A Bootstrap Assessment of Variability in Pedigree Reconstruction Based on DNA Markers

Anthony Almudevar*

January 18, 2000

Abstract

The problem of assessing the variability in pedigree reconstruction using DNA markers is considered for the special case of single generation samples with no parents present in the sample. Error in pedigree reconstruction is measured through a metric imposed on the space of partitions of the individuals into family groups. A confidence set can therefore be taken to be a neighbourhood of a point estimate, analogous to the estimation of a parameter in Euclidean space. The coverage probability is estimated using bootstrap techniques. Although the distributional properties of the sample depend on the population genotype frequencies, these are in practice usually unknown. Confidence sets conditioned on a statistic approximately sufficient for these frequencies are compared to confidence sets obtained by substituting frequency estimates directly into the sampling distribution. In two simulation studies the difference is found to be of some consequence.

*Department of Mathematics and Computing Science, St. Mary's University, Halifax, N. S., Canada, B3H 3C3

Key words: pedigree reconstruction, bootstrapping, DNA markers

1 Introduction

The reconstruction of unknown pedigrees of wild sampled individuals based on DNA marker data is an important technique in population biology. When the sampled pedigree is complete, in the sense that with the exception of (usually unrelated) founders the parents of all individuals are contained in the sample, and population genotype frequencies are known, a well defined likelihood can be constructed for any given pedigree. (Thompson 1976, Meagher and Thompson 1986).

Typically, pedigree estimation must proceed in the absence of some of the information required for the complete specification of the likelihood given in Thompson (1976). In this article we look specifically at the case in which all individuals are known to be of a single generation, so that there are no putative parents in the sample, and there are no reliable estimates of population genotype frequencies. Such a situation arises, for example, in larvae samples in fisheries research when interest is in establishing the degree of relatedness in the sample (Herbinger *et al.* (1997)). In practice, when individuals are known to be of a single generation, inference reduces to a specification of all mutual full sibling and half sibling relationships. Inference concerning more distant relationships is problematic (Brookfield and Parkin (1993)), and will not be considered here.

In this article the problem of assessing the variational error in this special case of pedigree reconstruction is considered. When the analysis is primarily pairwise, in the sense that relationships are infered separately for each pair, as in Blouin *et al.* (1996) or Herbinger *et al.* (1997), then the error assessment can also proceed on a pairwise basis. One option is to use a Bonferroni type correction for multiple comparisons (Herbinger *et al.* (1997), Rice

(1989)). Another approach is that of Painter (1996), in which a likelihood over the space of admissible pedigrees of 9 single generation falcon samples is constructed, and a posterior probability is assigned to the maximum likelihood pedigree.

We propose in this article an alternative procedure for assessing the accuracy of a single generation pedigree estimate, which will be based on a bootstrap scheme. The procedure can be applied to any type of pedigree estimator, as long as it does not construct pedigree estimates which are genetically inconsistent with the observed DNA markers.

To introduce the procedure we note that a single generation pedigree first partitions the individuals into full sibling groups. Half-sibling structure may subsequently be imposed on the partition, but the full sibling partition can always be unambiguously defined. We will concentrate specifically on the estimation of this full sibling partition. The object will be to construct a confidence set by first defining a metric on the space of partitions, the confidence set then taking the form of a neighbourhood generated by this metric centered at some initial estimate of the partition. This results in an estimate of how close the estimated partition is likely to be to the true partition. In effect, we treat the problem of estimating the true partition in a manner analogous to that of estimating a parameter when it is a point in Euclidean space.

The procedure proposed here may be used when population genotype or allele frequencies are unknown, which in practice is often the case. These frequencies can then be considered nuisance parameters. Two methods for dealing with the nuisance parameters are examined. In the first, a statistic which is (approximately) sufficient for the frequencies is be constructed and conditional confidence sets derived. This is then contrasted with confidence sets obtained by substituting estimates of the frequencies directly into the sampling distribution.

In Section 2 some of the issues concerning the construction of a metric on the space of partitions is discussed. In Section 3 we discuss how confidence sets can be constructed and

	U in A_2	R in A_2
U in A_1	8	0
R in A_1	1	6

	U in A_3	R in A_3
U in A_1	8	0
R in A_1	3	4

Table 1: Classification of pairs in $I = \{1, 2, 3, 4, 5, 6\}$ by (R)elatedness or (U)nrelatedness in two partitions, (A_1, A_2) and (A_1, A_3).

confidence levels estimated via a bootstrap procedure. In Section 4 two simulation studies are presented.

2 Distance between partitions

The techniques developed here depend on some quantifiable notion of error, or distance between the true partition and the estimated partition. One possibility is to base the distance on the number of pairwise relationships correctly identified, which would be natural for a pairwise analysis. This, however, suffers from a lack of clear interpretability. To see this consider three partitions of $I = \{1, 2, 3, 4, 5, 6\}$, $A_1 = (1, 2)(3, 4, 5, 6), A_2 = (1)(2)(3, 4, 5, 6), A_3 = (1, 2)(3, 4, 5)(6)$. For any two partitions A_i, A_j we can construct a 2×2 table classifying all pairs in I according to whether or not they are in the same set (that is, related) in A_i and A_j, as in Table 1.

Suppose A_1 is the correct partition and A_2 and A_3 are estimates of A_1. If we define the distance D_p between two partitions as the number of pairs classified differently, we then have $D_p(A_1, A_2) = 1$, but $D_p(A_1, A_3) = 3$. Note that both A_2 and A_3 can be obtained from A_1 by removing a member of I from one of the groups and placing it in a group of it's own (2 and 6 respectively). Nonetheless, the distances are different, being a indication of the size of the group from which the individual was removed. Unless there is an *a priori* reason to consider the removal of 6 to be a more consequential error than the removal of 2, this distance will lack clear meaning. (This distance, being reducible to counting measure imposed on symmetric differences of finite sets, is a true metric).

On the other hand, we may state that A_2 and A_3, considered as estimates of A_1, each have one error. A distance may then be defined as the minimum number of elements of I which must be removed in order to leave the remaining partitions equal. By removing 2 from A_1 and A_2, the remaining partitions are left equal, so $D_r(A_1, A_2) = 1$, and similarly $D_r(A_1, A_3) = 1$. Thus, this distance is more plainly interpretable as the number of errors made in the estimate of a partition.

We can show that D_r is a metric. It follows from the definition that $D_r(A, A) = 0$ and that $D_r(A, B) = D_r(B, A)$. It remains to show that the triangle inequality holds. Consider partitions A, B, C of a finite set I. There is a set E_{AC} of cardinality $D_r(A, C)$ which leaves A and C equal when E_{AC} is removed. Similarly, there is a set E_{BC} of cardinality $D_r(B, C)$ which leaves B and C equal when E_{AC} is removed. Consequently, when $E_{AC} \cup E_{BC}$ is removed from A, B and C, these partitions are left equal, hence

$$\begin{aligned} D_r(A, B) &\leq \operatorname{card}(E_{AC} \cup E_{BC}) \\ &\leq D_r(A, C) + D_r(B, C). \end{aligned}$$

It turns out that D_r is equivalent to another naturally defined distance (proposed in, for example, Painter (1994)). A *step* on a partition is defined to be the movement of an element of I from one group to another group (including to a newly formed group). Let $D_s(A, B)$ be the minimum number of steps needed to reach partition B from A. We now show that $D_s = D_r$. If E is a subset of I which when removed makes A and B equal, then we may also construct a sequence of steps, one for each element of E which reaches B from A, hence $D_s(A,B) \leq D_r(A,B)$. On the other hand, if the elements of I involved in a minimum number of steps from A to B form set E, then removing E from A and B leave the partitions equal, hence $D_r(A,B) \leq D_s(A,B)$.

An algorithm to compute D_r efficiently appears in Almudevar and Field (1999).

3 The bootstrap procedure

A "naive" bootstrap procedure will attempt to recreate as accurately as possible the process under which offspring genotypes are generated. First, parental genotypes are resampled, then the rules of Mendelian inheritence can be simulated to produce offspring genotypes. Pedigrees may be reconstructed from these offspring genotypes. If the process is repeated many times, the variability of the pedigree estimates can be examined. Hopefully, this variation is approximately equivalent to that obtainable under actual sampling conditions. We assume throughout that the offspring genotypes are the only observable data.

We will assume that there is a finite population of individuals labelled $\mathcal{I} = \{1, \ldots, M_{\mathcal{I}}\}$. Let $\mathcal{G}$ denote the set of genotypes in the population at a single locus under consideration (the methodology is easily extended to multiple unlinked loci). Let $\mathcal{A}$ denote the set of alleles in the population. In addition let $\mathcal{G}^k$ and $\mathcal{A}^k$ denote the k dimensional cartesian product spaces. For any set of individuals $I \subset \mathcal{I}$ we will define a *pedigree configuration* b to be an positive integer M_b, along with M_b subsets of I, $b_1, \ldots, b_{M_b}$, where M_b is the number of parents needed to completely specify the pedigree, and b_i represents all offspring in I of the ith parent. Let the set of all pedigree configurations of I be $\mathcal{B}_I$.

At a fixed locus the distribution of the genotypes of a sibling group conditional on the parental genotypes g_1, g_2 is independent of population frequencies, and dependent on g_1, g_2

through a straightforward application of the rules of Mendelian inheritance. For a single offspring, denote this distribution $Q(\cdot; g_1, g_2)$. Then conditional on parental genotypes g_1, g_2, the joint distribution of the genotypes in a sibling group is equivalent to an *i.i.d.* sample from $Q(\cdot; g_1, g_2)$.

Let I denote a sample of individuals selected in some fashion from $\mathcal{I}$. Assume it contains M_I individuals. Then let $B = (B_1, \ldots, B_{M_B}; M_B) \in \mathcal{B}_I$ be the true pedigree configuration of I. For convenience, assume that the individuals in I have been relabelled $I = \{1, \ldots, M_I\}$. Let $X_k \in \mathcal{G}$ be the genotype of the kth individual in I, and let $Y_j \in \mathcal{G}$ be the genotype of parent $j \in \{1, \ldots M_B\}$. Set $X = (X_1, \ldots, X_{M_I})$ and $Y = (Y_1, \ldots, Y_{M_B})$. Then, assuming that distinct parents of offspring in I are unrelated we have, for any $b = (b_1, \ldots, b_{M_b}; M_b) \in \mathcal{B}_I$ and

$$
\begin{aligned}
& P\{X = (x_1, \ldots, x_{M_I}) \mid B = b\} \\
&= \sum_{g_1, \ldots, g_{M_b} \in \mathcal{G}^{M_b}} P\{X = (x_1, \ldots, x_{M_I}) \mid Y = (g_1, \ldots, g_{M_b}), B = b\} \\
&\qquad \times P\{Y = (g_1, \ldots, g_{M_B}) \mid B = b\} \\
&= \sum_{g_1, \ldots, g_{M_b} \in \mathcal{G}^{M_b}} \left(\prod_{1 \le i < j \le M_b} \prod_{k \in b_i \cap b_j} Q(x_k; g_i, g_j) \right) \prod_{i=1}^{M_b} P_{g_i} \qquad (3.1)
\end{aligned}
$$

where P_g is the population frequency of genotype g and $x_i \in \mathcal{G}$.

The distribution of X ultimately depends on population allele or genotype frequencies. If these are known, they can be incorporated directly into calculations involving (3.1). In practice, they are often not known, in which case we present two approaches. The first approach is to substitute estimates of these frequencies into (3.1), although a potential problem is that the amount of information about the population frequencies depends solely on the number of parents, which may be much smaller than the number of offspring sampled. Hence, there is an unavoidable limit in the accuracy of frequency estimates, independent of the quality of the partition estimate. We use the term *unconditional* to refer to this approach.

A second approach is to define a statistic T which is sufficient for these frequencies, so that the distribution of X conditional on T does not depend on the population frequencies. If such a statistic T is unobservable then, as an approximation, we may substitute an estimate of T. Ideally, such an estimator would depend on the estimated pedigree, and would be accurate if the pedigree estimate itself was accurate. We use the term *conditional* to refer to this apporach, which is developed below.

To construct T, let N_g be the number of occurrences of genotype g among all parents of individuals in I, and let N^G denote the multinomial vector of genotype frequencies with probabilities P_g for $g \in \mathcal{G}$. Conditioning on N^G gives

$$\begin{aligned} &P\left\{X=(x_1,\ldots,x_{M_I}) \mid N^G=n, B=b\right\} \\ &= \sum_{g_1,\ldots,g_{M_b}\in\mathcal{G}^{M_b}} P\left\{X=(x_1,\ldots,x_{M_I}) \mid Y=(g_1,\ldots,g_{M_b}), N^G=n, B=b\right\} \\ &\quad \times P\left\{Y=(g_1,\ldots,g_{M_b}) \mid N^G=n, B=b\right\}. \end{aligned} \tag{3.2}$$

By convention we set terms in the above summation to 0 if

$$P\left\{Y=(g_1,\ldots,g_{M_b}) \mid N^G=n, B=b\right\}=0.$$

If this probability is not 0 then we must have

$$\{Y=(g_1,\ldots,g_{M_b})\} \subset \{N^G=n\}$$

so that

$$\begin{aligned} &P\left\{X=(x_1,\ldots,x_{M_I}) \mid N^G=n, B=b\right\} \\ &= \sum_{g_1,\ldots,g_{M_b}\in\mathcal{G}^{M_b}} P\left\{X=(x_1,\ldots,x_{M_I}) \mid Y=(g_1,\ldots,g_{M_b}), B=b\right\} \\ &\quad \times P\left\{Y=(g_1,\ldots,g_{M_b}) \mid N^G=n, B=b\right\}. \end{aligned}$$

As in equation (3.1) we have

$$P\{X = (x_1, \ldots, x_{M_I}) \mid Y = (g_1, \ldots, g_{M_b}), B = b\} = \prod_{1 \le i < j \le M_b} \prod_{k \in b_i \cap b_j} Q(x_k; g_i, g_j),$$

which does not depend on the population frequencies. Then

$$\begin{aligned} P\left\{Y = (g_1, \ldots, g_{M_b}) \mid N^G = n, B = b\right\} &= \frac{\prod_{i=1}^{M_b} P_{g_i}}{\frac{M_b!}{\prod_{g \in \mathcal{G}} n_g!} \prod_{g \in \mathcal{G}} P_g^{n_g}} \\ &= \left(\frac{M_b!}{\prod_{g \in \mathcal{G}} n_g!}\right)^{-1} \end{aligned}$$

unless the event has probability 0, so that X conditional on (n, b) does not depend on the population frequencies.

Another choice for the sufficient statistic T can be made. Suppose N_a is the number of times allele a appears among the parents. Let N^A be the vector of all such frequencies. Then under Hardy-Weinberg equilibrium N^A is a multinomial vector with multinomial probabilities P_a for all $a \in \mathcal{A}$. If a_{i1}, a_{i2} are the alleles which make up genotype g_i then, letting n_h be the number of heterozygotes among the parents we have

$$\begin{aligned} P\left\{Y = (g_1, \ldots, g_{M_B}) \mid N^A = n, B = b\right\} &= 2^{n_h} \frac{\prod_{i=1}^{M_b} P_{a_{i1}} P_{a_{i2}}}{\frac{M_b!}{\prod_{a \in \mathcal{A}} n_a!} \prod_{a \in \mathcal{A}} P_{a_i}^{n_a}} \\ &= 2^{n_h} \left(\frac{M_B!}{\prod_{a \in \mathcal{A}} n_a!}\right)^{-1} \end{aligned}$$

unless the event has probability 0.

The distribution of Y conditional on N^G may be reproduced by a random permutation of the original parental genotypes. Similarly, the distribution of Y conditional on N^A may be reproduced by a random permutation original parental alleles, assuming the alleles within a genotype to be ordered. Then the distribution of X conditional on N^A or N^G may be recreated by first randomly permuting the genotypes or alleles among the parents then applying the rules of random Mendelian inheritance according to pedigree B.

We will let $H(X)$ denote any estimator of the pedigree B that is a function of X only. It will be assumed that any relationship implied by $H(X)$ is consistent with the observed genotypes.

Since we are considering a sampling scenario in which N^A or N^G is unobserved we need to construct an estimate of these quantities based on X. We do so by first estimating the genotypes of the putative parents of all estimated sibling groups in $H(X)$. Let S be such a sibling group. The probability of the offspring genotypes conditioned on the parental genotypes is easily calculated. For example, suppose both parents have genotype $< a, b >$. Then genotypes $< a, a >, < b, b >, < a, b >$ appear in the offspring with frequencies $0.25, 0.25, 0.5$ respectively. Then the likelihood of observing n_{aa}, n_{bb}, n_{ab} offspring with these genotypes is proportional to $0.25^{n_{aa}} 0.25^{n_{bb}} 0.5^{n_{ab}}$. We reconstruct the parental genotypes by using the genotype pair which maximizes the likelihood of the offspring genotype frequencies.

One way of doing this is to first construct a list of the genotypes observed in S. There will be at most 4 genotypes if S is a feasible sibling group (FSG) (ie. individuals with genotypes compatible with the hypothesis of mutual full sibship). Define an *assignment* to be an assignment of the alleles in each genotype to the two distinct parents. There are 2^n (not necessarily unique) assignments for n genotypes. A feasible assignment is one which assigns no more than two alleles to a parent. Consider, for example, genotypes $< a, b >, < b, c >$ $, < c, a >$. Then the (unique and unordered) assignments are $abc/abc, ac/ab, ab/bc, ac/bc$. The first is not feasible, and the remaining three represent all parental genotype pairs which may have produced the genotypes observed in S. It is easy to verify that the maximum likelihood parental genotype pair will always be the one obtained from the assignment which maximizes the number of homozygous parents. If, for example, genotypes $< a, b >, < b, c >$ are observed, assignments lead to feasible parent genotype pairs bb/ac and ab/bc. The likelihood for bb/ac will be $0.5^{n_{bb}} 0.5^{n_{ac}}$, whereas the likelihood for ab/bc will be $0.25^{n_{bb}} 0.25^{n_{ac}}$, which is always smaller. This likelihood principle may be extended to the situation in which half sibships are infered by $H(X)$. Estimates of N^A or N^G can be easily constructed from the likelihood estimates of parental genotypes.

The conditional bootstrap procedure can be summarized by the following steps:

1. Estimate B from original data X using $\hat{B} = H(X)$.

2. Construct estimate $\hat{Y}$ of Y using X and $\hat{B}$.

3. Create a bootstrap replicate Y^* of parental genotypes by subjecting either the alleles or the genotypes in $\hat{Y}$ to a random permutation among the parents.

4. Generate a bootstrap replicate X^* of the offspring genotype by simulating inheritance events based on Y^* and $\hat{B}$.

5. Repeat the previous two steps to create a bootstrap sample $X_1^*, \ldots, X_K^*$ and $Y_1^*, \ldots, Y_K^*$ for some large K.

The unconditional bootstrap procedure is identical to the conditional, except that that the bootstrap replicate Y^* is constructed by sampling alleles or genotypes from the estimate $\hat{Y}$ with replacement.

In practice, DNA marker data is usually available for more than one locus. In this case steps 2, 3 and 4 of the bootstrap algorithm can be applied independently to each locus. The resulting single genotype arrays can then be assembled into multilocus genotype arrays X^* and Y^*. This is equivalent to the assumption that the loci are unlinked, and that therefore multilocus genotype probabilities are products of single locus genotype probabilities.

The bootstrap replicates can then be used to estimate the distribution of $D(B, H(X))$

by the empirical distribution function

$$F^*(t) = \sum_{i=1}^{K} \frac{I\{D(H(X), H(X_i^*)) \le t\}}{K}$$

with the bootstrap replicates generated either unconditionally, or conditionally on $T = N^A$ or $T = N^G$, using some partition metric D. A level u confidence set for the partition would take the form

$$\text{CS} = \{b \in \mathcal{B}_I, D(b, H(X)) \le d_u\}$$

where d_u is the $u100$th percentile from the bootstrap distribution F^*. Simlarly the hypothesis

$$H_0 : B = B_0$$

can be tested by rejecting H_0 if $D(B_0, H(X)) > d_u$.

4 Examples

We consider two examples. The first will use DNA marker data which comes from a salmon genetic improvement program run by the Atlantic Salmon Federation. The experiment was designed to assess the impact of confounding of the tank effect and the genetic family effect. Parents were available, and each of the offspring could be matched to a unique parent pair (Herbinger *et al.* (1999)), so that the pedigree is known exactly. There are 781 Atlantic salmon from a single generation which belong to one of 12 familes. Four loci were typed and the number of alleles observed in each of the four loci were 12,14,10 and 8. We assume throughout that there are no half siblings.

For the pedigree estimator $H(X)$, we use the algorithm proposed in Almudevar and Field (1999). In this algorithm the principle of genetic exclusion is used to enumerate all FSGs.

Likelihood based scores are assigned to each FSG, and a partition constructed on that basis. For a partition metric we use $D = D_r$ defined in Section 2.

As a demonstration 225 test samples I were constructed by sampling without replacement 40 individuals from the 781 available. The unconditional bootstrap and the bootstrap conditioned on N^A were applied to each test sample. Each trial i resulted in a 'true' pedigree B_i and offspring genotypes X_i. For both the conditional and unconditional bootstrap methods replications $X^*_{i,1}, \ldots, X^*_{i,K}$ were obtained, from which the estimate F^*_i of the distribution of $D(B, H(X))$ was derived, as discussed in Section 3. The bootstraps each used $K = 500$ replicates. We note that in Almudevar and Field (1999) under this sampling scheme $D(B, H(X))$ was estimated to have mean 4.774 and variance 6.320.

The above procedure was repeated for a second study scenario in which 5 parent pairs had 4 offspring each. For each trial parental alleles were randomly sampled at 4 loci. At each locus there are assumed to be 6 distinct alleles of equal frequency in the population. In this case 200 trials were used, with $K = 500$ bootstrap replicates.

To evaluate the accuracy of the bootstrap randomized confidence intervals were calculated. If F^* is a bootstrap estimate of the cumulative distribution function of $D(B, H(X))$ and f^* is the probability mass function of F^*, then for any $u \in (0, 1)$ let

$$
\begin{aligned}
a_u &= \max\{x : F^*(x) \leq u, f^*(x) > 0\} \\
b_u &= \min\{x : F^*(x) \geq u, f^*(x) > 0\}
\end{aligned}
$$

where the maximum of an empty set is defined as 0. Then the randomized confidence interval is defined as

$$
\phi_u(x; F^*) = \begin{cases} 1 & ; x < b_u \\ \frac{u - F^*(a_u)}{f^*(b_u)} & ; x = b_u \\ 0 & ; x > b_u \end{cases}
$$

if $a_u < b_u$ and

$$\phi_u(x; F^*) = \begin{cases} 1 & ; x \leq b_u \\ 0 & ; x > b_u \end{cases}$$

if $a_u = b_u$, with length

$$L_u(F^*) = a_u + (b_u - a_u)\frac{(u - F^*(a_u))}{f^*(b_u)}.$$

Conceptually, a value x is included in a randomized confidence interval with probability $\phi_u(x; F)$. For our purposes, if X is distributed as F^* then $E[\phi_u(X; F^*)] = u$ for all $u \in (0, 1)$, hence as a diagnostic tool we may plot the estimated actual coverage

$$\hat{\phi}(u) = \sum_{i=1}^{K} \frac{\phi_u\left(D(B_i, H(X_i)); F_i^*\right)}{K}$$

against predicted coverage u. In addition, we can calculate the average length of a level u confidence interval among all trials using

$$\hat{L}(u) = \sum_{i=1}^{K} \frac{L_u(F_i^*)}{K}.$$

Figure 1 contains a plot of $\hat{\phi}(u)$ against u for the conditional and unconditional bootstrap for the first study. Table 2 contains some tabulated figures from this graph. Both graphs reveal a marked departure from identity. The actual coverage is overestimated in the conditional bootstrap, and underestimated in the unconditional bootstrap. For coverages over 0.9, the conditional bootstrap appears to be more accurate. A similar plot is given in Figure 3 for the second study with tabulated figures in Table 2. Here also, the coverage is overestimated in the conditional bootstrap, and underestimated in the unconditional bootstrap. On the other hand, in this study the unconditional bootstrap seems to be more accurate for coverages over 0.9. We can say that each bootstrap is succesful in at least giving a rough indication of the amount of variation in the pedigree estimate.

As for the confidence interval length, Figure 2 contains a plot of average confidence interval length $\hat{L}(u)$ against predicted coverage u and against actual coverage $\hat{\phi}(u)$ (parametrized by u) for the first study. Tabulated figures are given in Table 3. As might be expected the confidence interval sizes are larger for the unconditional bootstrap when compared on the based of predicted coverage, since these confidence intervals were found to be conservative. When they are compared on the basis of the actual coverage, the difference is significantly reduced, with the unconditional confidence intervals remaining slightly larger. A similar effect was found in the second study, indicated in Figure 4 and Table 3.

There is a simple fact about sampling theory which may be of relevance in explaining the differences observed between the two bootstrap procedures. Suppose $p_1, \ldots, p_\nu$ represent the allele frequencies observed among the parents of the pedigree. Then the homozygosity is $\sum_i p_i^2$. If we draw a multinomial sample from these frequencies we obtain sample frequencies $\hat{p}_1, \ldots, \hat{p}_\nu$. It is a fact of elementary probability that

$$E[\sum_i \hat{p}_i{}^2] > \sum_i p_i^2$$

assuming $\nu > 1$ and all probabilities are greater than zero. Therefore the unconditional bootstrap will introduce a positive bias to the homozygosity of the resampled parents. The conditional bootstrap will not have this effect.

5 Conclusion

A procedure for constructing confidence sets for single generation pedigrees based on DNA markers was proposed. The procedure was designed to treat pedigree estimation in a manner analogous to the estimation of parameters in Euclidean space, in which confidence sets take the form of neighbourhoods defined by metric and centered at some point estimate. The coverage of the confidence sets is estimated using a bootstrap procedure. Two approaches are proposed, differentiated by the manner in which unknown population genotype frequencies are treated. In one (the unconditional bootstrap), frequencies are estimated then treated as known. In the second approach (the conditional bootstrap) an estimate of a statistic which is sufficient for the frequencies is calculated, then a conditional confidence set is constructed. The bootstrap technique can be applied to any pedigree reconstruction technique which gives a genetically feasible pedigree as an estimate.

The two approaches were each applied to two distinct sampling scenarios. In each case the conditional approach was found to overestimate the confidence interval coverage while the unconditional approach underestimated the coverage. When compared on the basis of true coverage, each approach gave approximately the same size confidence intervals. The studies therefore suggest that the exact manner in which the bootstrap is carried has implications at least for the accuracy with which coverage is estimated, and therefore this issue bears further scrutiny. It can also be said the the overall approach is a feasible one, which gives at least a rough indication of the accuracy of this type of pedigree estimation.

6 Acknowledgements

The author wish to thank Roger Doyle and Christophe Herbinger for their invaluable assistance to this project. As well, Patrick O'Reilly was involved in the collection and preparation of the salmon data.

7 Bibliography

ALMUDEVAR, A. and FIELD, C. (1999). Estimation of single generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological and Enviornmental Statistics* **4**, 2, 136-165

BLOUIN, M.S., PARSONS, M., LACAILLE, V. and LOTZ, S. (1996). Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology* **5**, 393-401.

BROOKFIELD, J.F.Y. and PARKIN, D.T. (1993). Use of single-locus DNA probes in the establishment of relatedness in wild populations. *Heredity* **70**, 660-663.

HERBINGER, C., O'REILLY, P.T., DOYLE, R.W., WRIGHT, J.M. and O'FLYNN,F. (1999). Early growth performance of Atlantic salmon full-sib families reared in single family tanks versus in mixed family tanks. *Aquaculture* **173**, 105-116,

HERBINGER, C., DOYLE, R.W., TAGGART, C.T., LOCHMANN, S.E., BROOKER, A.L., WRIGHT, J.M. and COOK, D. (1997). Family relationships and effective population size in a natural cohort of Atlantic cod (*Gadus morhua*) larvae. *Can. J. Fish. Aquat. Sci.* **54**, 11-18.

MEAGHER, R.M. and THOMPSON, E. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theor. Pop. Biol.* **29**, 87-106.

PAINTER, I. (1994) Analysis of full-sibship configurations in a maternal half-sibship.

Technical Report No. 272, Dept. of Statistics, University of Washington, Seattle, Washington.

PAINTER, I. (1996). Sibling reconstruction without parental information. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 2, 212-229.

RICE, W.R. (1989). Analysing tables of statistical tests. *Evolution* **43**, 223-225.

THOMPSON, E. (1976). Inference of genealogical structure. *Soc. Sci. Inform.* **15**, 2/3, 477-526.

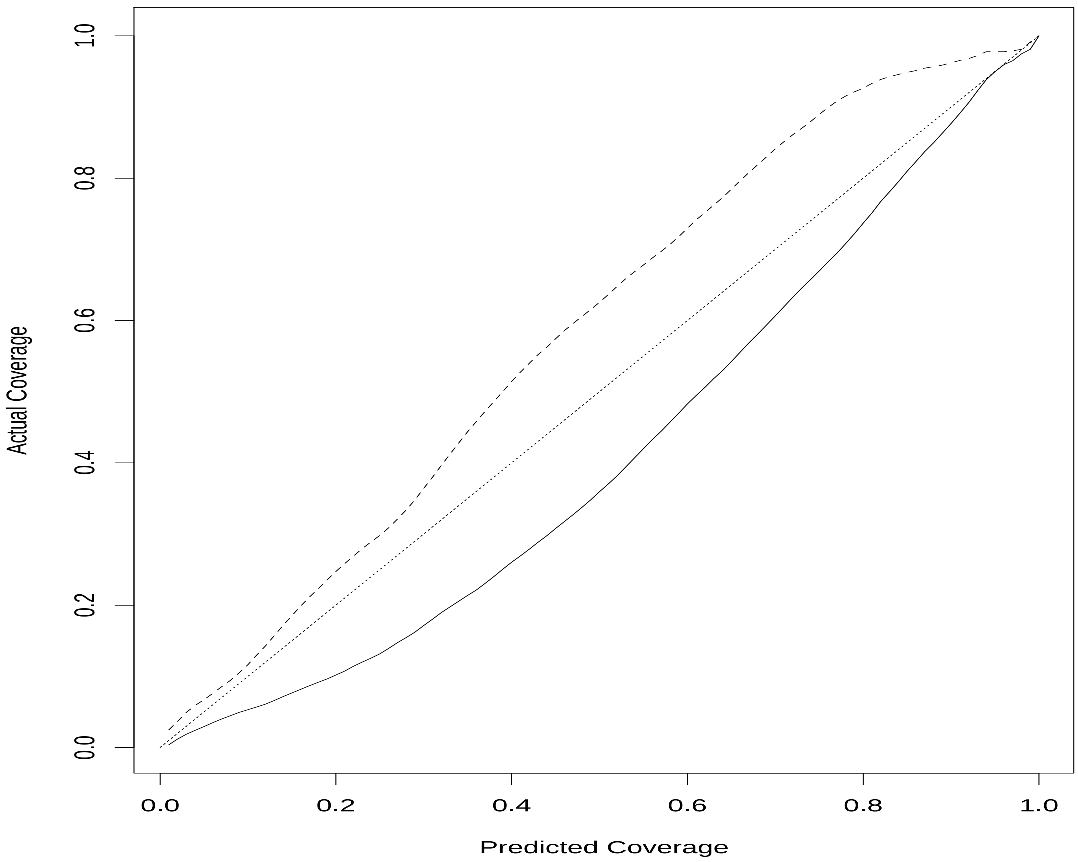


Figure 1: (Study 1) Plot of observed coverage against predicted coverage of randomized confidence intervals. Solid lines represent the conditional bootstrap and dashed lines represent the unconditional bootstrap. The identity is also indicated.

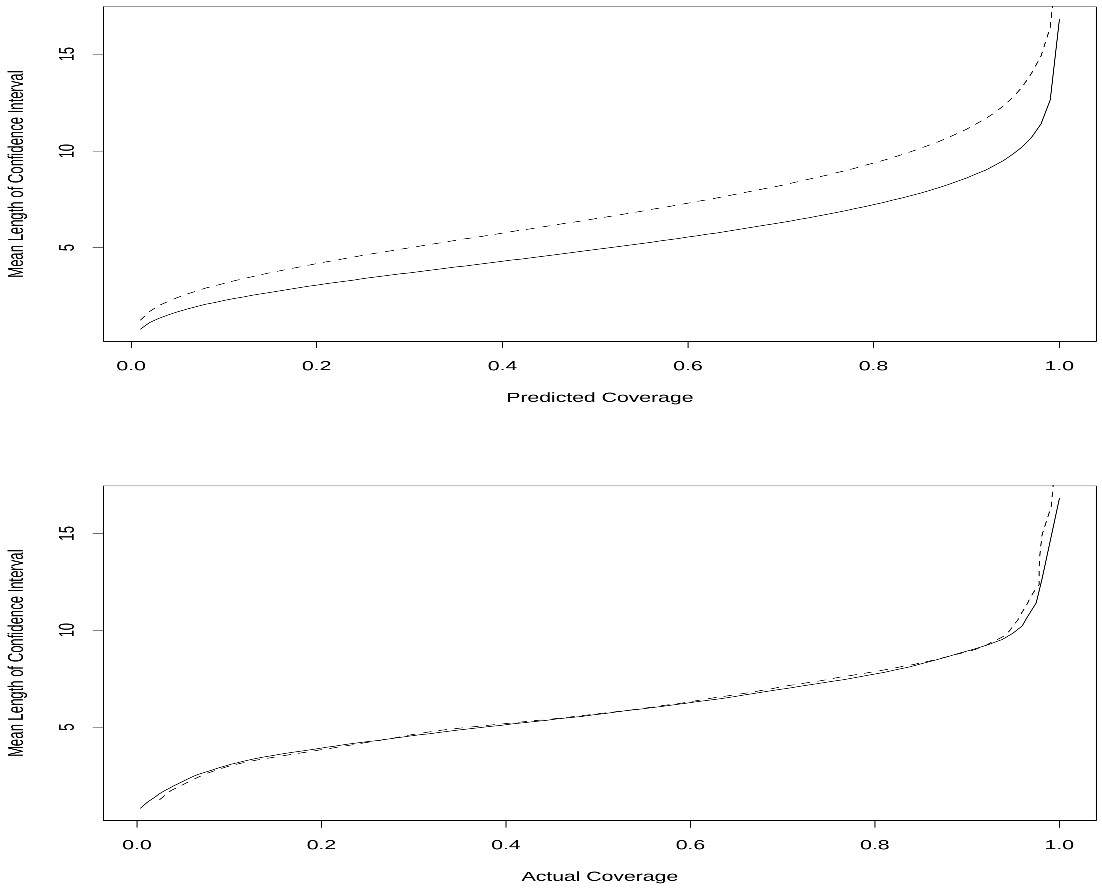


Figure 2: (Study 1) The first graph gives plot of mean randomized confidence interval length against predicted coverage. The second graph gives plot of mean randomized confidence interval length against actual coverage. The solid lines represent the conditional bootstrap and dashed lines represent the unconditional bootstrap.

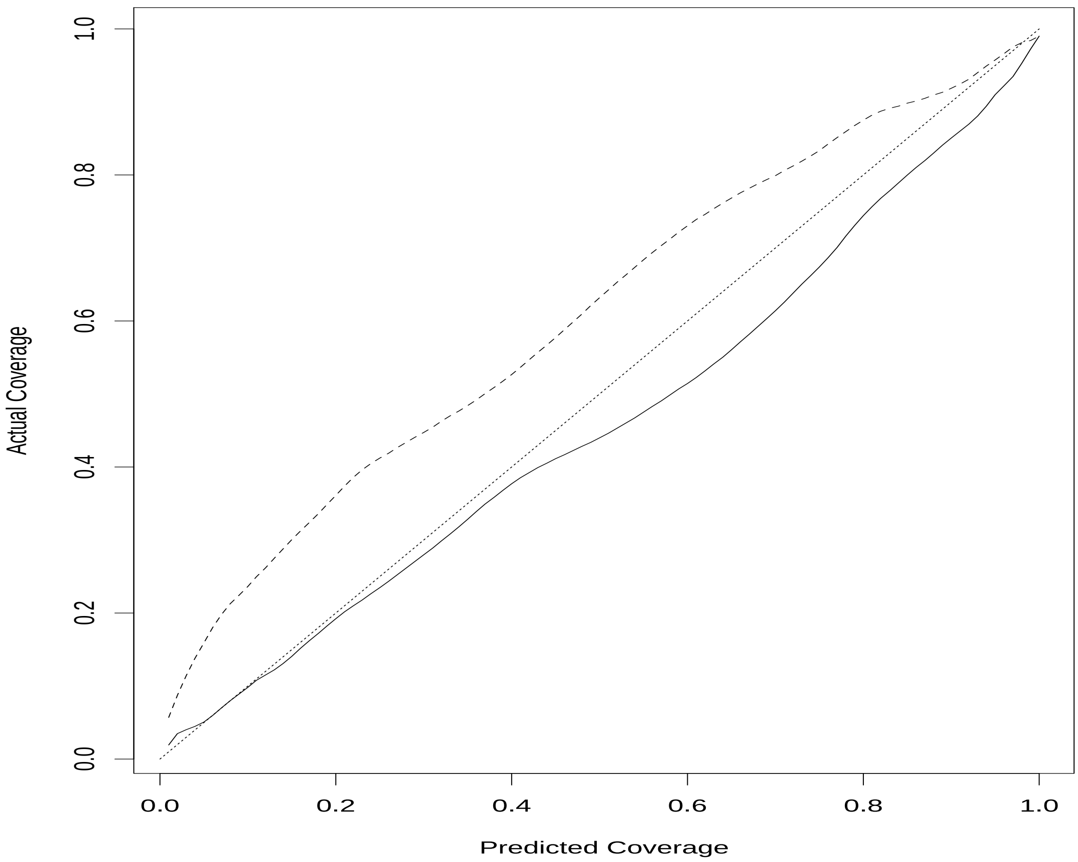
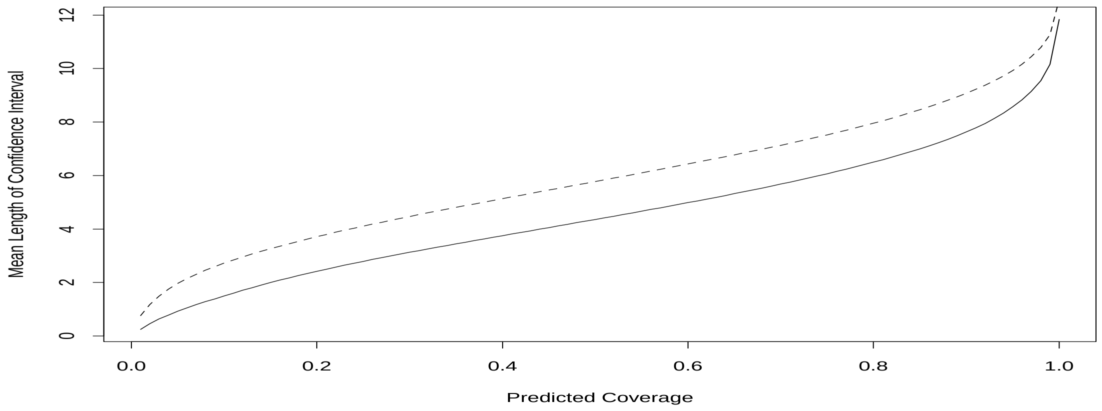


Figure 3: (Study 2) Plot of observed coverage against predicted coverage of randomized confidence intervals. Solid lines represent the conditional bootstrap and dashed lines represent the unconditional bootstrap. The identity is also indicated.



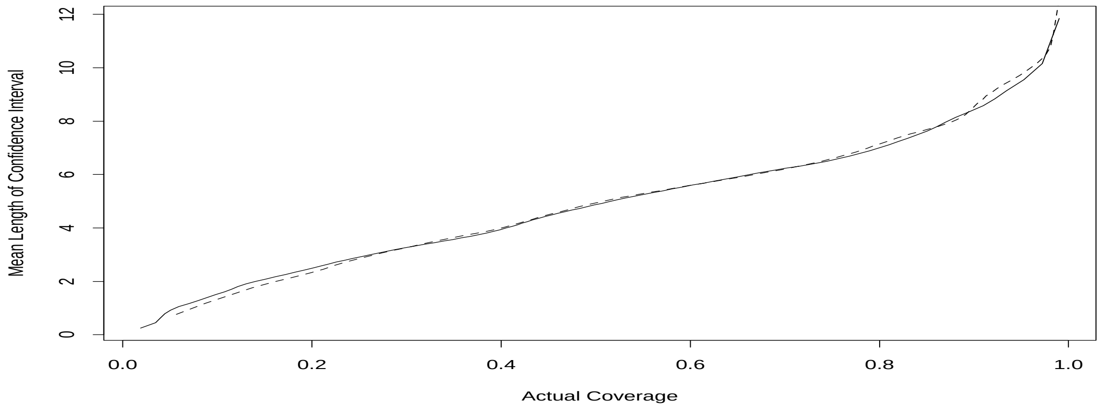


Figure 4: (Study 2) The first graph gives plot of mean randomized confidence interval length against predicted coverage. The second graph gives plot of mean randomized confidence interval length against actual coverage. The solid lines represent the conditional bootstrap and dashed lines represent the unconditional bootstrap.

Table 2: Tabulated values from Figures 1 and 3.

	Study 1		Study 2	
Predicted Coverage	Actual Coverage		Actual Coverage	
	Conditional	Unconditional	Conditional	Unconditional
0.80	0.737	0.927	0.744	0.875
0.81	0.751	0.933	0.757	0.882
0.82	0.767	0.939	0.768	0.887
0.83	0.781	0.943	0.779	0.891
0.84	0.795	0.946	0.789	0.894
0.85	0.810	0.949	0.800	0.898
0.86	0.823	0.951	0.810	0.901
0.87	0.837	0.955	0.820	0.905
0.88	0.849	0.957	0.830	0.909
0.89	0.863	0.959	0.840	0.913
0.90	0.877	0.962	0.851	0.919
0.91	0.891	0.966	0.860	0.924
0.92	0.906	0.968	0.869	0.931
0.93	0.922	0.972	0.880	0.940
0.94	0.938	0.978	0.894	0.949
0.95	0.950	0.978	0.910	0.957
0.96	0.960	0.978	0.922	0.966
0.97	0.965	0.979	0.935	0.975
0.98	0.975	0.981	0.953	0.981
0.99	0.981	0.991	0.972	0.984

Table 3: Tabulated values from Figures 2 and 4.

	Study 1		Study 2	
Actual Coverage	Length of Confidence Interval		Length of Confidence Interval	
	Conditional	Unconditional	Conditional	Unconditional
0.80	7.73	7.87	7.00	7.15
0.81	7.82	7.95	7.11	7.26
0.82	7.92	8.04	7.23	7.38
0.83	8.03	8.13	7.36	7.49
0.84	8.14	8.21	7.48	7.59
0.85	8.27	8.31	7.62	7.68
0.86	8.39	8.42	7.78	7.78
0.87	8.52	8.54	7.96	7.90
0.88	8.65	8.65	8.13	8.03
0.89	8.78	8.77	8.28	8.22
0.90	8.92	8.88	8.43	8.54
0.91	9.06	9.02	8.58	8.84
0.92	9.21	9.22	8.78	9.11
0.93	9.38	9.45	9.03	9.35
0.94	9.57	9.70	9.26	9.54
0.95	9.84	10.22	9.49	9.75
0.96	10.25	10.97	9.78	10.00
0.97	11.05	11.79	10.09	10.28
0.98	12.43	14.36	10.90	10.73
0.99	14.60	16.20	11.84	12.58