents and the relationship within samples.

Jones et al. (2007) present a further Bayesian approach to estimate reproductive success simultaneously with parentage in the presence of siblings and half siblings, hence incorporating within-nest relatedness rather than treating offspring as a random sample. The paper adds the idea of jointly estimating parentage along with demographic parameters.

Hadfield (2006) provides a Bayesian joint inference procedure that is computationally realized in the library *MasterBayes* [Hadfield (2008b)] for the statistical software package R [R Development Core Team (2008)]. Allowing a range of covariate models, *MasterBayes* jointly estimates parentage and population parameter(s) of interest by simultaneously making use of behavioural, phenotypic, spatial and genetic data in a Bayesian framework via MCMC and hence the uncertainty of parentage is incorporated to the estimate. We will denote the *MasterBayes* approach in later references with the abbreviation MB. *MasterBayes* assumes a random sample from the studied population and thereby that there are no relations between the sampled offspring beyond what is expected by chance. We will frequently use the term "independence" to express this assumption.

Often however a sample will not meet the requirements of a randomly collected sample. We may for example collect samples from nests (several sampled offspring per nest) where offspring from the same nest are likely to be siblings, or sample from other groups that are likely to contain siblings. This is frequently the most efficient and

cheapest and hence most economic way to collect data. If siblings are present in the sample, *MasterBayes'* assumption of independence (and hence a random population sample) is violated.

Under independent assumptions, one parent with say three sampled offspring is counted three times, whereas it would be more appropriate to count it indeed as one parent (hence unique parent). The following example illustrates the counting differences in the case of estimating mothers.

Independent versus Unique Mothers – A Counting Example

Following the fish example in Jones et al. (2007), assume a simplified sample collected from two nests. Each nest has one father but several participating females. Females are divided into two age groups. We are interested in the fraction of mothers belonging to age class 2.

Nest 1			
Offspring	**Father**	**Mother**	**Age**
O1	F1	M1	1
O2	F1	M1	1
O3	F1	M2	2
O4	F1	M2	2
O5	F1	M2	2
O6	F1	M3	1

Nest 2			
Offspring	**Father**	**Mother**	**Age**
O7	F2	M4	1
O8	F2	M4	1
O9	F2	M4	1
O10	F2	M5	2
O11	F2	M5	2
O12	F2	M6	1
O13	F2	M7	2

Figure 1.1.1 - Nest Structure

In this example, *MasterBayes'* counts each offspring's mother according to the assumption of independence by her age class, which leads to:

$$m_1 = 7 \qquad m_2 = 6$$

where m_1 is the counted number of mothers belonging to age class 1 and m_2 is the number of counted mothers belonging age class 2.

The probability λ_{indep} to find (by independent counting) a mother from age class 2 in the sample is:

$$\lambda_{indep} = \frac{m_2}{m_{tot}} \qquad \text{Equation 1.01}$$

$$\text{with} \quad m_{tot} = m_1 + m_2 \qquad \text{Equation 1.02}$$

Each mother represents a Bernoulli draw and hence the variance of λ_{indep} over both nests is:

$$\text{Var}(\lambda_{indep}) = \frac{\lambda_{indep}\ (1 - \lambda_{indep})}{m_{tot}} \qquad \text{Equation 1.03}$$

This independent counting scheme obviously ignores the fact that there are many siblings present in the sample and hence the number of unique mothers is actually less than m_{tot}.

Counting the unique mothers in each age class instead, we get:

$$um_1 = 4 \qquad um_2 = 3$$

where um_1 is the counted number of unique mothers belonging to age class 1 and um_2 is the number of counted unique mothers belonging to age class 2. The dependent counting approach using unique mothers acknowledges that one mating event produces several siblings.

The probability λ_{dep} to find (by counting unique mothers) a mother from age class 2 in the sample is:

$$\lambda_{dep} = \frac{um_2}{um_{tot}} \qquad \text{Equation 1.04}$$

$$\text{with} \quad um_{tot} = um_1 + um_2 \qquad \text{Equation 1.05}$$

Here the variance of λ_{dep} over both nests is:

$$Var(\lambda_{dep}) = \frac{\lambda_{dep}\ (1-\lambda_{dep})}{um_{tot}} \qquad \text{Equation 1.06}$$

In this example, we get the probability and variance for age class 2 as:

λ_{indep} = 0.4615	λ_{dep} = 0.4286
$Var(\lambda_{indep})$ = 0.0191	$Var(\lambda_{dep})$ = 0.0350

Though both age class 2 probability estimates are similar, their variances differ substantially. Thus, the assumption of independence within *MasterBayes*' framework leads to misjudgement of the parental presence for samples including siblings. As a result, the parameter estimate is though likely to provide a reasonable age class 2 probability, the estimate's variance however will be severely underestimated. Inferences to the population may be compromised since MB's variance estimate appears deceptively good.

The aim of this project is to adjust *MasterBayes'* parameter estimates for the presence of siblings. This is done on the simple example of estimating the parameter of reproductive success of two age groups (more specifically, the fraction of mothers belonging to one age group), which is in its aim identical to Jones et al. (2007). Beyond this specific application however, *MasterBayes* allows for more complicated covariate models, which can be adjusted for nest samples in a similar fashion.

In the process, we simulate nesting populations where the nest's fathers are assumed to be known and the mother remains to be estimated. Samples taken from nests will include siblings (unless sampling only one offspring from each nest, which would however be very uneconomic and furthermore equivalent to independent assumptions) and therefore provide an excellent example for the presence-of-siblings situation. Several approaches will be introduced which are to adjust the *MasterBayes* estimation procedure by accounting for siblings, as well as three importance sampling schemes that re-weight *MasterBayes* estimates to construct a posterior distribution based on a sibling model.

1.2 Introduction to Genetics

To understand how parentage analysis is able to determine parentage based on data extracted from sampled genotypes, a short introduction to genetics is necessary. See for example Hartl & Clark (1997).

Genes are, simplified speaking, the physical entity that is transmitted from parents to offspring and influences hereditary traits. Physically speaking, genes are code in the "hardware" of base pairs in regions along a molecule of DNA (deoxyribonucleic acid) that is shaped in the famous form of a right-handed double helix, connected by those base pairs. The actual gene consists of a specified region of base pairs along the DNA.

The DNA is wrapped up in microscopic bodies called chromosomes, of which several exist in sets to cover all genes, depending on species. Humans for example possess 46 chromosomes; the DNA molecule in the largest human chromosome consists of 230 Million base pairs. Higher species generally not only possess one set of chromosomes, but cells host two copies of each chromosome. Thus covering the same incomplete range of genes twice, the genetic content however does differ in each copy. Species with a double set of chromosomes are called diploid.

Diploid offspring inherit one set of chromosomes from the mother and the other set from the father. Sexual cells that transfer the genotype from parent to offspring carry

only one set of chromosomes (haploid) and hence discard one set of the diploid parental chromosomes, before delivering the remaining set to the offspring. By fusion of mothers and fathers cells, the offspring receives each parent's transmitted set and hence becomes diploid.

The location of a gene along the DNA in chromosomes is referred to as the locus (plural: loci) of the gene. A gene can be present in different coding and hence hold different information. This information influences the organism's phenotype (physical features). For example, a gene governing human eye colour can be present in several different colours, blue, grey, brown, etc. Those different possibilities of how the gene coding expresses the organism's features are the alleles. In terms of eye colour, blue is one allele, while grey is another allele, etc. The percentage in which an allele is present in the population is the allele frequency.

When analyzing the genotype, one is not dealing with the total of all genes of an organism, since this is a massive amount of data, but rather with small, manageable and, for the task adequate amount of loci. In parentage analysis, the analyst would typically look at highly polymorphic microsatellites. Those are generally co-dominant (the allele information in both chromosomes will be realized in the phentotype, whereas dominant alleles can obscure non-dominant ones) and possess a relatively large amount of alleles.

A large number of possible alleles provide a high level of diversity and therefore greater power to distinguish between individuals. On a locus with few alleles, one would by chance find the same allele on many individuals, while a locus with many alleles will show the same allele in fewer individuals.

Since an organism's locus contains characteristics on two chromosomes, two pieces of information (bits) are present on an individual's locus. We do know the parental genotype, but we do not know which bit is discarded during the transfer to the offspring. Consequently, four possible combinations exist, of how a father and a mother can pass their total four bits on to their offspring's two bits.

This is the key to parentage analysis. The parental combinations of allele on each locus are compared with the offspring's combination on the same locus. The

likelihood of being related is determined by how often the offspring's locus information can be found in the proposed parental 4-bit combination.

In the example below the offspring's locus is assumed to contain the allele "a" on each chromosome. We propose a female that contains the alleles "a" and "b" on both of her chromosomes on the same locus as well as a male that contains the alleles "a" and "c". We then determine the likelihood that the female and the male are the offspring's parents.

Fig. 1.2.1 shows that one locus only allows for limited permutations of likelihoods, namely 0, 0.25, 0.5 0.75 and 1. The likelihood of finding the genetic information in the offspring, given the examined father and mother is denoted as p(G|M,F).

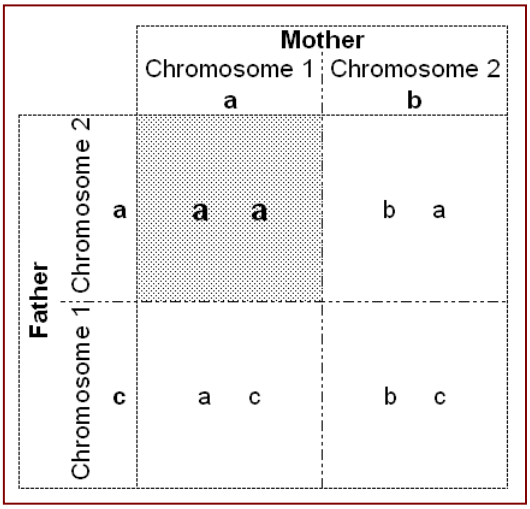

Figure 1.2.1 - Likelihood Determination

In the example illustrated in Fig 1.2.1, of the four possible combinations of parental bits, only one ("a" "a", shaded) matches the offspring's locus characteristic. One match out of four possible combinations yields a likelihood of p(G|M,F) = 0.25 for this specific locus.

The overall parental likelihood over several examined loci is obtained by multiplying each locus' likelihood:

$$p(G|M,F)_{\text{All Loci}} = \prod_{i=1}^{r} \left(p(G_i \mid M, F) \right) \qquad \text{Equation 1.07}$$

with r loci, $p(G_i|M,F)$ denoting the likelihood at the i'th examined locus.

Note that we wish to genotype independent (unlinked) loci, since dependence between examined would result in a decreased exclusion power, because knowledge

of a locus would also give information about a linked locus, which would decrease the amount of information, that this linked locus can introduce to the analysis.

It is clear that the likelihood is not on an absolute scale, but a relative likelihood that allows comparison with other potential parents. It is important to note that the likelihood p(G|M,F) = 0.25 in the above case does not mean that 0.25 is the actual probability that this is the parent's offspring, but just the likelihood given of finding the offspring's locus information given the proposed parents.

It is further clear that one locus has limited power to determine parentage, which depends on the number of alleles that are present in the specific locus. It is easy to see that the greater the number of allele present on the locus, the more informative a match becomes. With increasing number of alleles on a locus, the chance decreases that high likelihoods for actually unrelated individuals occur. If for example, there are only two alleles possible, many adults are expected to produce high likelihoods even if not being related to the offspring, while their number will drop, as the number of possible alleles increases. Therefore, highly polymorphic (containing many alleles) microsatellites are preferred.

The exclusion power however not only depends on the number of possible alleles, but also on the percentage with which an allele is present in the population (the allele frequency). Here we prefer ideally equifrequent alleles, or at least low frequencies for each allele present on the locus, because an allele with a high frequency would dominate in presence. For example, a 4-allele locus with frequencies 0.97, 0.1, 0.1 and 0.1 would provide a lower exclusion power than a locus with 2 equifrequent alleles, since 97% of all examined chromosomes would show the same allele and thus would be indistinguishable.

A genetic markers ability to exclude a given relationship between offspring and potential parents is expressed as exclusion probability [Gerber et. al (2000)]. We can exclude a candidate parent to be an offspring's true parent when its likelihood = 0. In order to determine parentage reliably, large exclusion probability is desired.

For co-dominant markers (as used in this project) three different approaches to determine exclusion probability exist, depending on available knowledge of pedigree

and goal. Single parent exclusion (assigning one parent to an offspring, the other parent remains unknown), paternity or maternity exclusion (assigning one parent to an offspring, the other parent is known) and parent pair exclusion (assigning both parents to an offspring and hence both are unknown).

As it will be described in chapter 2.1, fathers are assumed to be known in this project, while mothers remain to be estimated, therefore the appropriate approach is maternity exclusion. Jamieson & Taylor (1997) provide the maternity exclusion probability P_M for a co-dominant locus:

$$P_M = 1 - 2a_2 + 2a_3 + 2a_4 - 3a_5 - 2a_2^2 + 3a_2a_3 \qquad \text{Equation 1.08}$$

The a_k are calculated as:

$$a_k = \sum_{i=1}^{m} p_i^k \qquad \text{Equation 1.09}$$

where m is the number of different alleles on the locus, p_i is the frequency of each allele and index k allows for powers.

Exclusion probability can be increased either by examining loci with more allele, or by increasing the number of examined loci. The general overall maternal exclusion probability $P_{M,tot}$ when examining k independent (unlinked) loci is calculated as:

$$P_{M,tot} = 1 - \prod_{i=1}^{k} \left(1 - P_{M,i}\right) \qquad \text{Equation 1.10}$$

where $P_{M,i}$ is the maternal exclusion probability (equation1.08) at the i'th locus.

2. Methodology

2.1 Sample Simulation

When using real life data, sampled from actual populations, the pedigree and the true underlying distribution parameters remain unknown, which makes it virtually impossible to assess the accuracy of the analysis.

To verify the validity of the methods, we simulate population samples that provide known population parameters that the results of the analysis of the simulated sample can be checked against. Because the simulated samples are created according to pre-determined parameters, the true values of estimation are either known, or can be derived from the exact knowledge of the population parameters.

Therefore the simulation delivers the data to be analyzed that would be otherwise gathered from a real sample, but also provides knowledge of the true parentage and population parameters, which can be compared to the analyses results and so allows to assess the methods' estimation accuracy (which will however not reflect the part of estimated variance that is due to the uncertainty of parentage). Thus, simulated samples provide a powerful tool in assessing the ability of the analysis procedure.

The inadequacy of assumptions in *MasterBayes'* approach (denoted as MB) which we intend to correct is introduced by the presence of siblings in the sample. Note however that, if one was sure that the sample does not contain more siblings than one could expect from a random sample of a population, there was no need for adjustment. Consequently, the desired simulation procedure has to create samples that contain siblings. We choose the reproductive behaviour of fish to serve as a model, based on which the samples analyzed in this study will be simulated.

Cottus bairdi, the fish species modelled in Jones et al. 2007, is a ployamorous nest breeding species. A male usually guards his nests, not sharing nests with other males, while females only deposit their eggs in it. In the sample simulation, we assume one male per nest, who will be all the nest-offspring's father. Note that this is an idealization, since this assumption ignores the possibility of cuckolding fathers.

Though *MasterBayes* is able to incorporate cuckolding parents in its analysis, we ignore this possibility in this project for reasons of simplicity: We are only concerned with providing methods to adjust MB estimates, not however in validating the procedure itself. Similarly, we ignore occurrence of mutation or typing error.

Simulation Procedure

The following describes how the simulated sample is created. It is important to emphasis, that this is not a simulated population (since a further possibility could be, to simulate a population, from which a sample may be collected)! The idea of the approach is to characterize the underlying population by probability distributions, according to which the simulated sample will be drawn.

We define a number of nests to serve as sample. To each nest, we assign a male by a random draw with replacement from the pool of all available males. Since the males are drawn with replacement, any male can be assigned to several nests, each nest however has only one assigned male. The latter guarantees that every nest sample holds offspring that share the same father and therefore are at least half-siblings.

Though fatherhood can be determined from the pedigree knowledge provided by the sample, this assumption also provides for practical (non-simulated) purposes the knowledge of a nest's true father by taking the nest-guarding male as the exclusive father of the guarded nest's offspring. Each offspring's mother however remains to be estimated by *MasterBayes*.

Note that the constellation of known father and unknown mother is somewhat different to most mammal species, where one would usually know the offspring's mother, while the father would be subject to estimation. Here mammals' offspring are mostly taken care of by their mother and hence motherhood can be easily determined. The analysis principle however does not depend on genders; the approach is equivalent for any animal species where one parent is known.

Since we are modelling a ployamorous species, it is likely that not only one, but several females deposit their eggs in any given nest and hence a nest might consist of half and full siblings. In the simulation of the sample, the number of females

participating in a nest is governed by a Poisson (θ) distribution for which we choose a parameter θ = 2.87. The Poisson distribution is truncated at zero in order to avoid nests with no participating females. The truncation at zero causes the mean number of mothers we expect per nest not to be θ = 2.87, but rather:

$$E\left(\frac{\#\ Mothers}{Nest}\right) = \frac{\theta}{1-e^{-\theta}} = 3.043 \qquad \text{Equation 2.01}$$

The simulation so far has modelled a nest structure, where all offspring are at least half-siblings, sharing the same father and the participation of several mothers introduces groups of full-siblings. The number of sampled offspring per nest will be denoted as family size.

Note that until here only the number of participating females in each nest is determined, but not yet the assignment of specific female individuals (and hence age classes). Each nest-participating female is so far only a placeholder that yet has to be assigned an actual female.

While population parameters concerning the paternal assignments are somewhat unimportant, since male parameters are not the subject of estimation, we are interested in how well we can adjust MB's estimation of female parameters in the presence of siblings. As the parameter of interest, we choose the reproductive success of females depending on their age (fraction of unique mothers belonging to examined age), where age is divided into two classes, for example old versus young, and hence is a binary variable. Each age class's presence in the sample is specified. We choose to estimate and adjust the reproductive success of age class 2, though one could just as well choose age class 1.

The simulated samples used during this test series consist of 900 adults, divided into 57% (513) females and 43% (387) males. 70% (359) of all females are assigned age class 1, and hence 30% (154) to age class 2. Since not the *MasterBayes* procedure itself, but the adjustment to presence of siblings is tested, no unsampled parents are assumed to be present and therefore each offspring's true parents are present in the

sample. 30 sample nests are created, in which a certain amount of offspring (family size) are placed (all of which are part of the sample). The family size varies throughout the analysis.

As the parameter to estimate, we determine age class 2 females to gain maternity in 42.16% within the population and hence the female age class 2 parameter λ_{Pop} is set to 0.4216, i.e. the fraction of reproducing females in age class 2 (in the population).

In the simulation, this is realized by assigning specific mothers from each age class to the place-holding number of females in nests that was determined by the zero truncated Poisson (2.87) as explained above. We fill the place-holders for mothers by assigning a random age 2 female with probability $\lambda_{Pop} = 0.4216$, and an age class 1 female otherwise. A female may gain maternity in several nests.

Note that age class 2 assignment with probability λ_{Pop} is equivalent to λ_{Pop} being the true underlying population parameter. The actual number of mothers in each age class realized in each sample will be somewhat close to λ_{Pop}, but will however not quite turn out as such, because it is a probabilistic assignment (moving from population to sample).

λ_{Pop} governs the population age 2 assignment, the estimation however is done on a population sample and yields therefore, at best λ, the age class 2 parameter that is actually realized in the sample. Sufficiently increasing the number of sampled nests would cause λ to converge to λ_{Pop}.

We will estimate the sample age 2 parameter λ (the percentage of females in age class 2 that gain maternity), denoting the estimate as $\hat{\lambda}$. The true value of λ as realized in each sample can be calculated from the known, and hence not in real life applicable, number of sampled unique mothers by equation 1.04.

It is important to note that λ represents the fraction of mothers belonging to age group 2 in the sample, which however is *not* the actual reproductive success of age class 2, because it does not pay respect to the number of females of each age class in the sample. To illustrate this, note that $\lambda_{Pop} = 0.4216$ will lead to approximately 42% of all mothers being age class 2 and hence approximately 58% of all mothers will be age

class 1. We however chose the presence of age class 2 females to be 30% of all females and hence the actual reproductive success of age class 2 is actually greater than age class 1.

Since the scope of this project is to adjust MB estimates and all samples are simulated with equal population parameters, the estimates $\hat{\lambda}$ are comparable for all scenarios and hence suitable for assessing adjustment success.

The simulated sample now has fathers and mothers assigned to offspring. Offspring's genotypes are constructed from known parental genotype, where the diploid haplotypes of adults are constructed according to specified number of allele present and their frequency at each locus, as shown below. Note that the number of examined loci varies throughout the analysis. Parental genotypes are passed on to offspring by transferring one of both haploids from each parent to the offspring and hence the offspring's genotype will consist of half of its mother's and half of the father's genotype.

Adult genotypes are constructed using equal settings for each examined locus. We use m = 5 alleles, each allele with frequency $p_i = 0.2$. Thus, adult genotypes remain fixed throughout the different runs of simulations. Loci are unlinked and hence independent from each other (any given locus' properties do not influence any other locus' properties). As a result, every locus provides an equal maternal exclusion probability. Note that the overall exclusion probability reflects the knowledge of the genotype.

Though we will mostly refer to the number of the examined loci when assessing the quality of estimates at certain scenarios, the locus references are only valid for the specific loci settings of this project. Generalisations beyond the project-specific can however be made by referring to the maternal exclusion probabilities that each scenario provides by its number of examined loci and their definitions.

Equation 1.08 provides the maternal exclusion probability $P_M = 0.6352$ for each locus in the simulated sample. Further, we can calculate and plot (Fig. 2.1.1) the overall maternity exclusion probabilities $P_{M,tot}$ (equation 1.10) depending on the number of examined loci as far as this project is concerned.

The knowledge of having an offspring's true mother in the sample and the number of total females present in the sample (513) allows to calculate via exclusion probability the number of females (# Non Excl Fem) for any given offspring that, additionally to the true mother, are on average feasible for the offspring (based on genetic data). With equation 1.10 we can calculate this number as:

$$\#\ \text{Non Excl Fem} = (1 - P_{M,tot}) * f_{tot} \qquad \text{Equation 2.02}$$

where f_{tot} = 513 is the total number of females in the sample.

From the number of feasible non-excludable females, we can obtain the expected maternal assignment success-rate (the rate of correctly assigned mothers) given the examined loci by:

$$\text{Assignment Success-Rate} = \frac{1}{(\#\ Non\ Excl\ Fem) + 1} \qquad \text{Equation 2.03}$$

where (+ 1) accounts for the true mother.

Note that the so obtained success-rates are naïve estimates, since they are based on exclusion only and hence assume that each of the non-excluded females has an equal chance to be the true mother. The approach by MCMC however is expected to achieve higher success-rates, because the non-excluded females yet differ in their likelihoods, which leads to a better-than-random maternal assignment.

Fig. 2.1.1 on the following page shows for each independent locus in the analysis the numbers of non-excluded females in the table as well as maternity exclusion probabilities and the naïve maternal assignment success-rates plotted versus the number of independent loci.

# Independent Loci	# Non Excl Fem
1	187.142
2	68.270
3	24.905
4	9.085
5	3.314
6	1.209
7	0.441
8	0.161
9	0.059
10	0.021
11	0.008
12	0.003

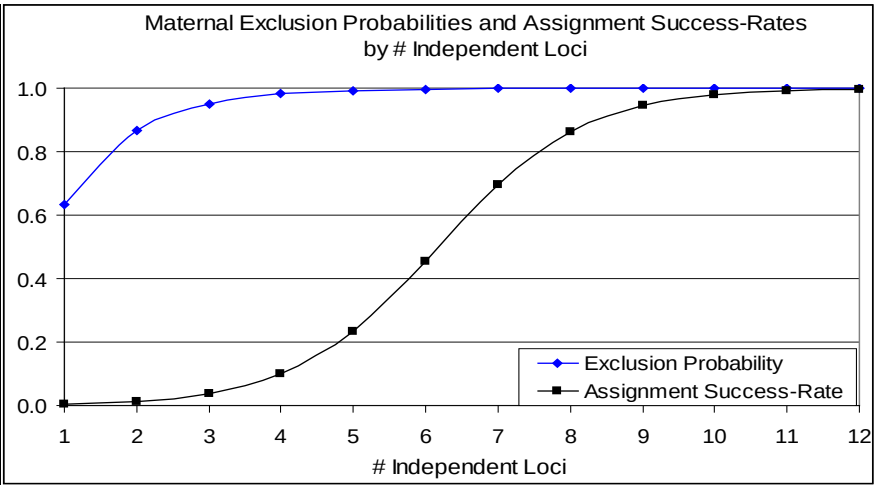

Fig. 2.1.1 - Number of Non-Excluded Females, Exclusion Probabilities and Naïve Success-Rates

2.2 Introduction to *MasterBayes*

MasterBayes [Hadfield (2008b)] is a software extension of the statistical package R [R Development Core Team (2008)]. *MasterBayes* depends on several further R extensions, namely the packages *coda* [Plummer et. al. (2008)], *genetics* [Warnes (2008)], *gtools* [Warnes (2008)] and *kinship* [Atkinson & Thereneau (2008)].

MasterBayes applies MCMC to integrate over uncertainty in pedigree configurations estimated from molecular markers and phenotypic data. Emphasis is put on the marginal distribution of parameters that relate phenotypic data to the pedigree. All computation is done in compiled C++ for efficiency.

Population parameters are estimated by a Bayesian MCMC approach, as described in the following chapters. *MasterBayes* allows to fit a range of log-linear covariate models [Hadfield (2008)] that can be written in the general case as:

$$P_{i,j}^{(O)} \propto e^{(\beta_1 x_1 \ldots)} \qquad \text{Equation 2.04}$$

where $P_{i,j}^{(O)}$ is the probability that proposed parents i and j are the true parents of offspring O and β is the vector of associated parameter(s). Any number of covariates

x may be included with regard to proposed mother, father or both. Possible expressions may be continuous variables [eg.: $\beta_1 * x_k$], sums [eg.: $\beta_1(x_i + x_j)$], differences [eg.: $\beta_1(x_i - x_O)$], absolute differences [eg.: $\beta_1(|x_i - x_j|)$], distances [eg.: $\beta_1 \sqrt{(lat_j - lat_O)^2 + (long_j - long_O)^2}$ if x are coordinates] between mates or mates and offspring, and more. For further details on available covariate models, see Hadfield (2008).

An adequate model for this project, where the aim is to estimate and adjust for maternal age classes with two categorical levels (age 1, age 2), is provided by:

$$P_{i,j}^{(O)} \propto e^{(\beta_1 \delta_i)} \qquad \text{Equation 2.05}$$

where δ_i takes the value 1 if mother i's age class is 2, and 0 otherwise. Index i denotes females. Since paternity is assumed to be known, index j, which refers to males, is redundant in the model.

MasterBayes estimates the age class 2 parameter within the framework of a logistic model and expresses it as parameter $\hat{\beta}$, which is the logarithm of age class 2 vs age class 1:

$$e^{\hat{\beta}} = \frac{m_2 / f_2}{m_1 / f_1} \qquad \text{Equation 2.06}$$

where m_1 is the number of age class 1 mothers, counted according to the independent scheme, and m_2 the number of age class 2 mothers respectively. Note that independent counting in the presence of siblings introduces the underestimation of the parameter variance (refer to chapter 1.1, section *Independent versus Unique Mothers – A Counting Example*) that we intend to correct. f_1 is the number of age class 1 females and f_2 is the number of age class 2 females in the sample.

By inclusion of f_1 and f_2, $\hat{\beta}$ accounts for imbalanced presence of each age group. As mentioned in chapter 2.1, we will work with the probability estimate $\hat{\lambda}$. The logit transformation allows to convert $\hat{\beta}$ to $\hat{\lambda}$ by:

$$\hat{\lambda}_{indep} = \frac{e^{\hat{\beta}}}{C + e^{\hat{\beta}}} \qquad \text{Equation 2.07}$$

$$\text{with} \quad C = \frac{f_1}{f_2} \qquad \text{Equation 2.08}$$

where the constant C accounts for the number of females in each age class, hence the imbalance of age class presence.

Since independent counting is applied by *MasterBayes*, the age 2 parameter estimate is denoted as $\hat{\lambda}_{indep}$.

Vice versa $\hat{\lambda}_{indep}$ can be re-transformed to $\hat{\beta}$ by:

$$\hat{\beta} = \ln\left(\frac{\hat{\lambda}_{indep}}{1 - \hat{\lambda}_{indep}} C\right) \qquad \text{Equation 2.09}$$

Not that the procedure is easily applicable to estimate the parameter for age class 1 instead or, if maternity is known instead of paternity, estimate male age parameters. The following chapters 2.2.1 and 2.2.2 will show in detail, how $\hat{\lambda}_{indep}$ and maternal assignments are estimated.

2.2.1 MCMC Parental Assignment Estimation via Gibbs Sampler in *MasterBayes*

Mothers are assigned to offspring via Gibbs sampling [S. Geman, D. Geman. (1984)]. For each offspring, the Gibbs sampler generates a candidate mother M_p conditional on $\hat{\lambda}_{indep}$ from the pool of m available females in the sample.

The probability $p(M_p \mid \hat{\lambda}_{indep})$ of a female to become an offspring's candidate mother is calculated for every female in the sample, the actual candidate mother is then assigned according to her probability $p(M_p \mid \hat{\lambda}_{indep})$.

$p(M_p \mid \hat{\lambda}_{indep})$ is obtained using Bayes Theorem as:

$$P(M_p \mid \hat{\lambda}_{indep}, G) = \frac{p(G \mid M_p, \hat{\lambda}) * p(M_p \mid \hat{\lambda}_{indep})}{\sum_{i=1}^{m} \left(p(G \mid M_i, \hat{\lambda}) * p(M_i \mid \hat{\lambda}_{indep}) \right)} \qquad 1 \le p \le m \qquad \text{Equation 2.10}$$

where the sum in the denominator over all m females in the sample provides the normalizing constant.

The components $p(G \mid M, \hat{\lambda})$ and $p(M \mid \hat{\lambda}_{indep})$ are obtained as follows:

$\underline{p(G \mid M, \hat{\lambda})}$

This is equivalent to the likelihood of the parental combination $p(G|M,F)_{All\ Loci}$ (equation 1.07), which we however will denote as $p(G|M, \hat{\lambda})$ for convenience of display. Therefore $p(G|M,F)_{All\ Loci} = p(G|M, \hat{\lambda})$. We can make this simplification in notation because within this project's framework we treat the fathers as known. Though we still look at the likelihood of the examined genetic data, given the examined females and males, the fathers are known and hence the changes over parental combination introduced to $p(G|M, \hat{\lambda})$ are due to proposal of mothers only and hence the female genetic part.

Typically, several loci are examined in this fashion, where the accuracy of parentage exclusion increases with the number of loci.

$p(M \mid \hat{\lambda}_{indep})$

The probability $p(M \mid \hat{\lambda}_{indep})$ is the best guess without knowledge of genetic data, what the probability of a female given her age class is, to be the offspring's true mother. Knowledge of a female's age group and her age group's reproductive success allows to give the probability as the chance of the female being the true mother is as good as the chance of any other female from the same age group. Both age groups however differ in their probability of providing the true mother, according to $\hat{\lambda}_{indep}$.

Being provided with $\hat{\lambda}_{indep}$, a female belonging to age class 1 has a non-genetic probability of being the true mother:

$$p(M_{(age\ class\ 1)} \mid \hat{\lambda}_{indep}) = \frac{1 - \hat{\lambda}_{indep}}{f_1} \qquad \text{Equation 2.11}$$

A female belonging to age class 2 has the non-genetic probability of being the true mother:

$$p(M_{age\ class\ 2)} \mid \hat{\lambda}_{indep}) = \frac{\hat{\lambda}_{indep}}{f_2} \qquad \text{Equation 2.12}$$

The first MCMC iteration provides an estimate $\hat{\lambda}_{indep}$, which updates $p(M \mid \hat{\lambda}_{indep})$ in the second iteration. The second iteration's estimate $\hat{\lambda}_{indep}$ again updates $p(M \mid \hat{\lambda}_{indep})$ in the third iteration, and so on.

Note that *MasterBayes* assigns candidate parents conditional on $\hat{\lambda}_{indep}$. This links successive iterations by a somewhat weak bond that nevertheless influences the maternal assignments. Since $\hat{\lambda}_{indep}$ is obtained under independent counting, the assignment of mothers depends on *MasterBayes*' independent assumptions.

Adjustment techniques for sibling structure introduced later, use maternal assignments and hence still will suffer in quality, because each iteration's parental assignments are influenced by independent assumptions through $p(M \mid \hat{\lambda}_{indep})$.

2.2.2 *MasterBayes* Age 2 Parameter Estimate (MB)

This original *MasterBayes* approach we will denote with the abbreviation MB.

Being provided with the estimated maternal assignments, the number of estimated mothers of each age class can be easily obtained by counting. Since the iterations' assignments come from sample data, inferences to the age parameter of the population need to incorporate a more sophisticated statistical approach.

MasterBayes uses Bayes Theorem to make inferences to the population via the Metropolis-Hastings scheme [Hastings (1970)]. The idea behind the Metropolis-Hastings scheme is to propose in every iteration a parameter estimate $\hat{\lambda}_{indep,p}$ and compare it with the previously accepted parameter estimate $\hat{\lambda}_{indep,c}$, making a decision, whether to accept or reject the proposal.

The reasoning of the Bayesian approach is that the maternal age class 2 assignments observed in the sample are generated by the underlying true age class 2 probability λ_{Pop}. Bayes Theorem allows by incorporating a prior to infer the posterior distribution, which is the underlying age class 2 probability $\hat{\lambda}_{indep}$ (under independent assumptions) given the observed data, denoted $p(\hat{\lambda}_{indep,i} \mid m_{2,i}, G)$, where index i indicates the iteration.

Since the normalizing constant is difficult to obtain, the posterior is modeled up to a constant of proportionality as:

$$\underbrace{p(\hat{\lambda}_{indep,i} \mid m_{2,i}, G)}_{posterior} \propto \underbrace{p(\hat{\lambda}_{indep,i})}_{prior} * \underbrace{p(m_{2,i} \mid \hat{\lambda}_{indep,i}, G)}_{likelihood} \qquad \text{Equation 2.13}$$

The likelihood term represents the knowledge we gain from observing the sample. This knowledge from the sample can be expressed by the parental assignments conditional on $\hat{\lambda}_{indep,i}$ that are provided by the Gibbs step, as shown in the previous section. The likelihood is modeled by a binomial distribution with probability $\hat{\lambda}_{indep,i}$, $m_{tot,i}$ draws and $m_{2,i}$ successes:

$$p(m_{2,i} | \hat{\lambda}_{indep,i}, G) = \binom{m_{tot,i}}{m_{2,i}} \hat{\lambda}_{indep,i}^{m_{2,i}} (1 - \hat{\lambda}_{indep,i})^{m_{1,i}} \quad \text{Equation 2.14}$$

The prior $p(\hat{\lambda}_{indep,i})$ is realized by a draw from a uniform (0,1) distribution, by proposing a candidate age class 2 estimate $\hat{\lambda}_{indep,p}$. The proposed parameter estimate $\hat{\lambda}_{indep,p}$ is compared with the current age class 2 parameter estimate $\hat{\lambda}_{indep,c}$ that was accepted in the previous iteration (and hence $\hat{\lambda}_{indep,c} = \hat{\lambda}_{indep,i-1}$) in the Hastings Ratio (HR).

Because the parental assignments are equal in numerator and denominator, the normalizing constant is equal and hence cancels out of HR. As a result, HR can be written as equality:

$$HR = \underbrace{\frac{p(\hat{\lambda}_{indep,p} | m_{2,i}, G)}{p(\hat{\lambda}_{indep,c} | m_{2,i}, G)}}_{\textit{likelihood ratio}} * \underbrace{\frac{p(\hat{\lambda}_{indep,p} \rightarrow \hat{\lambda}_{indep,c})}{p(\hat{\lambda}_{indep,c} \rightarrow \hat{\lambda}_{indep,p})}}_{\textit{ratio of proposal densities}} \quad \text{Equation 2.15}$$

where the ratio of proposal densities (see more details below in section HR < 1) in two directions ($\hat{\lambda}_{indep,p}$ to $\hat{\lambda}_{indep,c}$ and vice versa) evaluates to 1, since the proposal is symmetric and thus numerator and denominator are equal. This special case of the Metropolis-Hastings scheme is also known as Metropolis algorithm.

Using equation 2.13 and the fact that the ratio of proposal densities evaluates to 1, we can rewrite HR as:

$$HR = \frac{p(\hat{\lambda}_{indep,p}) * p(m_{2,i} \mid \hat{\lambda}_{indep,p}, G)}{p(\hat{\lambda}_{indep,c}) * p(m_{2,i} \mid \hat{\lambda}_{indep,c}, G)} \qquad \text{Equation 2.16}$$

Proposed $p(\hat{\lambda}_{indep,p})$ and current $p(\hat{\lambda}_{indep,c})$ priors are both uniform (0,1), from which it follows that $p(\hat{\lambda}_{indep,p}) = p(\hat{\lambda}_{indep,c})$, and hence the ratio of priors evaluates to 1. Thus, HR simplifies to only comparing the likelihoods:

$$HR = \frac{p(m_{2,i} \mid \hat{\lambda}_{indep,p}, G)}{p(m_{2,i} \mid \hat{\lambda}_{indep,c}, G)} \qquad \text{Equation 2.17}$$

which can be written with equation 2.14 as:

$$HR = \frac{\lambda_{indep,p}^{m_{2,i}} \ (1-\lambda_{indep,p})^{m_{1,i}}}{\lambda_{indep,c}^{m_{2,i}} \ (1-\lambda_{indep,c})^{m_{1,i}}} \qquad \text{Equation 2.18}$$

HR provides a comparative measure, which age class 2 parameter ($\hat{\lambda}_{indep,p}$ or $\hat{\lambda}_{indep,c}$) is more likely to be the true parameter given the maternal assignments obtained from the sample. Depending on the value of HR, the Metropolis-Hastings scheme either accepts $\hat{\lambda}_{indep,p}$ to become the iteration's parameter estimate $\hat{\lambda}_{indep,i}$, or rejects it and keeps $\hat{\lambda}_{indep,c}$ as the iteration's parameter estimate $\hat{\lambda}_{indep,i}$. The decision rules whether or not to accept $\hat{\lambda}_{indep,p}$ are summarized below.

<u>HR ≥ 1</u>

$\hat{\lambda}_{indep,p}$ is more than, or at least as likely to be the true parameter as $\hat{\lambda}_{indep,c}$ is. The Metropolis-Hastings algorithm accepts the proposal and sets $\hat{\lambda}_{indep,p} = \hat{\lambda}_{indep,i}$.

Therefore, whenever the algorithm finds an equally or more likely proposal, it will accept it.

HR < 1

$\hat{\lambda}_{indep,p}$ is less likely to be the true parameter than $\hat{\lambda}_{indep,c}$. It is however not impossible that $\hat{\lambda}_{indep,p}$ is the true parameter. Consider that, if the algorithm would strictly reject proposals for HR < 1, it would stop moving once it draws the proposal with the highest probability and thus pay no respect to this possibility.

The Metropolis-Hastings scheme copes with this issue by drawing a random value x from the uniform (0,1) proposal distribution. Note that technically x is also generated for HR ≥ 1, since however $x_{max} = 1$, the draw can be ignored because HR ≥ 1 is larger than $x_{max} = 1$ and hence equivalent to accepting the proposal without a draw.

If x < HR, $\hat{\lambda}_{indep,p}$ is accepted, though being less likely the true estimate than $\hat{\lambda}_{indep,c}$ and hence the iteration's parameter estimate is set to $\hat{\lambda}_{indep,p} = \hat{\lambda}_{indep,i}$. Thus, the algorithm is still able to move, even if the current estimate $\hat{\lambda}_{indep,c}$ is the more likely estimate.

If x > HR, $\hat{\lambda}_{indep,p}$ is rejected and hence the current estimate $\hat{\lambda}_{indep,c}$ is set to be the iteration's parameter estimate by $\hat{\lambda}_{indep,c} = \hat{\lambda}_{indep,i}$. This is equivalent to keeping the previous iterations parameter estimate, since the current estimate was previously introduced by $\hat{\lambda}_{indep,i-1} = \hat{\lambda}_{indep,c}$.

HR < 1 not only provides the information that the proposal is less likely than the current estimate, but the value of HR also reflects, how much less likely the proposal is. Therefore, slightly less likely proposals will be accepted more often than much less likely proposals and thus the probabilistic character is maintained.

The more loci are examined in the analysis, the greater the exclusion power of maternal assignments and hence the less variation we expect in the maternal assignments. As a result, we expect fewer accepted proposals, the more loci we examine in the analysis. The constancy with which parameter estimates are accepted over the course of the MCMC is a measure for how likely they are, to be the true parameter.

Reviewing the Metropolis-Hastings Algorithm Steps

1. Initialize MCMC by drawing a parameter estimate $\hat{\lambda}_{indep,i-1}$ from a uniform (0,1)
2. Gibbs step. Generate candidate parents for each offspring conditional on $\hat{\lambda}_{indep,i-1}$
3. Set the former parameter estimate as current parameter estimate

   $\hat{\lambda}_{indep,i-1} = \hat{\lambda}_{indep,c}$
4. Draw proposal estimate $\hat{\lambda}_{indep,p}$ from a uniform (0,1)
5. Calculate HR
6. If HR ≥ 1 then $\hat{\lambda}_{indep,p} = \hat{\lambda}_{indep,i}$

   If HR < 1 then draw x from proposal uniform (0,1)

   If x < HR then $\hat{\lambda}_{indep,p} = \hat{\lambda}_{indep,i}$

   If x > HR, then $\hat{\lambda}_{indep,c} = \hat{\lambda}_{indep,i}$
7. Index i = i + 1
8. Repeat steps 2 to 8 until convergence.

The posterior distribution of the age class 2 parameter estimate $p(\hat{\lambda}_{indep,i} \mid m_{2,i}, G)$ consists of all $\hat{\lambda}_{indep,i}$ collected by the chain, which could for example be visualised by plotting in a histogram. Note that *MasterBayes* works in the framework of an exponential model (equation 2.05), and hence $\hat{\lambda}_{indep,i}$ are converted into $\hat{\beta}_i$ (equation 2.09) and stored. In this project, we will however refer to $\hat{\lambda}_{indep,i}$ only.

The age class 2 estimate can be summarized by expected value and variance of the posterior.

We get the posterior expected value as sample mean:

$$\mathrm{E}_{\mathrm{MB}}(\hat{\lambda}_{indep} \mid m_2, G) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\lambda}_{indep,i} \right) \qquad \text{Equation 2.19}$$

and the posterior variance as sample variance:

$$\mathrm{Var}_{\mathrm{MB}}(\hat{\lambda}_{indep} \mid m_2, G) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{\lambda}_{indep,i} - E_{MB}(\hat{\lambda}_{indep} \mid m_2, G) \right)^2 \qquad \text{Equation 2.20}$$

where n is the number of MCMC iterations.

It is clear that more iterations will produce more $\hat{\lambda}_{indep,i}$ draws and hence the posterior will increase in resolution with increasing number of iterations.

2.3 Dependent Estimation Approach (DEP)

This approach we will denote with the abbreviation DEP.

The idea behind the dependent estimation approach is, to construct each iteration's posterior distribution from the maternal assignments estimated by the Gibbs sampler, followed by combining the iteration wise posteriors to the final age class 2 parameter posterior estimate.

Unlike MB, which builds the posterior on the basis of counting the independent number of mothers, the DEP approach constructs the posteriors from the number of unique mothers and thus includes the presence of siblings in its assumptions.

Note however that maternal assignments are influenced by independent assumptions, as described in chapter 2.2.1, which compromises the dependent assumption to a certain extent.

Constructing the DEP Posterior

Analogous to equation 2.14 we get the dependent likelihood, counting unique mothers as:

$$p(um_{2,i}|\hat{\lambda}_{dep,i},G) = \binom{um_{tot,i}}{um_{2,i}} \hat{\lambda}_{dep,i}^{um,_{2,i}} (1-\hat{\lambda}_{dep,i})^{um_{1,i}} \qquad \text{Equation 2.21}$$

The DEP posterior is obtained analogous to equation 2.13, using dependent equivalents, however. Note that the following is based on the same principals as the MB approach, though different in appearance.

The prior $p(\hat{\lambda}_{dep,i})$ obtained from a uniform (0,1) distribution is equivalent to a beta($\alpha = 1$, $\beta = 1$) distribution. A beta (1,1) distribution is a conjugate prior for the binomial distribution (the likelihood term), leading to a beta-distributed posterior [Gelman et al. (2004)] as follows:

$$p(\hat{\lambda}_{dep,i} \mid um_{2,i}, G) \propto \frac{\hat{\lambda}_{dep,i}^{um_{2,i}+\alpha-1} (1-\hat{\lambda}_{dep,i})^{um_{1,i}+\beta-1}}{B(\alpha+um_2, \beta+um_1)} = \frac{\hat{\lambda}_{dep}^{um_{2,i}} (1-\hat{\lambda}_{dep,i})^{um_{1,i}}}{B(1+um_{2,i}, 1+um_{1,i})} \qquad \text{Equation 2.22}$$

where $B(\alpha + um_{2,i}, \beta + um_{1,i})$ is the beta function:

$$B(\alpha+um_{2,i}, \beta+um_{1,i}) = \frac{\Gamma(\alpha+um_{2,i})\ \Gamma(\beta+um_{1,i})}{\Gamma(\alpha+um_{2,i}+\beta+um_{1,i})} \qquad \text{Equation 2.23}$$

After transforming into the general notation of a beta distribution by taking $\alpha_{dist,dep,i} = \alpha + um_{2,}i$ and $\beta_{dist,dep,i} = \beta + um_{1,i}$ we get each iteration's posterior as:

$$p_i\left(\hat{\lambda}_{dep,i} \mid um_{2.i}, G\right) \propto \frac{\Gamma(\alpha_{dist,dep,i}+\beta_{dist,dep,i})}{\Gamma(\alpha_{dist,dep,i})\Gamma(\beta_{dist,dep,i})} \hat{\lambda}_{dep,i}^{\alpha_{dist,dep,i}-1} (1-\hat{\lambda}_{dep,i})^{\beta_{dist,dep,i}-1} \qquad \text{Equation 2.24}$$

Since α = 1 and β = 1, the beta posteriors distribution parameters $\alpha_{dist,dep,i}$ and $\beta_{dist,dep,i}$ can be expressed in numbers unique mothers counted in each iteration's maternal assignments for each age class as:

$$um_{2,i} = \alpha_{dist,dep,i} - 1 \Rightarrow \alpha_{dist,dep,i} = um_{2,i} + 1 \qquad \text{Equation 2.25}$$

$$um_{1,i} = \beta_{dist,dep,i} - 1 \Rightarrow \beta_{dist,dep,i} = um_{1,i} + 1 \qquad \text{Equation 2.26}$$

where $um_{1,i}$ is the number of unique age group 1 mothers, $um_{2,i}$ is the number of unique age group 2 mothers in the i'th iteration. Expected value and variance are obtained for each iteration from the well known beta distribution as:

$$E_{DEP,i}\left(\hat{\lambda}_{dep,i} \mid um_{2,i}, G\right) = \frac{um_{2,i}+1}{um_{tot,i}+2} \qquad \text{Equation 2.27}$$

$$Var_{DEP,i}\left(\hat{\lambda}_{dep,i} \mid um_{2,i}, G\right) = \frac{(um_{2,i}+1)(um_{1,i}+1)}{(um_{tot,i}+2)^2(um_{tot,i}+3)} \qquad \text{Equation 2.28}$$

We combine the iteration-wise mean and variance to obtain the overall posterior mean and variance.

The overall expected value is constructed by taking the average over all iterations expected values:

$$E_{DEP}\left(\hat{\lambda}_{dep} \mid um_2, G\right) = \frac{1}{n}\sum_{i=1}^{n}\left(E_{DEP,i}(\hat{\lambda}_{dep,i} \mid um_{2,i}, G)\right) \qquad \text{Equation 2.29}$$

The overall DEP posterior variance has not only to account for each iteration's posterior variances, but also for the change in location of the posteriors between iterations. This is somewhat different to the MB approach, which accounts for variances and location change the draws accepted by Metropolis-Hastings over the course of the MCMC. Note however that an equivalent to the MB approach for DEP could be realized by drawing values $\hat{\lambda}_{dep,i}$ from each iteration's posterior. Vice versa

MasterBayes could abandon the Metropolis-Hastings scheme and draw each iteration's estimate from the posterior constructed from the estimated mothers delivered by the Gibbs Sampler.

If we view the parent assignments at each iteration as multiple imputations of missing variables, Meng (1994) provides a way of combining the within and between iteration variances. We construct the DEP final posterior variance from the variance of expected values (between component) and the posterior variances in each iteration (within component).

The within imputation component $\overline{W_K}$ is calculated by averaging the posterior variances (equation 2.28) of all iterations:

$$U_{K,DEP} = \frac{1}{n}\sum_{i=1}^{n}\left(Var_{DEP,i}\left(\hat{\lambda}_{dep,i} \mid um_{2,i}, G\right)\right) \qquad \text{Equation 2.30}$$

The between imputation component B_K is calculated by obtaining the sample variance across the expected values $E_{DEP,i}\left(\hat{\lambda}_{dep,i} \mid um_{2,i}, G\right)$ of all iterations:

$$B_{K,DEP} = \frac{1}{n-1}\sum_{i=1}^{n}\left(E_{DEP,i}(\hat{\lambda}_{dep,i} \mid um_{2,i}, G) - E_{DEP}(\hat{\lambda}_{dep} \mid um_2, G)\right)^2 \qquad \text{Equation 2.31}$$

The overall MCMC estimated posterior variance is then calculated by:

$$Var_{DEP}\left(\hat{\lambda}_{dep} \mid um_2, G\right) = U_{K,DEP} + \frac{n+1}{n} B_{K,DEP} \qquad \text{Equation 2.32}$$

Though this method avoids direct use of $\hat{\lambda}_{indep,i}$, the independence assumption still influences it through the parental assignments sampled by the Gibbs Sampler, where successive iteration's parental assignments are linked by the $p(M \mid \hat{\lambda}_{indep}, G)$ term.

2.4 Analytical Derivation of the True Dependent Parameters from Simulation Data (TRUE)

This approach we will denote with the abbreviation TRUE.

The TRUE approach analytically constructs the posterior distribution analogous to DEP, using the known parentage structure of the simulation though. Therefore, the TRUE approach is not of practical relevance as an adjustment method, since the parentage structure is either not known, or if known the parameters would be estimated easily by classical methods and hence make the MCMC approach redundant.

The TRUE approach however provides a tool to validate the success of adjustment approaches applied on simulated samples. Note that the knowledge of pedigree eliminates the need of repeated sampling and thus is simply constructed from the numbers of unique mothers in each age group.

Analogous to equation 2.24 TRUE posterior has pdf:

$$p(\hat{\lambda}_{dep} \mid um_{2,sim}, G) \propto \frac{\Gamma(\alpha_{dist,sim} + \beta_{dist,sim})}{\Gamma(\alpha_{dist,sim})\Gamma(\beta_{dist,sim})} \lambda_{dep}^{\alpha_{dist,sim}-1} (1-\lambda_{dep})^{\beta_{dist,sim}-1} \quad \text{Equation 2.33}$$

with parameters:

$$um_{2,sim} = \alpha_{dist,sim} - 1 \Rightarrow \alpha_{dist,sim} = um_{2,sim} + 1 \quad \text{Equation 2.34}$$

$$um_{1,sim} = \beta_{dist,sim} - 1 \Rightarrow \beta_{dist,sim} = um_{1,sim} + 1 \quad \text{Equation 2.35}$$

where λ_{dep} is the dependent age class 2 parameter, $um_{1,sim}$ is the number of unique age group 1 mothers, $um_{2,sim}$ is the number of unique age group 2 mothers in the sample and $m_{tot,sim}$ is the sum of both.

Expected value and variance of the TRUE posterior are obtained by:

$$E_{TRUE}(\lambda_{dep}|um_{2,sim},G) = \frac{um_{2,sim}+1}{um_{tot,sim}+2} \quad \text{Equation 2.36}$$

$$Var_{TRUE}(\lambda_{dep}|um_{2,sim},G) = \frac{(um_{2,sim}+1)(um_{1,sim}+1)}{(um_{tot,sim}+2)^2(um_{tot,sim}+3)} \quad \text{Equation 2.37}$$

Those are the true population parameters, which will provide the guideline to evaluate the MCMC adjustment approaches.

It is however important to note that the TRUE approach delivers results from the known distribution and thus does not reflect any uncertainty about maternal assignments. Any successful estimating technique should ideally either deliver greater variances than the TRUE approach does (thereby reflecting uncertainty in maternal assignments), or at best delivers an equal variance (when all mothers are correctly assigned and hence there remains no uncertainty in maternal assignments).

As a result, the TRUE approach does not deliver posterior variances that are accurate for every tested scenario, but rather supplies a lower bound for the estimate's variance that should be approached with increasing accuracy of maternal assignments.

2.5 Importance Sampling Schemes

Importance sampling is a technique to approximate a distribution from which either no direct draws can be made, or are difficult to make. We want to evaluate an integral of the form:

$$\int_x q(x) f(x) dx \quad \text{Equation 2.38}$$

where f(x) may be an expensive or impossible to find density [Hörmann & Leydold (2005)]. The key to approximation is to find an importance sampling or proposal density g(x) that is cheaper or possible to draw from. Sampling from g(x) instead, allows evaluation of the integral as:

$$\int_x q(x)w(x)g(x)dx \qquad \text{Equation 2.39}$$

where w(x) is the weight function:

$$w(x) = \frac{f(x)}{g(x)} \qquad \text{Equation 2.40}$$

which supplies the weights assigned to each iteration of the MCMC.

Approximate proportionality of f(x) and g(x) is desirable, since otherwise greatly differing weights are obtained and hence a few observations dominate the sample [Monahan, J. F. (2001)]. See more details in chapter 3.1.5.

In this project, the idea behind importance sampling is to adjust age class 2 estimates for the presence of siblings by weighting each of the MCMC iterations according to the proposed importance sampling schemes and thus assigning more weight to iterations that the schemes deem to be “better” and less weight to those the schemes deem to be “worse”.

The adjusted age class 2 parameter and hence its expected value and variance are then provided by the importance sampled posterior. Three importance sampling schemes are proposed.

2.5.1 Importance-Sampling on MB (W1)

This approach we will denote with the abbreviation W1.

The W1 importance sampling approach re-weights the MB age 2 parameter estimates $\hat{\lambda}_{indep,i}$ by the posteriors MB [g(x) = $p(\hat{\lambda}_{indep,i} \mid m_{2,i}, G)$] and DEP [f(x) = $p(\hat{\lambda}_{indep,i} \mid um_{2,i}, G)$], obtained in each iteration. Thus, W1 re-weights the MB approach by the DEP approach.

Each iteration's weights are obtained as:

$$W_{1,i} = \frac{f(x)}{g(x)} = \frac{p(\hat{\lambda}_{indep,i} \mid um_{2,i}, G)}{p(\hat{\lambda}_{indep,i} \mid m_{2,i}, G)} \qquad \text{Equation 2.41}$$

which, after applying equations 2.13, 2.14 and 2.24, simplifies to:

$$W_{1,i} = \hat{\lambda}_{indep,i}^{\,um_{2,i} - m_{2,i}} \; (1 - \hat{\lambda}_{indep,i})^{um_{1,i} - m_{1,i}} \qquad \text{Equation 2.42}$$

We fit the weights in a comparable scale by standardizing:

$$w_{1,i} = \frac{W_{1,i}}{\sum_{i=1}^{n} (W_{1,i})} \qquad \text{Equation 2.43}$$

The importance sampling process weights q(x) = $\hat{\lambda}_{indep,i}$ and hence expected value of the W1 importance sampled age class 2 parameter expected value is obtained by:

$$E_{W1}(\hat{\lambda}_{\text{indep}} \mid m_2, G) = \sum_{i=1}^{n} \left(w_{1,i} * \hat{\lambda}_{indep,i} \right) \qquad \text{Equation 2.44}$$

For obtaining the age class 2 parameter variance $B_{K,W1}$, we have to consider each iteration's weight and further account for the standardization of the weights by multiplying with the number of iterations (n).

$$B_{K,W1} = \frac{n}{n-1} \sum_{i=1}^{n} \left(w_{1,i} \left(\hat{\lambda}_{indep,i} - E_{W1}(\hat{\lambda}_{indep} \mid m_2, G) \right)^2 \right) \qquad \text{Equation 2.45}$$

Preferably, we want the variability in weights to be small (have similar weights). Since we cannot alter the weights and have to work with what we get though, we can merely take their variability into account. Because the weights may differ considerably, they add extra variability and thus the number of iterations may be a poor measure for sample size. Meng (1993) provides a simple remedy to account for variability in weights that leads to the overall W1 age class 2 parameter's variance:

$$Var_{W1}(\hat{\lambda}_{indep} \mid m_2, G) = \left(1 + \frac{1 + s_{W1}^2}{n} \right) B_{K,W1} \qquad \text{Equation 2.46}$$

where

$$s_{W1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(w_{1,i}\ n\ -1 \right)^2 \qquad \text{Equation 2.47}$$

2.5.2 Importance-Sampling on Draws from Dependent Posterior (W2)

This approach we will denote with the abbreviation W2.

Different to W1, where we re-weighted $\hat{\lambda}_{indep,i}$, the W2 scheme re-weights dependent estimates $\hat{\lambda}_{dep,i}$. Each iteration's $\hat{\lambda}_{dep,i}$ are obtained by draws from the corresponding DEP posteriors (equation 2.24) that are constructed from each iteration's numbers of unique mothers.

The weight function re-weights $p(m_{2,i} \mid \hat{\lambda}_{indep,i}, G) * p(\hat{\lambda}_{dep,i} \mid um_{2,i}, G)$, which is a joint density of dependent estimates and independent parent assignments, to obtain

$p(\hat{\lambda}_{dep,i}, um_{2,i} |, G)$, the joint density of dependent estimates and maternal assignments. Incorporating dependent assumptions in the sampling distribution g(x) appears more consistent with the dependent assignment of siblings and thus is expected to improve the importance sampling process. The weighting function is:

$$W_{2,i} = \frac{f(x)}{g(x)} = \frac{p(\hat{\lambda}_{dep,i}, um_{2,i} |, G)}{p(m_{2,i} | \hat{\lambda}_{indep,i}, G) * p(\hat{\lambda}_{dep,i} | um_{2,i}, G)} \quad \text{Equation 2.48}$$

where the numerator $p(\hat{\lambda}_{dep,i}, um_{2,i} |, G)$ is a joint density and hence can be rewritten as $p(um_{2,i} | \hat{\lambda}_{dep,i}, G) * p(\hat{\lambda}_{dep})$. Since the prior on $p(\hat{\lambda}_{dep})$ is a uniform(0,1) distribution, the marginal density evaluates to $p(\hat{\lambda}_{dep})$ = 1, which simplifies the weighting function to:

$$W_{2,i} = \frac{p(um_{2,i} | \lambda_{dep,i}, G)}{p(m_{2,i} | \lambda_{indep,i}, G) * p(\lambda_{dep,i} | um_{2,i}, G)} \quad \text{Equation 2.49}$$

Using equations 2.21, 2.14 and 2.24, we can express the weight function as:

$$W_{2,i} = \frac{\lambda_{dep,i}^{um_{2,i}} (1-\lambda_{dep,i})^{um_{2,i}}}{\lambda_{indep,i}^{m_{2,i}} (1-\lambda_{indep,i})^{m_{2,i}} * \frac{\Gamma(\alpha_{dist,dep,i} + \beta_{dist,dep,i})}{\Gamma(\alpha_{dist,dep,i})\Gamma(\beta_{dist,dep,i})} \lambda_{dep,i}^{um_{2,i}} (1-\lambda_{dep,i})^{um_{2,i}}} \quad \text{Equation 2.50}$$

where it is easy to see that the dependent likelihood terms cancel and the weighting function simplifies to:

$$W_{2,i} = \frac{B(\alpha_{dist,dep,i}, \beta_{dist,dep,i})}{\lambda_{indep,i}^{m_{2,i}} (1-\lambda_{indep,i})^{m_{2,i}}} \quad \text{Equation 2.51}$$

The weights $W_{2,i}$ are standardized to $w_{2,i}$ analogous to equation 2.43.

The W2 importance sampled age class 2 parameter's expected value is obtained by:

$$E_{W2}(\hat{\lambda}_{\text{dep}} \mid um_2, G) = \sum_{i=1}^{n} \left(w_{2,i} \; \hat{\lambda}_{dep,i} \right) \qquad \text{Equation 2.52}$$

The weighted age class 2 parameter's variance is obtained analogous to W1, using dependent equivalents for $B_{K,W2}$ and weights $w_{2,i}$ for s^2_{W2} however.

$$Var_{W2}(\hat{\lambda}_{dep} \mid um_2, G) = \left(1 + \frac{1 + s^2_{W2}}{n} \right) B_{K,W2} \qquad \text{Equation 2.53}$$

2.5.3 Rao-Blackwellized Importance-Sampling on Dependent Posterior (W3)

This approach we will denote with the abbreviation W3.

W3 goes one step further than W2 by eliminating the need for sampling $\hat{\lambda}_{\text{dep,i}}$. The scheme instead weights each iteration's DEP posterior expected value $E_{DEP,i}\left(\hat{\lambda}_{dep,i} \mid um_2, G\right)$ (equation 2.27). Note that $\hat{\lambda}_{dep,i}$ is not observed, but rather depends on the maternal assignments estimated in each iteration and hence W3 provides a Rao-Blackwellized estimator for the age class 2 parameter.

The weight function is:

$$W_{3,i} = \frac{f(x)}{g(x)} = \frac{\int_{\lambda_{dep}} p(um_2 \mid \lambda_{dep,i}, G) * p(\lambda_{dep,i})}{p(m_2 \mid \lambda_{indep,i}, G)} \qquad \text{Equation 2.54}$$

From Bayes Theorem we know that $p(um_2 \mid \lambda_{dep}, G) * p(\lambda_{dep}) = p(\lambda_{dep} \mid um_2, G)$, the DEP beta posterior.

Thus, using equation 2.24 we get:

$$W_{3,i} = \frac{\int_{\lambda_{dep}} \left(\frac{\Gamma(\alpha_{dist,dep,i} + \beta_{dist,dep,i})}{\Gamma(\alpha_{dist,dep,i})\Gamma(\beta_{dist,dep,i})} \lambda_{dep,i}^{um_{2,i}} (1-\lambda_{dep,i})^{um_{2,i}} \right) \partial\lambda_{dep}}{\lambda_{indep,i}^{m_{2,i}} (1-\lambda_{indep,i})^{m_{2,i}}} \qquad \text{Equation 2.55}$$

which solves to:

$$W_{3,i} = \frac{1}{B(\alpha_{dist,dep,i}, \beta_{dist,dep,i}) \ \lambda_{indep,i}^{m_{2,i}} \ (1-\lambda_{indep,i})^{m_{2,i}}} \qquad \text{Equation 2.56}$$

The weights $W_{3,i}$ are standardized analogous to equation 2.43 to $w_{3,i}$.

The W3 importance sampled age class 2 parameter expected value is obtained by:

$$E_{W3}(\hat{\lambda}_{\text{dep}} \mid um_2, G) = \sum_{i=1}^{n} \left(w_{3,i} * E_{DEP,i}(\hat{\lambda}_{dep,i} \mid um_{2,i}, G) \right) \qquad \text{Equation 2.57}$$

Unlike schemes W1 and W2, where the importance sampled posterior draws were weighted, scheme W3 provides a similar environment like DEP (chapter 2.3), where the age class 2 parameter's variance has to account for each iteration's posterior variance (within component), as well as the variance of the expected values over iterations (between component). Also here we look at the maternal assignments at each iteration as multiple imputations of missing variables and hence apply Meng (1994) to combine the within and between iteration variances.

The within-imputation variability is given by $U_{K,W3}$:

$$U_{K,W3} = \sum_{i=1}^{n} \left(w_{3,i} * Var_{DEP,i}(\hat{\lambda}_{dep,i} \mid um_{2,i}, G) \right) \qquad \text{Equation 2.58}$$

using the DEP age class 2 parameter variance $Var_{DEP,i}(\hat{\lambda}_{dep,i} \mid um_{2,i}, G)$ (equation 2.28).

The between-imputation variability is given by $B_{K,W3}$:

$$B_{K,W3} = \frac{n}{n-1}\sum_{i=1}^{n}\left(w_{3,i} \left(E_{DEP,i}(\hat{\lambda}_{dep,i} \mid um_{2,i}, G) - E_{W3}(\hat{\lambda}_{dep} \mid um_2, G)\right)^2 \right) \qquad \text{Equation 2.59}$$

Analogous to equation 2.47, using weights $w_{3,i}$ however, we account for the extra variability caused by the weights via s^2_{W3}.

The overall age class 2 parameter's variance is calculated as:

$$Var_{W3}(\hat{\lambda}_{dep} \mid um_2, G) = \left(1 + \frac{1 + s_{W3}^{\ 2}}{n}\right) B_{K,W3} \ + \ U_{K,W3} \qquad \text{Equation 2.60}$$

2.6 Overview of Methods

In the previous chapters four methods were introduced that aim to improve the estimates provided by the independent *MasterBayes* (**MB**) approach.

1. Dependent Estimation Approach (**DEP**) *Chapter 2.3*

   The approach constructs the dependent posterior, using estimated maternal assignments.

2. Importance-Sampling on MB (**W1**) *Chapter 2.5.1*

   The W1 scheme re-weights MB estimated $\hat{\lambda}_{indep,i}$ based on likelihoods obtained by independent assumptions and dependent parental assignments.

3. Importance-Sampling on Draws from Dependent Posterior (**W2**) *Chapter 2.5.2*

   The W2 scheme provides a sampling distribution that is more consistent with the dependent assignment of siblings, re-weighting $\hat{\lambda}_{dep,i}$.

4. Rao-Blackwellized Importance-Sampling on Dependent Posterior (**W3**) *Chapter 2.5.3*

The W3 scheme progresses from W2 by eliminate the need for sampling $\hat{\lambda}_{dep,i}$ and Rao-Blackwellizing the weighted estimate.

Chapter 2.4 additionally provides a method (**TRUE**) to extract the true estimate parameters from the simulated samples, paying however no respect to uncertainty in maternal assignments, which cannot be avoided in an estimation process. Though not applicable in real life, this approach is substantially important to assess the quality of each of the above methods.

3. Analysis and Discussion

3.1 MCMC Settings and Verification

The following chapters describe test set-up and some MCMC diagnostics.

3.1.1 Set-Up of Scenarios

To evaluate the effectiveness of the developed methods, several population samples with varying numbers of examined loci and family sizes were simulated and analyzed.

The number of examined loci was chosen between 3 and 12. Scenarios using 12 examined loci provide nearly completely correct maternal assignments. Therefore using more than 12 loci would not considerably increase the assignment quality.

Family size is given by the numbers of offspring, which are sampled from each nest. Since offspring were sampled from 30 different nests (refer to chapter 2.1, section *Simulation Procedure*), a total of 30 * family size offspring were present in each scenario's sample. Family size was varied between 3 and 20 offspring.

Fig. 3.1.1.1 below shows all loci/ family size combinations that were simulated and analyzed during this project, as well as the total number of offspring present in each scenario's sample.

		Family Size								
		3	4	5	6	7	9	12	16	20
Examined Loci	3	90	120	150	180	210	270	360	480	600
	4									
	5									
	6									
	7									
	8									
	12									

Fig. 3.1.1.1 - Examined Combinations and Number of Offspring

3.1.2 Burn-In and its Sufficiency

As described in chapter 2.2.1, the Gibbs Sampler assigns candidate parents conditional on $\hat{\lambda}_{indep}$. After initiation, the Markov chain is likely to wander through parameter space, accepting unlikely proposals of $\hat{\lambda}_{indep}$ until it approaches convergence around the best estimate. Because the proposal of candidate mothers is conditional on $\hat{\lambda}_{indep}$, the maternal assignments are influenced by this, which again influences the acceptance of $\hat{\lambda}_{indep}$. The chain needs to run a certain number of iterations to produce better $\hat{\lambda}_{indep}$ estimates and maternal assignments. Once the chain arrives, we say the chain has converged.

Since the parameter estimates and maternal assignments obtained before convergence are of low quality, it is common practice to discard those from the analysis and construct the final posterior only of estimates that were obtained after the chain's convergence. Those discarded iterations are referred to as burn-in.

The determination of the burn-in sufficiency, and hence the number iterations needed to convergence, is a difficult problem. There is no known method to pre-determine the burn-in length. Further one can never verify with absolute certainty that the chain has converged, since the chain may have converged around a local maximum. We test burn-in sufficiency by setting a burn-in according to guesswork and then check the post burn-in results for consistency.

In the following test, the burn-in was set to discarding the first 3,000 MCMC iterations. The MCMC was run for 3,000 post burn-in iterations. Those were split up into the first 100 post burn-in iterations and the succeeding 2,900 iterations. For both, the maternal assignment success-rates (see next chapter 3.1.3) were obtained. The idea behind testing for convergence is that if the maternal assignment success-rates of both groups are similar, it seems that the chain has arrived and remains at the high probability mothers and hence has converged. The possibility though, that the chain has only converged around a local maximum and might wander further in future iterations, can never be excluded.

The following table shows the success-rates of the first 100 post-burn-in and the succeeding 2,900 iterations of randomly chosen loci/family size combinations:

Number of examined Loci	Family Size	Success-Rate 100	Success-Rate 2,900	Difference
3	3	0.039	0.056	0.0170
4	20	0.133	0.134	0.0012
5	4	0.265	0.272	0.0070
6	16	0.474	0.473	0.0004
7	12	0.753	0.753	0.0007
8	5	0.867	0.864	0.0030
12	9	0.999	0.999	0.0001

Figure 3.1.2.1 - Burn-In Success-Rates

The difference between the first 100 post-burn-in iterations and the succeeding 2,900 post-burn-in iterations appear to decrease, the more loci are examined. This observation fits reasonably to what would be expected, since the success-rates themselves strongly depend on the number of examined loci and hence fewer examined loci would show greater variability.

Conclusive, all scenarios provide comparable success-rates for the first 100 and the succeeding 2,900 post burn-in iterations and thus it appears relatively safe to assume that a burn-in of 3,000 iterations is sufficient to make the MCMC converge at maternal assignments. This is however neither a guarantee of actual convergence. We may call it educated guesswork.

Based on this finding, all analysis was undertaken with a 3,000 iterations burn-in.

3.1.3 Maternal Assignment Success-Rate

It is of interest to see how well the Gibbs sampler captures maternal assignments in each scenario, since the quality of those is essential to the quality of the estimates. Note that since fathers are known, it is only of interest to assess the quality of maternal assignments.

Maternal uncertainty is ideally expected to result in a wider age class 2 parameter estimated variance, by adding the uncertainty of maternity (a greater variety of plausible maternal assignments across iterations results in a greater overall

variance). The higher the achieved success-rate though, the closer the age class 2 parameter estimates of successfully adjusted approaches should approach the TRUE variance (see however this chapter's section *Family Size Effect on Unique Mothers* below), whereas we expect the MB approach to underestimate the age class 2 parameter variance (see chapter 1.1, section *Independent versus Unique Mothers – A Counting Example*).

Note that in the MB approach a misassigned mother still has a certain chance to belong to the same age class than the true mother. In the case of equal age classes of misassigned and true mothers, the MB age class 2 parameter estimate would not change, since the approach would still count the same total numbers in each age class. In reality, however, the chance of a misassigned mother to be assigned an age class depends on the age balance in the sample.

The issue of miss-assignments is more complex for counting unique mothers. For details, see this chapter's section *Family Size Effect on Unique Mothers* below.

From all scenarios analysed in this project, the maternal success-rates were obtained and plotted in Fig. 3.1.3.1 on the next page. The combinations vary the number of examined loci as well as family size. The maternal assignment success-rates are given as the percentages of correctly assigned mothers. These percentages include all iterations and hence are an average over the whole course of the MCMC. The maternal assignment success-rates are plotted bold.

The naïve maternal success-rates found in chapter 2.1 are plotted thin in corresponding colours and line types.

The graph shows clearly, that increasing the number of examined loci improves the maternal assignment success. Fewer loci result in lower power and thus, the fewer loci examined, the lower is the expected maternal assignment success-rate.

Family size has no visible effect on the assignment success, nor is it expected to.

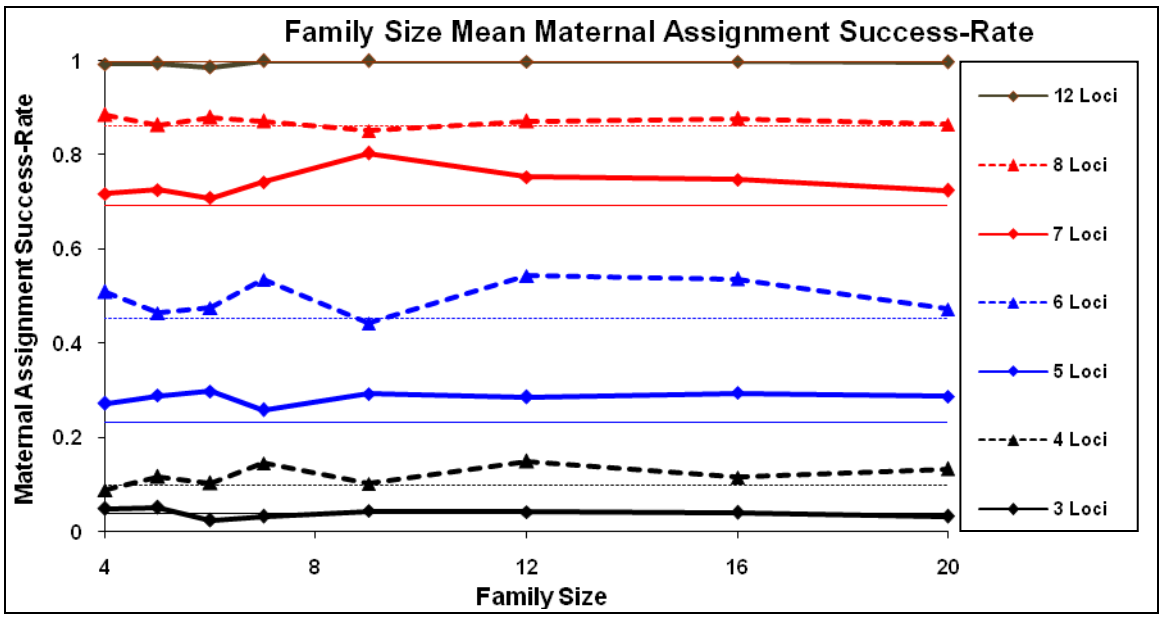

Figure 3.1.3.1 - Family Size Mean Maternal Assignment Success-Rate

The Gibbs sampler assigns mothers notably better than the naïve success-rate, based on maternal exclusion only (see chapter 1.2) predicts. Note however that the largest differences between MCMC success-rate and naive success-rate are observed at assignment success-rates of around 0.5. The differences get smaller the closer the assignment success gets to either total assignment success (1), or random assignment ($1/f_{tot}$).

Approaching total success by naïve maternal exclusion leaves less opportunity for the Gibbs sampler to improve by the advantage of likelihood, since better-than-full assignment success cannot be achieved. On the other hand, the Gibbs sampler's likelihood advantage shows only small improvements for very bad assignment successes, since here maternal assignment cannot be worse than random.

The Gibbs sampler's advantage of likelihood shows up most efficient against the naïve approach at medium assignment success.

At 12 examined loci, the Gibbs sampler approaches nearly fully correct assignment. Though the candidate mothers are generated conditional on $\hat{\lambda}_{indep}$ and hence under

independent assumptions, the genetic exclusion power provided by 12 loci is overwhelming against the independent influence.

Family Size Effect on Unique Mothers

It is important to note that, though the maternal assignment success-rates remain constant over family size, increasing family size has a negative impact on the estimation of unique mothers. The assignment success-rate provides information about how many mothers are correctly assigned to offspring. With increasing family size, the samples hold more offspring for each unique mother (since the number of females participating in a nest is held constant, as shown in chapter 2.1, equation 2.01).

In the independent case, a miss-assignment would only lead to a different assigned mother, the total number of independent mothers however would not change.

A miss-assignment in the dependent case on the other hand would lead to the introduction of a new, but false unique mother. If the misassigned unique mother is already present as mother (say we missassign a mother to an offspring, but the misassigned mother has actually offspring in another sampled family), the total number of unique mothers present in the sample will not increase. If the misassigned mother is not already present as mother of another sampled offspring, the miss-assignment increases the total number of unique mothers. Generally, we would expect the latter case happening more often, since usually a sample would consist of much more females than mothers.

With larger family size (and hence more offspring of each unique mother) the chance of misassignment remains the same for each offspring, but the chance of introducing false unique mothers increases. The following example illustrates this.

Assume a maternal assignment success-rate of 0.75 and hence a misassignment-rate of 0.25. For illustrative simplification, we ignore the possibility that a false assigned unique mother is already present as a unique mother, but rather assume that a misassignment will introduce an additional false unique mother. For a unique mother of 4 offspring we would expect 1 misassignment, which would result in the

introduction of 1 false unique mother. For a unique mother of 20 offspring however, we would expect 5 misassignments that would result in the introduction of 5 false unique mothers. We see that for equal assignment success-rates larger family size introduces more false unique mothers.

We also expected more false unique mothers, the fewer loci we examine simply for the reason of lower maternal assignment success.

Estimating more unique mothers than actually are present will cause the posterior variance to be underestimated, by the same mechanisms that cause the independent approach to underestimates the posterior variance (as shown in chapter 1.1, section *Independent versus Unique Mothers – A Counting Example*).

Note that though uncertainty of maternal assignments should result in increased parameter variance, the introduction of false unique mothers causes the opposite by decreasing the parameter variance. Increased uncertainty of maternal assignments is a proper effect, reflecting our state of knowledge of the pedigree, decreased variance by introduction of false unique mothers however does not reflect any real knowledge we have about the sample, but introduces a bias.

We cannot determine, how much influence each of those effect has on the estimated variance, which is a serious limitation to the technique. For any variance result, it would be difficult to tell, which part of the estimated variance is due to parameter variance (in the simulations supplied by TRUE, but not applicable in real life), which part is due to pedigree uncertainty and how much variance reduction is introduced by false unique mothers.

3.1.4 Thinning Interval

MasterBayes allows to specify a thinning interval, which is not to record every $\hat{\lambda}_{indep,i}$ and corresponding parental assignments, but rather recording only every z^{th} estimate (and hence thinning = z). Thinning is not a necessity from the theoretical point of view, but practically available computer memory and hence data storage problems might dictate the use of thinning.

The Markov chain's moving speed in parameter space influences the iterations needed until convergence is achieved. Slow travel is indicated by high autocorrelation, since the new proposals depend on the current state. If the chain was to run indefinitely, every region would have the same influence on the result; practically however, the needed data storage space might run out before convergence is achieved.

Alternatively to record a given number of successive iterations, one can use a thinning and thereby record only every z^{th} iteration's results and discards all iterations in between. Since then no longer successive MCMC iterations are recorded, the autocorrelation will decrease and hence the recorded iterations will depend less on the previous iteration. High autocorrelation is undesirable, since the stronger the dependence on the previous value is, the less "fresh" information is supplied by the new iteration.

The thinned MCMC covers the same parameter space than the un-thinned MCMC, by demanding far less data storage capacity (by discarding iterations). The effective iterations that the analysis is based on are related to the actual number of iterations that the MCMC runs through by:

$$\text{Effective Iterations} = \frac{1}{\text{Thinning}} * \text{MCMC Iterations} \qquad \text{Equation 3.01}$$

where Effective Iterations are the number of iterations that are recorded, Thinning is the thinning interval and MCMC Iterations the number of actual iterations that the Markov chain travels.

Assessing convergence is generally a problem for which no absolute solution exists. One can be more confident about convergence the longer the result appears to settle at the same value, though this is no solid proof of actual convergence. Thus, when using the term convergence we mean that the results appear to have settled at a

parameter estimate, we can however never be quite sure that this would not change if we were to run the chain longer.

A test was undertaken on the MB estimated combination of 7 examined loci and family size 3, having 30 nests in total. Two MCMC's were run, the first run was close to the upper limit of RAM capacity (2GB RAM on Windows Vista) by running 100,000 iterations without thinning hence also 100,000 effective iterations. The second run also used 100,000 MCMC iterations but was run with Thinning = 20 (recording every 20^{th} iteration) and hence 5,000 effective iterations.

We look at the variance of the estimate $\hat{\lambda}_{indep}$, as assessed at different running length of the Markov chains. The dotted line shows the TRUE variance, obtained from knowledge of the actual maternal structure in the sample.

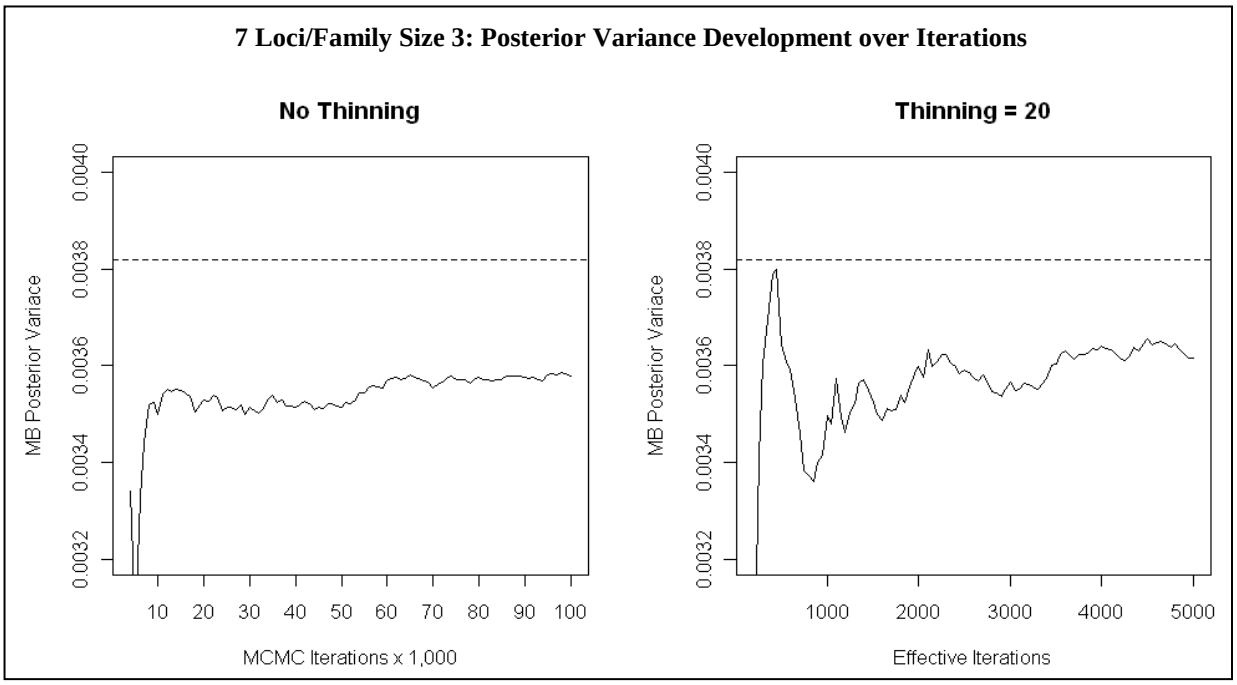

Figure3.1.4.1 - 7 Loci/Family Size 3: Posterior Variance Development over Iterations

Neither of both runs looks perfectly converged, both appear to show a slight upwards trend. Thus it appears that also the burn-in did not deliver full convergence.

The un-thinned run appears to get steady (in the sense of calming down and from there on still showing an upwards trend) at around 10,000 iterations.

Steadiness (again in the sense of calming, yet not converging) happens for the thinned MCMC at about 2,000 effective iterations (40,000 MCMC iterations).

The thinned run uses comparatively little memory, whereas the un-thinned run is close to the limit the computer can handle. Additionally the thinned run provides a steady (yet not converged) chain at less than a quarter of effective iterations than the un-thinned run provides (keeping in mind however that the MCMC of both approaches are equally long).

Though neither MCMC converges, we receive from the thinned MCMC a similar picture as from the un-thinned MCMC, however using far less memory.

Both runs do underestimate the TRUE variance, which is however expected, since MB is the independent approach that this project intends to improve. All following MCMCs will be conducted with a thinning interval of thinning = 20.

3.1.5 Importance Sampling Weights Distribution and Expected Sample Size

The importance sampling approaches W1, W2 and W3 introduce an additional source of potential quality loss by the weighting process. Large density differences in the densities of the weighting function may lead to large weights for only a few observations, which then will dominate the sample [Monahan, J. F. (2001)]. We desire approximately uniform weights.

Though the weights are supposed to be different, the case of few extremely dominating weights is undesirable since conclusions will be based on few observations, while the vast part of the MCMC is practically discarded.

Note that if all sampled maternal assignments are of low probability under the target distribution, this would lead to approximately uniform weights, simply by the fact that all assignments are equally bad. In this case, we would not have sampled the distribution we intended to sample from, the weights obtained would however appear much like the desired weights. Therefore, obtaining approximately uniform weights is not a guarantee for successful importance sampling.

Fig. 3.1.5.1 below shows an acceptable standardized W1 weights distribution, obtained from a combination of 7 loci/ family size 4 with a maternal assignment success-rate of 71.73%:

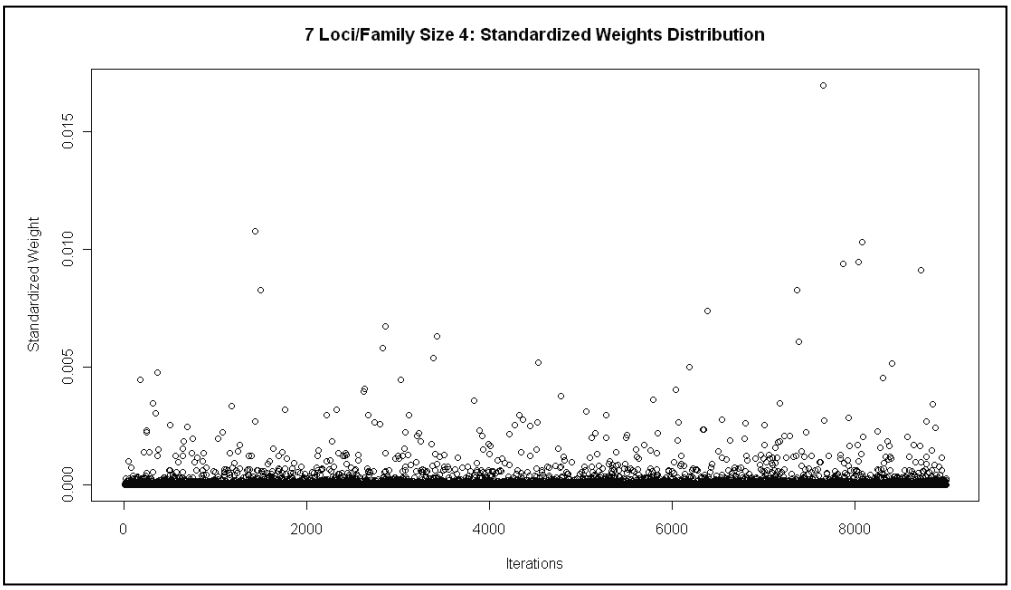

Figure 3.1.5.1 - 7 Loci/Family Size 4: Standardized Weights Distribution

The large amount of weights is crowded close to zero, those weights are near to being meaningless for the importance sampled result. There is however a cloud of larger weights, that mainly count for the result. One weight at about 7,500 iterations peaks out to about 0.018. Thus, since the weights are standardized, 1.8% of the parameter estimate is due to this particular iteration.

The occurrence of large weights is due to chance, since it depends on how well proposal and sampling density correspond, the chances do depend however on the number of examined loci and family size. The fewer the examined loci, the more we expect to find few dominating weights. Since the exclusion probability drops and hence parental assignments go worse, we expect to find greater differences between f(x) and g(x) that will lead to few observations dominating the sample.

Increasing family size increases the risk of large weights since, though every offspring has the same chance to have a mother misassigned, every additional offspring provides a further chance to introduce a false unique mother. Thus, despite the percentage of miss-assignments remains constant with increasing family size, a

greater absolute number of miss-assignments does impact on unique mothers and hence the estimated sibling structure (see chapter 3.1.3, section *Family Size Effect on Unique Mothers*).

The problem of large weights is strongly connected to the concept of expected sample size (ESS). Since importance-sampling assigns more weight to some iterations and less weight to others, the sample size cannot be regarded as the number of iterations anymore. Hesterberg (1995) suggests the ESS:

$$ESS = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} \left(w_i\right)^2} \qquad \text{Equation 3.02}$$

If for example the importance sample is dominated by few large weights, this is equivalent to having only few observations and thus the ESS will be small.

Fig. 3.1.5.2 on the following page shows the development of ESS over effective iterations in the same W1 importance sample (7 loci/ family size 4) that Fig. 3.1.5.1 above shows the weights.

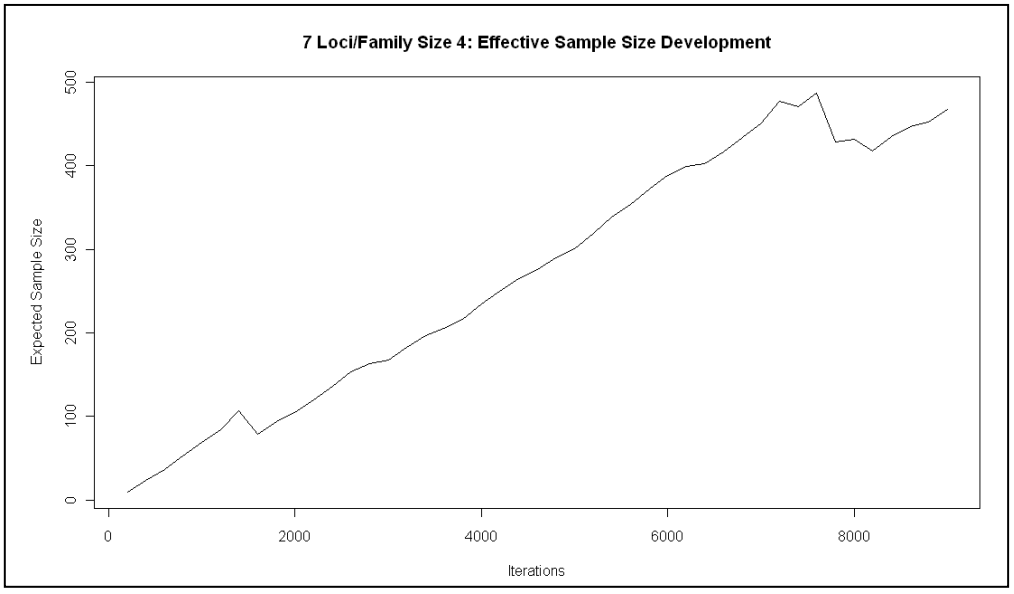

Figure 3.1.5.2 - 7 Loci/Family Size 4: Expected Sample Size Development

The ESS is steadily increasing, though not with the same rate as the iterations do, which indicates that many iterations produce small weights and thus correspond to the weights being near meaningless. There are two major breaks in this tendency, corresponding with large weights in fig. 3.1.5.1. The first at about 1,500 iterations is caused by two comparatively large weights in this region. The second, larger, break at about 7,500 iterations is caused by the largest weight, plus a couple of relatively large weights in the same area.

This illustrates nicely the effects of large weights on the ESS hence ESS gives a numeric indicator of samples containing dominating weights.

Fig. 3.1.5.3 on the next page shows W1 of the 5 loci/ family size 3 scenario, where indeed a dominating weight is present.

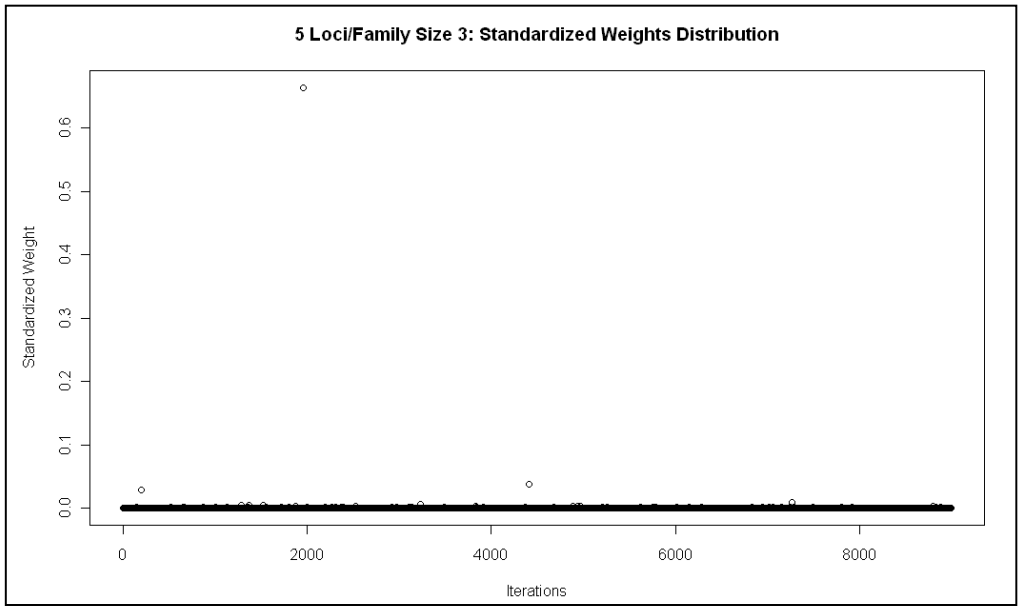

Figure 3.1.5.3 - 5 Loci/Family Size 3: Standardized Weights Distribution

At around 2,000 iterations an extremely large weight appears, accounting for more than half of the total standardized weight (≈ 0.67). This observations weights more than the sum of all remaining MCMC iterations. The ESS development over iterations shows for this scenario:

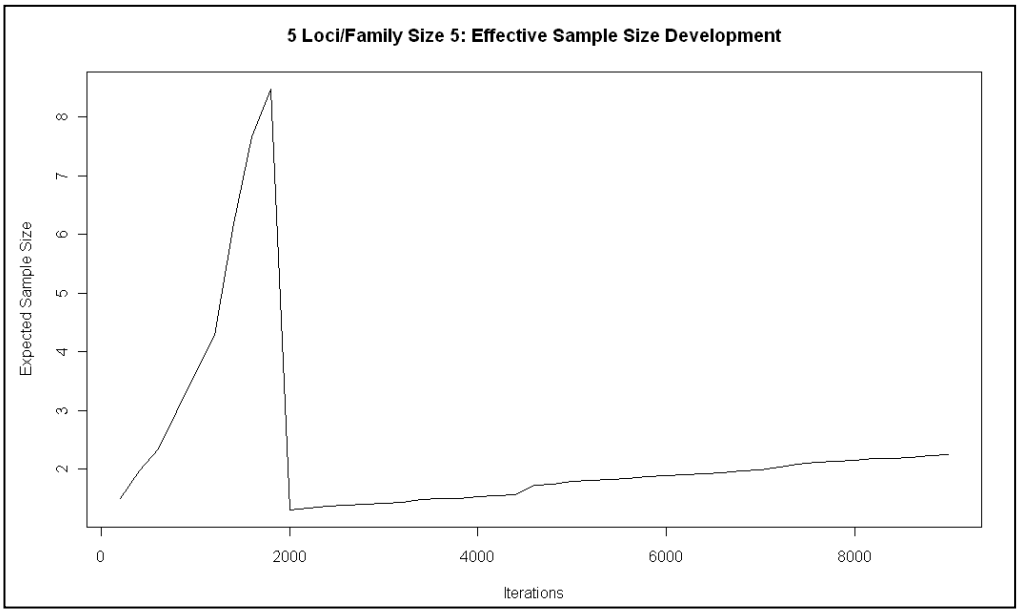

Figure 3.1.5.4 - 5 Loci/Family Size 5: Effective Sample Size Development

The ESS first increases slow but steadily, breaks down at 2,000 iterations, corresponding to the large weight observed in fig. 3.1.5.3. From there on the ESS increases very slowly, being dominated by the single large weight, finishing at the equivalent of around only two observations.

The importance sampling schemes aim to deliver large weights for “good” maternal assignments, as determined by the importance sampling schemes, and thereby increase their corresponding iterations’ influence whereas “bad” maternal assignments are down-weighted.

With an increasing number of examined loci, the maternal assignments are expected to find “good” maternal assignments more often, simply because the power to estimate the true mother (and thus exclude wrong mothers) does increase with more genetic information.

Therefore, more examined loci increase the chance of receiving large ESS. Increasing family size on the other hand increases the number of estimable, but yet false maternal combinations and thus decreases the chance of finding “good” maternal combinations. Therefore, smaller family sizes increase the chance of receiving large ESS.

The longer the MCMC is running, the more “good” combinations are expected to be found. Scenarios running comparable MCMC lengths are expected to deliver lower ESS (and hence dominating weights) with lesser examined loci and greater family size.

Equal scenarios running varying MCMC lengths are expected to deliver greater ESS (and hence more homogenous weights), the longer the MCMC runs, since more iterations with “good” maternal assignments are expected to be found.

A chain showing low ESS may have been stopped, before more “good” maternal assignments were found that caused more similar high weights. Because running time and available computer memory are limiting factors, the length of the MCMC may not be subject to free choice.

Fig. 3.1.5.5 on the next page gives the ESS of all three weighting approaches analyzed for 4, 8 and 12 examined loci versus family size (between 7.000 and 13.000 effective iterations).

In fig. 3.1.5.5 we see that the ESS changes drastically over the scenarios, depending on number of examined loci and family size. Since the ESS decreases quickly, all graphs are represented in a linear scaled plot (left), as well as in a logarithmic scaled plot (right). The fast decay with increasing number of examined loci makes it necessary to plot the linear scaled graphs for each number of loci on different scales. The logarithmic plots however allow to plot all presented scenarios on the same scale and thus are directly comparable for each number of loci.

4 examined loci show consistently low ESS between 1 and ≈ 20, W2 making an exception at family size 7, going up to ≈ 50. Only family size 3 appears to yield larger ESS for W1 (≈ 60) and W2 (≈ 95). Generally, there appears to be only a slight tendency for the ESS to decrease with increasing family size. As the ESS however cannot go below 1, greater family sizes cannot decrease the ESS considerably, since all importance-sampling schemes are already very low.

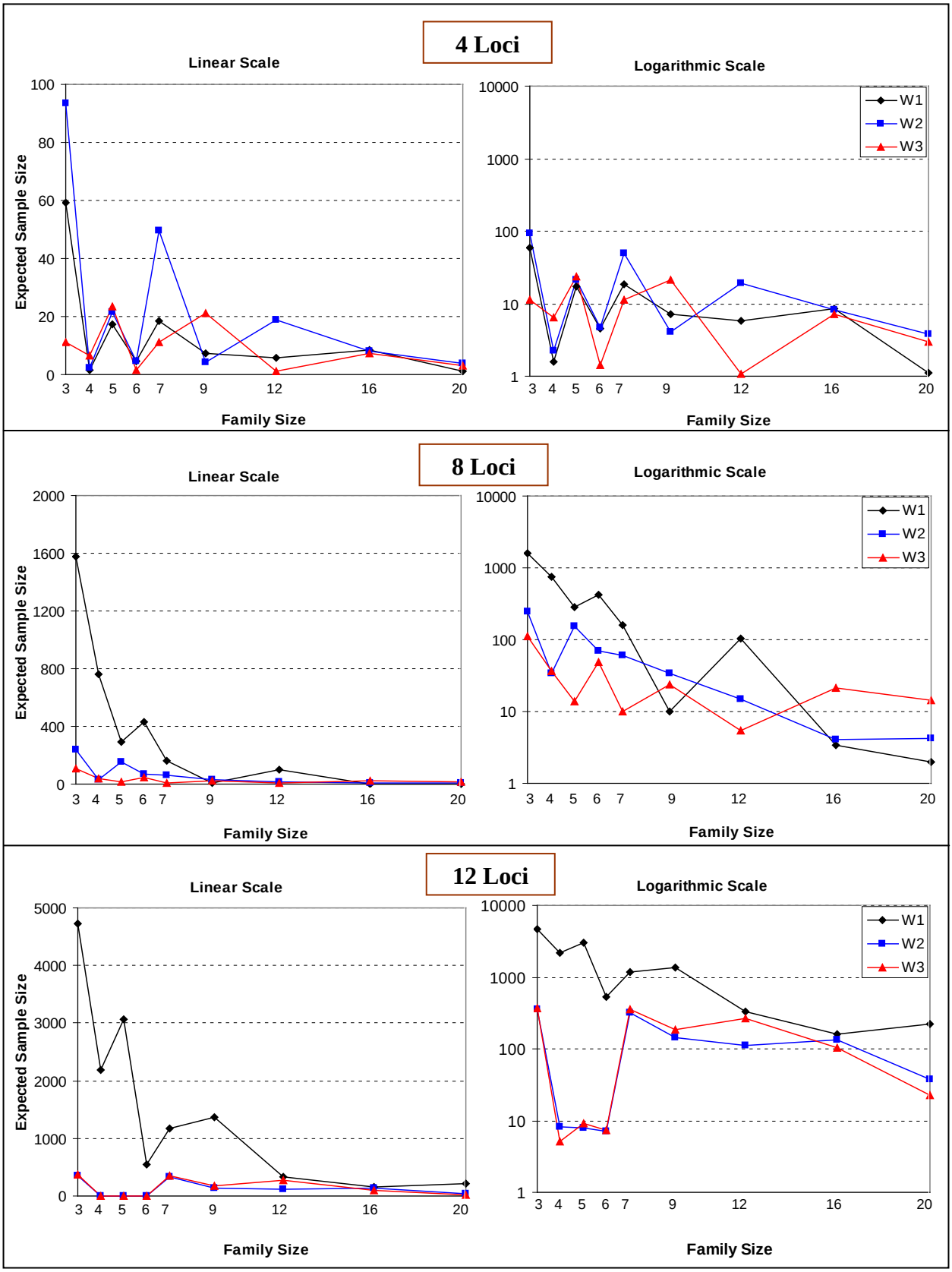

Figure 3.1.5.5 - Expected Sample Sizes

8 examined loci show decreasing ESS tendency with increasing family size. Note that the scale is now larger on the ESS linear axis. W1 shows ESS ≈ 1,600 in the family size 3 scenario and is much greater than W2 and W3. W1 however drops faster than W2 and W3, arriving at similar ESS for all importance-sampling schemes at family size 9.

12 examined loci again demand a larger scale of the linear graph, showing W1 at an ESS of nearly 5,000 in the family size 3 scenario. Also here, W1 is quickly decreasing, meeting W2 and W3 however now at family size 12. Note the unusual low ESS for W2 and W3 at 4, 5 and 6 examined loci. Here we seem to face a flux where 3 lower than expected ESS are obtained in successive family sizes.

Each plot shows the decrease of ESS by increasing family size. The logarithmic plots give a comparative picture of ESS behaviour depending on examined loci. We see that increasing number of examined loci do increase the ESS substantially. Increasing the number of examined loci also appears to allow greater family sizes, before W1 drops to comparable ESS to W2 and W3. W2 and W3 appear to behave very similar to each other, which is not surprising giving the fact that both schemes are closely related to each other.

Large ESS enlarges our trust in the results of each weighting scheme, it remains however merely an observation. Further, the ESS does not indicate any superiority of one weighting scheme over another, since for each scheme, we apply different methods and hence comparison of schemes based on ESS would be faulty.

3.2 Results

The running times of each loci/family size scenario ranged between one and three days (Windows Vista, 2.0 GHz, 2GB RAM), depending on thinning interval, number of iterations, number of examined loci and family size. Besides the running time of the MCMC itself, additional algorithms, written to extract and arrange the data to a workable format, occupied a large amount of the total running time. To evaluate the

adjustment success of all approaches, we will look separately first at the age class 2 parameter estimate's expected values and then at the variances.

3.2.1 Adjusted Expected Value

The estimation the age class 2 parameter's expected value itself is of secondary interest, since *MasterBayes*' independent assumption is not expected to yield a large location change in the parameter estimate, but rather to underestimate the variance. Fig. 3.2.1.1 on the next page shows however considerable improvement of estimating the parameter, when only little genetic information is available.

Note that the actual in each sample realized value of the age class 2 paramneter λ does vary, since each scenario is based on a new sample from the population. λ_{Pop} is the underlying age class 2 parameter in all scenarios. Since we sample from the population, each sample's λ will differ within the range of variability from λ_{Pop} (refer to chapter 2.1, section *Simulation Procedure*). TRUE gives λ that is actually realized in each sample.

Consequently, we do not look out for the actual expected values of λ that each method yields, but rather the differences of applied methods to the TRUE approach.

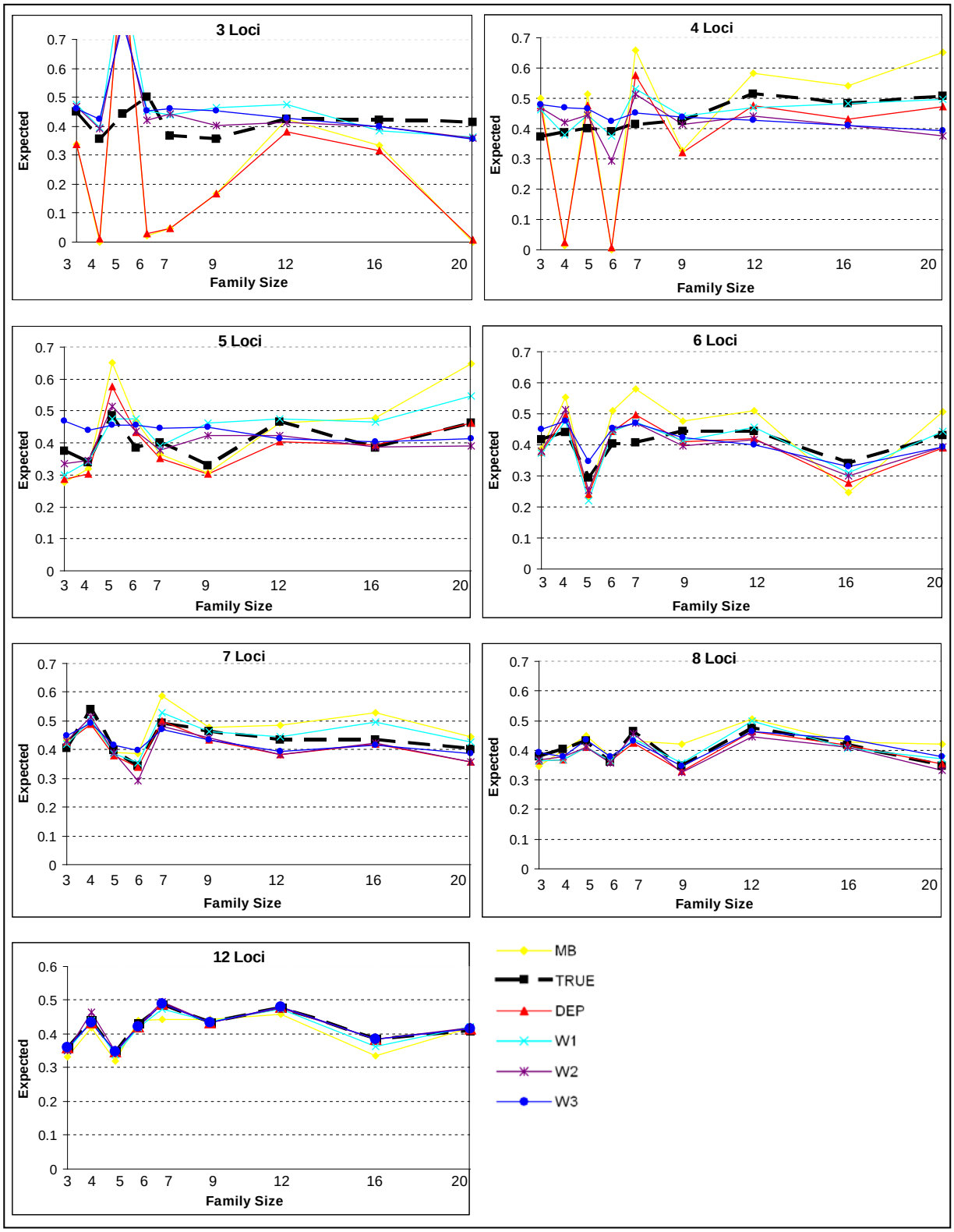

Figure 3.2.1.1 - Parameter Estimate

Clearly, increasing the number of examined loci leads to decreasing spread of the estimating schemes around the TRUE expected value. At 12 loci we achieve almost fully identical results, showing mainly MB to disagree slightly.

The most interesting results are achieved at lower numbers of loci.

At 4 loci we find that MB and DEP (hence the non importance sampling schemes) show to fail badly at family sizes 4 and 6, whereas all importance-sampling approaches appear to cope well and remain relatively close to the TRUE expected value. For greater family sizes, the expected value of MB and DEP show no clear inferiority to importance-sampling approaches. All importance sampling approaches behave similar over the whole range of tested family sizes, at low family sizes rather overestimating the expected value and at large family sizes rather underestimating.

The scenarios using 3 examined loci, and hence less genetic information, show MB and DEP behaving over the whole range of family size unreliable. Family sizes 12 and 16 come relatively close to the TRUE expected value, which may be however just happen by chance. We find MB and DEP expected values over the whole possible range of probability, from almost 0 to nearly 1. Again all importance-sampling approaches estimate the expected value comparatively well, with the exception of family size 5, where every approach fails badly to estimate the expected value.

For large numbers of examined loci all methods converge to deliver very similar results, when genetic information however does not allow for nearly full correct maternal assignment, the importance sampling schemes appear to be considerably superior to non importance sampling schemes.

Within the three available importance sampling schemes W1, W2 and W3 there appears to be none superior over another. Ideally, one would apply all available approaches, where consistency of estimates would increase the trust in the result. This however is not a failsafe method, as the combination of 3 examined loci and family size 5 shows, where estimates are somewhat consistent, but yet far off the TRUE expected value. In case of inconsistency, the proposed importance sampling schemes provide the more trustable estimate than MB or DEP.

Especially at little knowledge of the genotypes, the importance sampling schemes provide relatively good estimates of the expected value, whereas MB and DEP both badly fail.

3.2.2 Adjusted Variance

While the expected value of the age class 2 parameter generally does not suffer greatly from the assumption of independence, the parameter variance will be underestimated, as shown in chapter 1.1, section *Independent versus Unique Mothers – A Counting Example*.

By adjusting the estimates' variance for the presence of siblings, we finally approach the heart of the project. Figures 3.2.2.1 and 3.2.2.2 show the graphs of the estimated variances obtained from the all presented approaches.

Note that by assessing the variances obtained by the estimation approaches in comparison to the TRUE variance, we have to keep in mind that the TRUE approach does not reflect maternal uncertainty, since it is based on the actual mothers known from the simulation. Uncertainty in parentage thus adds to any estimated variance and thus we expect the adjustment approaches to deliver greater variance than the TRUE approach. See also chapter 2.4 for comments on this issue.

Scenarios using 3 examined loci produced extremely large variances for MB and DEP, which are not displayable on common gridlines of all loci's graphs. Thus, the following fig. 3.2.2.1 shows the parameter variance plot of 3 loci on a separate grid line spacing than fig. 3.2.2.2, the grid lines themselves however are corresponding to the same variance in all graphs.

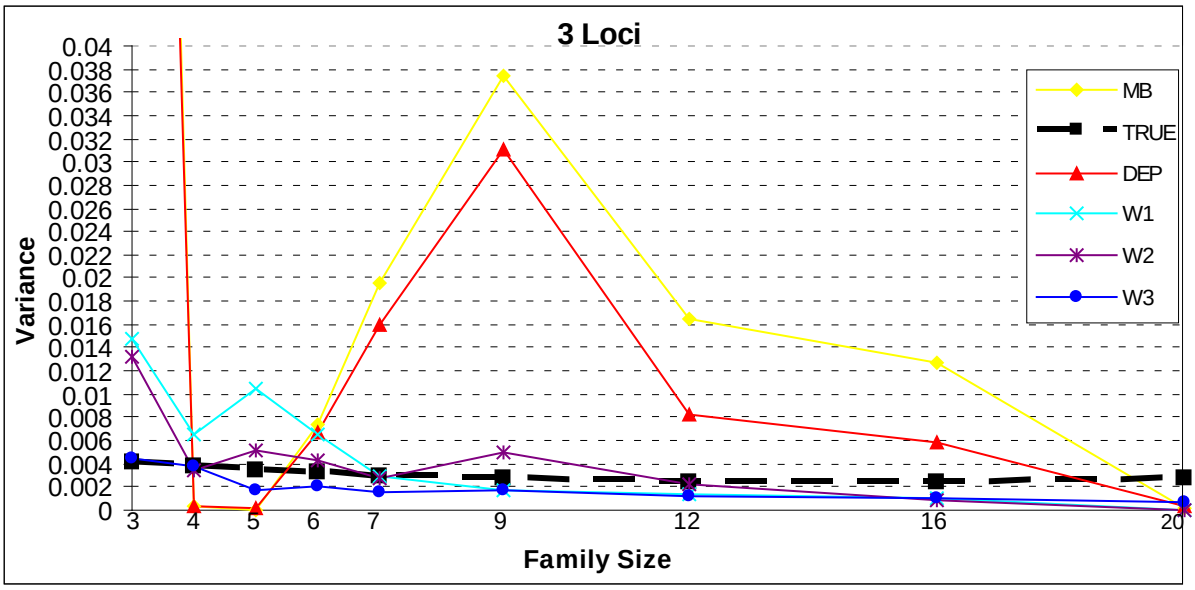

Figure 3.2.2.1 - Parameter Variance of 3 Examined Loci

With 3 examined loci and hence little information of maternity structure (maternal assignment success is at around 5%), the original MB approach as well as the DEP approach show to be extremely unreliable. There appears to be no systematic bias, the variance might be under- or overestimated. Both, MB and DEP, do however roughly correspond with each other.

The importance sampling approaches on the other hand deliver comparatively good estimated variances, considering the extreme uncertainty of MB and DEP. W1 overestimates considerably for small family sizes, but comes close to the TRUE variance at 7 loci. W2 only shows similar overestimation at family size 3, but then approaches TRUE. W3 goes well along with TRUE for low family sizes 3 and 4, but then drops below TRUE at family size 5.

All approaches appear to yield smaller variances for increasing family sizes. See the following fig. 3.2.2.2 for a “magnified” plot of the 3 examined loci scenario.

The estimated variances of all scenarios (including the 3 loci scenario, which however appears incomplete when plotting in common grid line spacing) are plotted in fig. 3.2.2.2 on the next page in comparable scale. Note however that the grid lines in fig. 3.2.2.2 correspond with the grid lines in fig. 3.2.2.1 above.

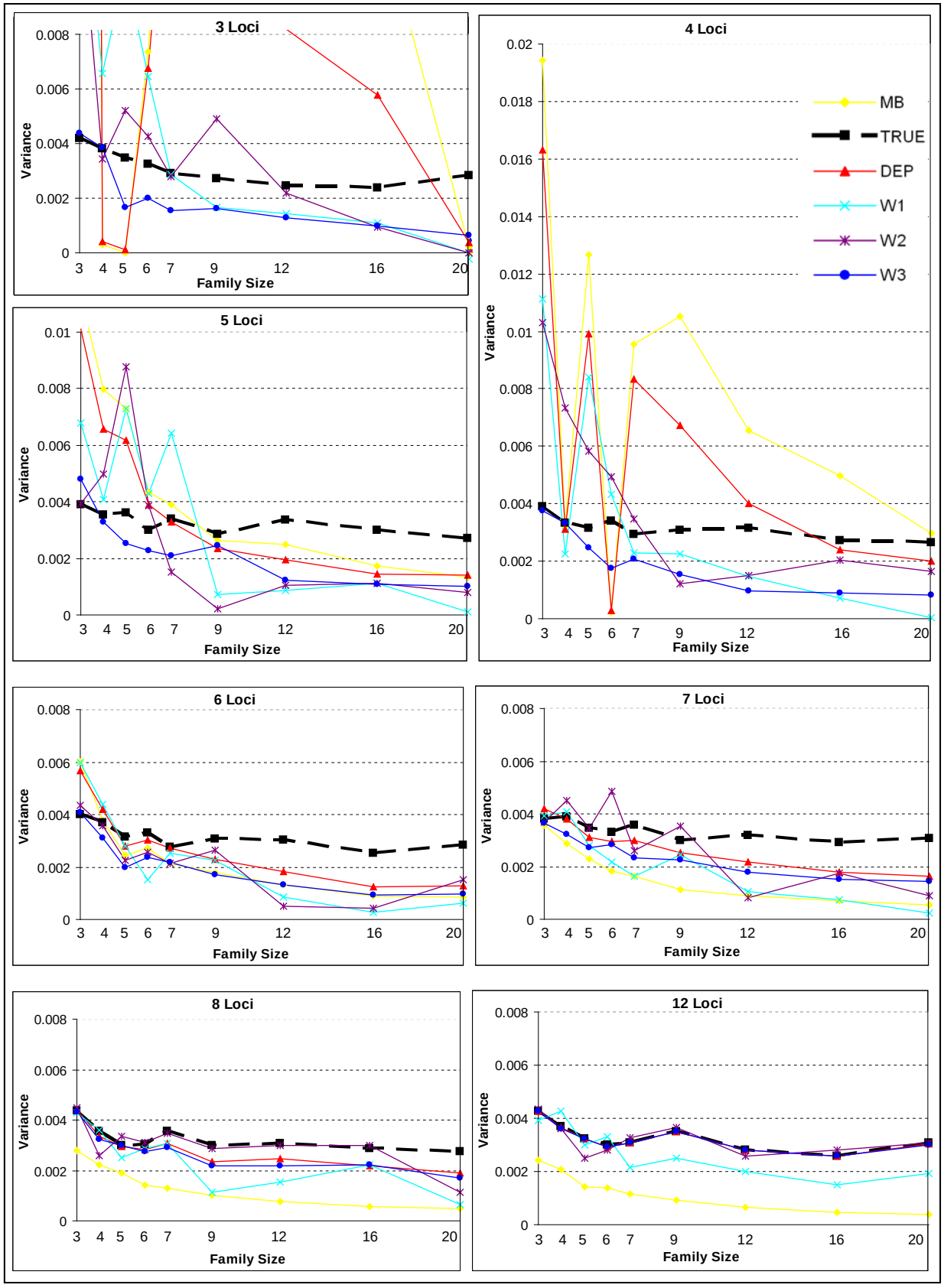

Figure 3.2.2.2 - Parameter Variance

As the number of examined loci, and hence the accuracy of maternal assignments, increases, the spread of all approaches' estimated parameter variances narrows down, in case of DEP, W2 and W3 to converge to the TRUE approach, in case of MB and W1 however underestimating TRUE. At low numbers of examined loci the importance sampling schemes provide parameter estimate variances notably closer to TRUE than, than MB and DEP.

Increasing family size on the other hand, results in decreasing estimated parameter variances, leading to greater underestimation, the larger the family is. Though the maternal assignment success-rates remain about constant for a given number of loci, bigger family size results in more misassigned mothers and hence the introduction of false unique mothers (as shown in chapter 3.1.3, section *Family Size Effect on Unique Mothers*). This causes the variance to decrease for larger family sizes by the same process that causes the MB estimates to underestimate the posterior variance (as shown in chapter 1.1, section *Independent versus Unique Mothers – A Counting Example*).

From 6 loci on to greater numbers of examined loci, the underestimation of the parameter variance is comparatively small, for family sizes up to 6. For greater family sizes, the underestimation becomes then increasingly severe. Increasing numbers of examined loci however reduces the underestimation for larger family sizes.

The plot of 12 loci illustrates nicely the underestimation of posterior variance of the MB approach that we intend to correct. With nearly full maternal assignment success and hence only little uncertainty about parentage, all adjustments besides W1 are able to identify (almost) all unique mothers and thus the adjustment procedures provide an adequate variance. The original *MasterBayes* MB approach does underestimate the TRUE parameter variance. Also here MB shows for increasing family sizes increasing differences between the parameter variance and TRUE.

At 12 examined loci we further note that the importance sampling scheme W1 does family size 7 on to larger family sizes underestimate the parameter variance, though being closer to TRUE than MB. W2 shows only at family size 5 a departure from TRUE, but otherwise follows TRUE well. DEP and W3 prove to estimate the posterior

variance excellent over the whole range of examined family sizes at 12 loci. Also approach W2 provides satisfying results, W1 however fails to adjust MB well.

Including fewer loci to an analysis increases the uncertainty about the estimated maternal structure. The development of estimation accuracy between high maternal assignment uncertainty and nearly full knowledge of the maternal structure, which we approach at 12 loci scenarios, is the most interesting part, since any estimation of population parameters could be done with less effort by classical methods, when the pedigree is known (eg. at 12 loci).

The approach that seems to be least influenced by the assignment success and hence knowledge of maternal assignments appears to be the importance sampling scheme W3. It provides the most consistent parameter variance estimates over the whole range of examined number of loci/family size scenarios.

W3 approaches TRUE with increasing certainty of maternal assignments, for greater uncertainty it tends to underestimate the TRUE variance. Greater uncertainty however should be reflected by increased estimate variance, which W3 fails to capture, instead showing smaller variance, the greater uncertainty becomes.

The low ESS (see chapter 3.1.5), not only for W3, but also for W1 and W2, suggests that pedigree configurations that may have large genetic likelihood (since exclusion power is low, many females that are not related to offspring may yet show considerably large likelihood) are ruled out by down-weighting, whereas MB and DEP treat all visited configurations equally. We expect more false unique mothers by examining few loci (simply by false assignment), which causes the variance to be underestimated (see chapter 1.1, section *Independent versus Unique Mothers – A Counting Example*). Uncertainty in pedigree on the other hand causes the variance to increase by spreading the parameter estimates over iterations (since bad assignments allow worse proposals to be accepted). The importance sampling however shows low ESS and thus only few important iterations that count, having most iterations ruled out and thereby largely eliminated the mechanism that increases the variance by uncertainty of pedigree. Therefore, the importance sampling approaches do not properly reflect the uncertainty in pedigree when ESS is low,

incorporating the mechanism that decreases the variance, but eliminating the mechanism that increases variance.

The greater the uncertainty of pedigree, the more W3 does underestimate the posterior variance, by suppressing the process that increases variance (as indicated by decreasing ESS) and supporting the process that decreases variance. It is difficult to tell however, how much the variance should be overestimated due to maternal assignment uncertainty.

Though other approaches do occasionally approach TRUE better than W3, they fail in showing the same consistence over tested scenarios. Below 6 loci, all approaches besides W3 appear to yield large parameter variances at low family sizes and quickly decreasing variances when increasing the family size, which they however do not deliver in a consistent manner.

The importance sampling approaches W1 and W2 are less affected by this than MB and DEP. W2 is affected at high uncertainty of parentage at small family sizes. It shows less consistency than W3, sometimes estimating larger and sometimes smaller posterior variance. It however converges nearly fully at 12 loci.

W1 appears to be the most unreliable importance sampling approach, yet delivering relatively good results at low numbers of loci, compared to MB and DEP. At nearly full knowledge of pedigree, W1 however still severely underestimates the posterior variance.

From 6 examined loci on to greater numbers of loci, DEP becomes the superior parameter variance estimation approach, still like W3 underestimating TRUE, but being closer to TRUE than W3 is. One has to keep in mind however, that DEP proved to be extremely unreliable at smaller numbers of loci, where W3, though underestimating, still showed relatively good consistency. The examined number of loci and hence the maternal assignment success determines when DEP becomes superior to W3.

Though we cannot determine the assignment success-rates without knowledge of the true pedigree, the naïve assignment success- rates (equation 2.03), presented in

chapter 2.1 and compared to actual maternal success-rates in fig. 3.1.3.1, provide an by-eye estimate, what maternal success-rates we can expect to achieve in the MCMC, that can be calculated from in real life known values.

Since 6 loci provide a naïve maternal assignment success-rate of 0.45, it appears feasible to generally expect the DEP approach to become superior to W3 at about this success-rate. This however assumes ideal conditions, such as no typing error, no cuckolding fathers, zero mutation rate, no unsampled parents in the sample, etc.

It is recommendable to run a series of simulations, designed on the studied population, to determine the behaviour of all approaches under similar conditions that the actual research provides and from there infer, at what maternal naïve success-rate DEP become superior to W3.

4. Conclusions and Recommendations

Three of the presented approaches to adjust *MasterBayes*' parameter estimate's variance to the presence of siblings only succeed satisfyingly, when the pedigree is nearly perfectly estimated. Those approaches are DEP, W2 and W3. Increasing uncertainty of the pedigree still yields acceptable adjustments to the parameter variances for medium pedigree certainty at low family sizes.

All presented approaches however do improve the parameter's variance as estimated by the original *MasterBayes* approach.

While uncertainty in pedigree increases the estimated variance (which is a "justified effect", representing this uncertainty), increasing family sizes does cause the variance to underestimated (introduced by false unique mothers, which is a "unjustified effect", since it does not represent improved knowledge of the pedigree or the parameter variance). This puts limits to the number of siblings (and hence sample sizes that can be taken from a nest) that the presented approaches can reasonably adjust for, depending on the desired quality of the estimated variances. Consequently, the presented approaches adjust less well for the presence of siblings the more siblings are present, resulting in an underestimation of the parameter variance.

At great uncertainty of the pedigree, importance sampling schemes perform more reliable than DEP. Especially W3 provides consistent parameter variance estimates, that however do underestimate the true posterior variance the more, the greater the uncertainty of pedigree (whereas greater uncertainty in parentage should ideally be reflected by increased variance).

When the pedigree is estimated sufficiently for DEP to deliver consistent parameter estimate variances, DEP proves to be the superior scheme to adjust the variance for the presence of siblings. At nearly perfect pedigree assignment, W2 and W3 become similar successful when approaching total maternal assignment success. The naïve assignment success-rate provides a by-eye method to decide whether W3 or DEP is the best parameter variance estimator.

No approach however satisfyingly adjusts the parameter estimate variance at less than perfect maternal assignments, either failing in delivering consistent results, or in failing to pay respect to the uncertainty of pedigree by underestimating the parameter's variance, instead of overestimating.

Though the adjustment success is less than fully satisfactory, all presented approaches show considerable improvements to *MasterBayes* original parameter estimate's variance.

Beyond the scope of adjusting the estimate's variance to the presence of siblings, we achieve considerable improvements to *MasterBayes*' parameter estimate, when uncertainty of the pedigree is large and hence knowledge of genotypes and therefore pedigree is little. Here all introduced importance-sampling procedures deliver expected values for the parameter estimates close to the true sample parameter, whereas non-importance sampling approaches MB and DEP fail by delivering expected values that span over nearly the full range of probability.

The importance sampling schemes provide a more reliable method for obtaining the expected parameter estimates when the pedigree is uncertain, than *MasterBayes* itself can.

5. References

B. Atkinson (atkinson@mayo.edu) for pedigree functions. T. Therneau (therenau@mayo.edu) for all other functions. (2008). kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. R package version 1.1.0-22.

A. M. Emery, I. J. Wilson, S. Craig, P. R. Boyle, L. R. Noble. (2000). Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. Molecular ecology (2001) 10, 1265-1278.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin. (2004). Bayesian data analysis (Ed. 2). Chapman & Hall/CRC

S. Gerber, S. Mariette, R. Streiff, C. Bodenes, A. Kremer. (2000). Comparison of microsatellites and AFLP markers for parentage analysis. Mol Ecol 9: 1037-1048.

J. C. Faria. (2008). Resources of Tinn-R GUI/editor for R environment. UESC, Ilheus, Brasil.

S. Geman, D. Geman. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE transactions of pattern analysis and machine intelligence, 6:721–741.

J. D Hadfield, D.S. Richardson, T. Burke. (2006). Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. Molecular ecology (2006) 15 ,3715 –3730

J. D. Hadfield. (2008a). MasterBayes: Maximum likelihood and markov chain monte carlo methods for pedigree reconstruction, analysis and simulation.

J. D. Hadfield. (2008b). MasterBayes: ML and MCMC Methods for Pedigree Reconstruction and Analysis. R package version 2.3.

D. L. Hartl, A. G. Clark. (1997). Principles of population genetics (Ed. 3). Sinauer Associates, Inc.

W. K. Hastings. (1970). Monte carlo sampling methods using Markov chains and their applications. Biometrika 57: 97-109

T. Hesterberg. (1995). Weighted average importance sampling and defensive mixture distributions. Technometrics, 37(2): 185-195.

W. Hörmann, J. Leydold. (2005). Monte carlo integration using importance sampling and gibbs sampling. Austrian science foundation (FWF), project no. P16767-N12.

A. Jamieson, S. S. Taylor. (1997). Comparison of three probability formulae for parentage exclusion. Animal genetics 28: 397-400

B. Jones, G. D. Grossman, D. C. I. Walsh, B. A. Porter, J. C. Avise, A. C. Fiumera. (2007). Estimating differential reproductive success from nests of related individuals, with application to a study of the mottled sculpin, cottus bairdi. Genetics 176: 2427-2439 (August 2007)

S. M. Lynch. (2007). Introduction to applied Bayesian statistics and estimation for social scientists. New York: Springer.

T. C. Marshall, J. Slate, L. E. B. Kruuk, J. M. Pemberton. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. Molecular ecology (1998) 5, 639-655.

X. L. Meng. (1993). Coherent multiple-imputation inference under incoherent models. Technical report 359, Dept. statistics, Univ. Chigao.

X. L. Meng. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical science, vol. 9, no. 4, pp. 538-558

J. F. Monahan. (2001), Numerical methods of statistics. Cambridge university press.

M. Plummer, N. Best, K. Cowles & K. Vines. (2008). coda: Output analysis and diagnostics for MCMC. R package version 0.13-2.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

D. B. Rubin. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, Vol. 91, No. 434. (Jun., 1996), pp. 473-489.

G. Warnes, with contributions from G. Gorjanc, F. Leisch and M. Man. (2008). Genetics: population penetics. R package version 1.3.3.

G. Warnes. (2008). Includes R source code and/or documentation contributed by B. Bolker and T. Lumley (2008). gtools: Various R programming tools. R package version 2.5.0.

I. Wilson. (2001). Parentage version 1.0 – users guide. http://www.mas.ncl.ac.uk/~nijw/parentage