Improving Common Population Genetic Estimators by Using Shrinkage

Futschik, Andreas
Inst. f. Statistik, Univ. Wien
Universitaetsstr. 5/9
A-1010 Vienna, Austria
E-mail: andreas.futschik@univie.ac.at

Introduction

Statistical inference in population genetics can be quite challenging. In practice, a complex mix of random drift, demographic and selective forces shapes the history of a population. Modeling this history usually requires to choose an appropriate level of model complexity that captures the essential evolutionary forces, but is not too complex given the data at hand. As the exact likelihood of the data is intractable under more complex models, statistical inference is either based on summary statistics, or some approximations to the likelihood. Summary statistics also play an important role with methods of approximate inference, such as Approximate Bayesian Computation.

A convenient way to model the genealogical history of a sample is by using coalescent trees. In the most basic setup, coalescent trees arise as an asymptotic approximation to the Wright–Fisher model letting the population size N tend to infinity and measuring time in units of $2N$ generations. Under the neutral Wright–Fisher model, the resulting stochastic process is also called simple coalescent, as more complex population genetic models lead to more complex stochastic behavior. Coalescent trees trace the history of a DNA sample up to a most recent common ancestor, and mutations generated by a Poisson process imposed on the coalescent tree are the source of variation in the observed sequence data.

A convenient unit of observation is a locus, i.e. a stretch of DNA that is short enough such that recombination can be ignored. Here, our focus will be on population genetic inference for a single locus. The extension to multiple independent loci is straightforward.

An important quantity that captures the variation in sequence data is the scaled mutation parameter

$$\theta = 2N\mu.$$

Here N is the population size measured in the number of DNA sequences that exist for a given locus in the population, and μ is the rate of mutation per generation for a DNA sequence taken at the considered locus. For a diploid species, the number of sequences N equals twice the number of individuals in the population.

In the coalescent framework, θ is introduced most easily by imposing a Poisson process on the coalescent tree that generates mutations. The rate of this Poisson process is $\theta/2$. Consequently, the number of mutations S occurring on the tree has a Poisson distribution with parameter $l_n\theta/2$, given a coalescent tree for a random sample of size n of total length $L_n = l_n$. If a mutation, say from the DNA base "A" to "T", occurs at a given position, the mutated base "T" is called derived allele. The frequency of "T" in the sample depends on where on the tree the mutation occurred.

Here, our focus is on estimating θ, using a sample consisting of DNA sequences from a locus. Several estimators of θ have been proposed for the case where n individual reads are available each covering the whole locus. Under the simple coalescent model, Futschik and Gach (1998) showed that the MSE of Watterson's estimate and Tajima's π can be uniformly improved by shrinkage. Here, we extend these results to cover further population genetic estimators.

Then we will show how some estimates of θ can be optimized with next generation sequencing

reads from pooled data. We model these reads by a second stage of local independent sampling with replacement. Simulation results will demonstrate the amount of possible improvement. As the analysis of next generation sequencing data is becoming more and more important, this can be viewed as the main contribution of this paper.

Estimating θ

r Several estimators of θ can be found in the literature. Watterson's estimator $\hat{\theta}_W$, proposed by Ewens (1974) and Watterson (1975), is among the most popular. Given a sample of size n containing S_n segregating sites at which mutations occurred, $\hat{\theta}_W$ is defined as

$$\hat{\theta}_W := \frac{2S_n}{\mathrm{E}(L_n)},$$

where L_n is the total length of the coalescent tree. The motivation behind the normalization constant $\mathrm{E}(L_n)$ is to obtain an unbiased estimator of θ in the absence of knowledge of the actual total tree length. Under the neutral Wright-Fisher model, $\mathrm{E}(L_n) = \sum_{i=1}^{n-1} 1/i$.

Tajima's (1983) estimator $\hat{\theta}_\pi$ is defined as the average number of differences between all $\binom{n}{2}$ pairs of the sequences. It is less efficient than Watterson's estimator, and even inconsistent. However, $\hat{\theta}_\pi$ is an ingredient to Tajima's D (Tajima (1989)), where the normalized difference between $\hat{\theta}_\pi$ and $\hat{\theta}_W$ is used as a test for neutrality.

Further estimator's of θ have been proposed such as those by Fay and Wu (2000) or by Zeng et. al. (2006). A common property of all these estimators is that they are unbiased under the neutral Wright-Fisher model, and that their variance is of the form

$$a_n\theta + b_n\theta^2. \tag{1}$$

See section 2.2. in Durrett (2008) for details. As will be shown below, the term $b_n\theta^2$ in the above formula, implies that all these estimates are inadmissible and can be improved uniformly by shrinkage. The following result implies that improvement by shrinkage is possible for any estimator with a variance structure as in (1). It generalizes results by Futschik and Gach (2008) for $\hat{\theta}_W$ and $\hat{\theta}_\pi$.

Lemma 1. *Let $\hat{\theta}$ denote an estimate of a parameter $\theta > 0$. Assume furthermore that $E(\hat{\theta}) = \theta$ and*

$$Var(\hat{\theta}) = a\theta + b\theta^2 \tag{2}$$

with $a, b \geq 0$. Then with $c := [a/\theta + (b+1)]^{-1}$

$$MSE(c\hat{\theta}) \leq MSE(\hat{\theta}),$$

and strict inequality holds, if $c < 1$, i.e. unless $a = b = 0$. If $b > 0$, an estimator uniformly better than $\hat{\theta}$ is given by $\hat{\theta}_s := \frac{\hat{\theta}}{b+1}$.

Proof: A decomposition of the MSE into variance and squared bias leads to

$$\mathrm{MSE}(c\hat{\theta}) = c^2\theta(a + b\theta) + (c-1)^2\theta^2.$$

By taking the derivative with respect to c, we obtain $c := [a/\theta + (b+1)]^{-1}$. By observing our the objective function is unimodal, it follows that $\frac{\hat{\theta}}{b+1}$ is uniformly better than $\hat{\theta}$ for $b > 0$. □

Notice that $b = 0$ in a classical Poisson model. A uniform improvement of the maximum likelihood estimator $\hat{\theta}$ is not possible here, as c depends on the unknown parameter θ. Substituting an estimator of θ into c does not lead to a uniform improvement, as $\hat{\theta}$ is admissible with respect to the MSE in a one-dimensional setting, see Johnstone (1984).

Although the total length L_n of the coalescent tree is unknown in practice, its probability distribution is determined by the considered model. For the simple coalescent in particular, there

exist explicit formulas for the distribution as well as the expected value and the variance of L_n. The randomness of the observation interval of the Poisson process generating mutations leads to a situation, where $b > 0$ in (1). Thus shrinkage by the factor $1/(b+1)$ leads to a uniform improvement of unbiased estimators.

In the table below, we illustrate this for several well known estimates of θ. We also state the variance of these estimates, so that the shrinkage constant b can be read off easily. For further details on the estimates see for instance section 2.2. in Durrett (2009). In the table, η_i denotes the number of sites where the mutant allele is present i times in a sample of size n. Furthermore $c_n := \sum_{i=1}^{n-1} i^{-1}$.

estimate	formula	variance
Watterson (1975)	$\hat{\theta}_W$	$\theta/c_n + \sum_{i=1}^{n-1} i^{-2}/c_n^2\theta^2$
Tajima (1983)	$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i)\eta_i$	$\frac{n+1}{3(n-1)}\theta + \frac{2[n^2+n+3]}{9n(n-1)}\theta^2$
Fu and Li (1993)	$\hat{\theta}_{FL} = \eta_1$	$\theta + 2\frac{nc_n-2(n-1)}{(n-1)(n-2)}\theta^2$
Zeng, Fu, Shi, Wu (2006)	$\hat{\theta}_L = \frac{1}{n-1} \sum_{i=1}^{n} i\eta_i$	$\frac{n}{2(n-1)}\theta + \left[2\frac{n^2}{(n-1)^2}(\sum_{i=1}^{n} i^{-2} - 1) - 1\right]\theta^2$
Fay and Wu (2000)	$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2\eta_i$	$\theta + \frac{2[36n^2(2n+1)\sum_{i=1}^{n} i^{-2} - 116n^3 + 9n^2 + 2n - 3]}{9n(n-1)^2}\theta^2$

Next Generation DNA Sequencing

In recent years, massively parallel high throughput sequencing became extremely popular. This new methodology permits to obtain a large amount of sequence information at relatively low cost, especially when compared to classical Sanger sequencing. This may be the reason why the technology is also known as next generation sequencing. The decreased sequencing cost makes is possible for population geneticists to obtain data at a much larger scale, and even time series of sequence data are now obtained in experimental evolution experiments. One practical challenge is that the reads obtained are relatively short, which can make sequence alignment to a reference genome ambiguous in come cases, especially at genome positions that involve repetitive sequence elements. Another challenge is that the sequence reads contain a certain proportion of reading errors that need to be taken into account. A popular way to assess the reliability of reads is to use Phred quality scores that can be translated into a probability that the nucleotide obtained at a certain position is correct. See Li et al. (2008) for details.

To avoid sequencing errors, individuals are often sequenced separately. An important parameter describing the quality of sequencing is the coverage. The number of reads at a given DNA position is usually assumed to follow a Poisson distribution, and the coverage is then defined as the expected value λ of this Poisson distribution. A large enough value of λ ensures that a given position is usually read sufficiently often such that the genotype at this position can be determined with a negligible error probability. For diploid organisms, the coverage needs to be higher than for haploids, as two bases need to be determined at each position.

The cost of next generation sequencing is still an important constraint when sequencing a larger number of individuals. Thus especially for non-model organisms, pooling became popular as a cost effective design. A pooled sample consists of several individuals that are sequenced simultaneously. Sequencing is done by drawing with replacement small sequence chunks from the pool of individual sequences. While being a cost saving design, data from pooling experiments are more challenging to analyze. Besides the difficulty of distinguishing between rare alleles and sequencing errors, the sampling process with replacement adds an additional layer of randomness. Futschik and Schlötterer (2010) and Kofler et al. (2011) show that standard population genetic estimates become biased as a consequence. For a further methodological discussion of the analysis of pooled samples see for instance

Kolaczkowski et al. (2011).

Estimating θ using Pooled Next Generation Sequencing Data

Suppose we have a pool of n DNA sequences covering a locus G of interest. This leads to a partition of G into k random intervals $G = I_1 \cup I_2 \cup \cdots \cup I_k$ such that within each subinterval I_j, any position is covered by the same number or reads M_j. Let u_j denote the length of I_j. Then the length of G is given by $u := \sum_{j=1}^{k} u_j$. Furthermore on I_j the scaled mutation parameter is $\theta\frac{u_j}{u}$, leading to a total of $\theta = \sum_{j=1}^{k} \theta_j$ for the whole locus. It is thus natural to estimate θ_j by $\hat{\theta}_j$ on I_j and define $\hat{\theta} = \sum_{j=1}^{k} \hat{\theta}_j$.

To avoid biases caused by sequencing errors, the reads are filtered first, and only reads with a certain minimum quality implied by the Phred score are used. (See Li et al. (2008).) After filtering out those reads with insufficient quality, we may assume that the error probability for the remaining reads is bounded by some constant ϵ. Here, values such as $\epsilon = 0.01$ are common. Given M_j reads for a particular subinterval I_j, a position within the interval is a candidate for a segregating site, if not all reads at the position are identical. In such a case, $1 \leq X < M_j$ of the reads are from the so called derived allele, and $M_j - X$ from the ancestral allele. (We follow the infinite sites model, and assume that the probability of more than two alleles at a position is negligible.) To further protect against sequencing errors, a site at which reads are polymorphic is only used, if the minor (less frequent) allele has a certain minimum frequency d. The intention behind this second stage of filtering is not to loose too many reads in the first stage by setting the inclusion threshold too high. The threshold d is chosen as to keep the number of wrongly identified segregating sites small. See Futschik and Schlötterer (2010) for a discussion of the choice of d, and its consequence on the probability of wrongly identifying segregating sites.

Suppose now that our pool of reads contains a random number S of segregating sites $w_1, \ldots, w_S$ within subinterval I_j after the first stage of filtering. Then the following versions of Watterson's θ and Tajima's π achieve the desired protection against sequencing errors on subinterval I_j with an appropriately chosen threshold $d \geq 1$ and $M = M_j$.

$$\hat{\theta}_{W,j}^{(d)} := \sum_{i=1}^{S} \frac{\mathbf{1}_{[d \leq X_i \leq M-d]}}{c_M}$$

where $c_M = \sum_{i=1}^{M-1} 1/i$

$$\hat{\theta}_{\pi,j}^{(d)} = \sum_{i=1}^{S} \binom{M}{2}^{-1} X_i(M - X_i)\mathbf{1}_{[d \leq X_i \leq M-d]}$$

Here $\mathbf{1}_{[E]}$ denotes the indicator function of an event E. More generally, we consider estimates of the form

$$\hat{\theta}^{(d)} = \sum_{i=1}^{S} W(M, X_i).$$

Notice that S is unknown in practice, and the sum is taken over all segregating sites found within the considered subinterval. If we assume that a sufficient protection against sequencing errors by choosing the quality score and d large enough has been chosen, the practical approach leads to essentially equivalent estimates.

Both requiring a minimum minor allele frequency d, and the sampling with replacement leads to biased estimators of θ. For Tajima's π and Watterson's θ, the following bias correction terms $\gamma(M)$ have been derived in Futschik and Schlötterer (2010): For Tajima's π,

$$\gamma(M) = \binom{M}{2}\Big[\sum_{m=d}^{M-d}\sum_{r=1}^{n-1} m(M-m)P(X = m|Y_n = r)r^{-1}\Big]^{-1}. \tag{3}$$

Here X is the number of reads from the minor allele at a segregating site, and Y_n the number of minor alleles at this position in the pool of size n. According to the model of our reading process as binomial sampling

$$P(X = m|Y_n = r) = \binom{M}{i} \left(\frac{r}{n}\right)^i \left(1 - \frac{r}{n}\right)^{M-i}. \tag{4}$$

Let furthermore $F_{(B)}(x, M, p)$ be the binomial cumulative distribution function

$$F_{(B)}(x, M, p) = \sum_{i=0}^{x} \binom{M}{i} p^i (1-p)^{M-i}.$$

Then for Watterson's θ,

$$\gamma(M) = \frac{\sum_{i=1}^{M-1} i^{-1}}{\sum_{r=1}^{n-1} [F_{(B)}(M-d, M, r/n) - F_{(B)}(d-1, M, r/n)] \frac{1}{r}}. \tag{5}$$

We denote the bias corrected version of such estimates by

$$\hat{\theta}^{(d)*} = \gamma(M) \sum_{i=1}^{S} W(M, X_i).$$

On I_j, it holds that $\mathrm{E}(\hat{\theta}^{(d)*}|M) = \theta \frac{u_j}{u}$.

Given the resulting bias corrected estimates, shrinkage can again be used as in Lemma 1 to reduce the MSE.

Lemma 2. *Under the Wright-Fisher model, the unbiased estimator* $\hat{\theta}_j^{(d)*}$ *of* $\theta \frac{u_j}{u}$ *has a MSE which is uniformly larger than that of*

$$\hat{\theta}_{s,j}^{(d)*} \frac{\hat{\theta}^{(d)*}}{b_j + 1}$$

where $b_j = [\gamma(M_j) E(W(M_j, X_1))]^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$.

Proof: The computations are done conditionally on M. For simplicity, we drop the subscript j. Conditioning on the number of segregating sites in the pool, we obtain the following lower bound on the variance:

$$Var(\hat{\theta}^{(d)*}) = \mathrm{E}(Var(\hat{\theta}^{(d)*}|S)) + Var(E(\hat{\theta}^{(d)*}|S)) \geq Var(E(\hat{\theta}^{(d)*}|S)).$$

Furthermore

$$\mathrm{E}(\hat{\theta}^{(d)*}|S)) = \gamma(M)\mathrm{E}(W(M, X_1)|M)S$$

and

$$Var(E(\hat{\theta}^{(d)*}|S)) = [\gamma(M)\mathrm{E}(W(M, X_1)|M)]^2 Var(S). \tag{6}$$

Now under neutral Wright-Fisher model

$$Var(S) = \theta \frac{u_j}{u} \sum_{i=1}^{n-1} \frac{1}{i} + [\theta \frac{u_j}{u}]^2 \sum_{i=1}^{n-1} \frac{1}{i^2}.$$

As in Lemma 1, it now follows that $\frac{\hat{\theta}^{(d)*}}{b_j+1}$ is a estimate uniformly better than $\hat{\theta}^{(d)*}$, if b_j is chosen as the coefficient b in front of $[\theta \frac{u_j}{u}]^2$ in (6). The result follows since $\gamma(M)\mathrm{E}(W(M, X_1) = 1$. □

For the bias corrected versions of Watterson's estimate $\hat{\theta}_{W,j}^{(d)*}$ and Tajima's $\hat{\theta}_{\pi,j}^{(d)*}$, Lemma 2 leads to easily implementable improved estimates.

Indeed under the Wright-Fisher model, the shrinkage terms are given as

$$b_j = [\gamma(M)\mathrm{E}(W(M, X_1))]^2 \sum_{i=1}^{n-1} \frac{1}{i^2} = \frac{\sum_{i=1}^{n-1} \frac{1}{i^2}}{[\sum_{i=1}^{n-1} \frac{1}{i}]^2}$$

both for $\hat{\theta}_{W,j}^{(d)*}$ and $\hat{\theta}_{\pi,j}^{(d)*}$. Notice that b_j is the same for all subintervals. Therefore it is not hard to see that

$$\sum_{j=1}^{k} \hat{\theta}_{s,j}^{(d)*} = \frac{\sum_{i=1}^{n-1} \frac{1}{i^2}}{[\sum_{i=1}^{n-1} \frac{1}{i}]^2} \sum_{j=1}^{k} \hat{\theta}_{j}^{(d)*}$$

is an estimator for the overall mutation parameter θ of the locus that has a MSE uniformly smaller than that of $\sum_{j=1}^{k} \hat{\theta}_{j}^{(d)*}$

Simulation Results

In order to investigate the amount of gain achieved in practice by shrinkage estimators for pooled next generation sequencing samples, we carried out a simulation experiment involving 1000 simulated samples consisting of NGS reads from a locus. We estimated θ using both the bias corrected estimates $\hat{\theta}_{W,j}^{(d)*}$ and Tajima's $\hat{\theta}_{\pi,j}^{(d)*}$. The simulation results shown in Figures 1-3 below show that shrinkage can lead to a substantial improvement. The gain achieved by shrinkage is larger for Watterson's θ than for Tajima's π. When sample size and sequencing effort increases, the improvement achieved by the shrinkage estimate tends to decrease (Figure 3). Finally estimates tend to become better when the threshold d is made smaller (while still keeping sufficient control of sequencing errors), but the relative improvement achieved by shrinkage remains similar.

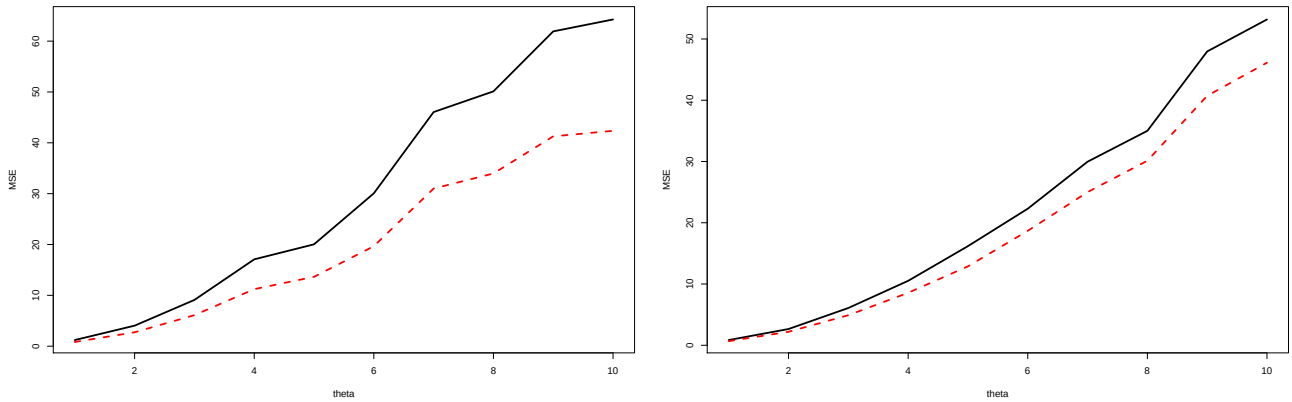

Figure 1: MSE of Watterson's estimate (left panel) and Tajima's π (right panel) in dependance of θ for pooled samples. Pool size $n = 10$, expected coverage $\lambda = 20$, minimum minor allele frequency $d = 3$. Solid line bias corrected estimate $\hat{\theta}^{(d)*}$, dashed line shrinkage estimate $\hat{\theta}_s^{(d)*}$

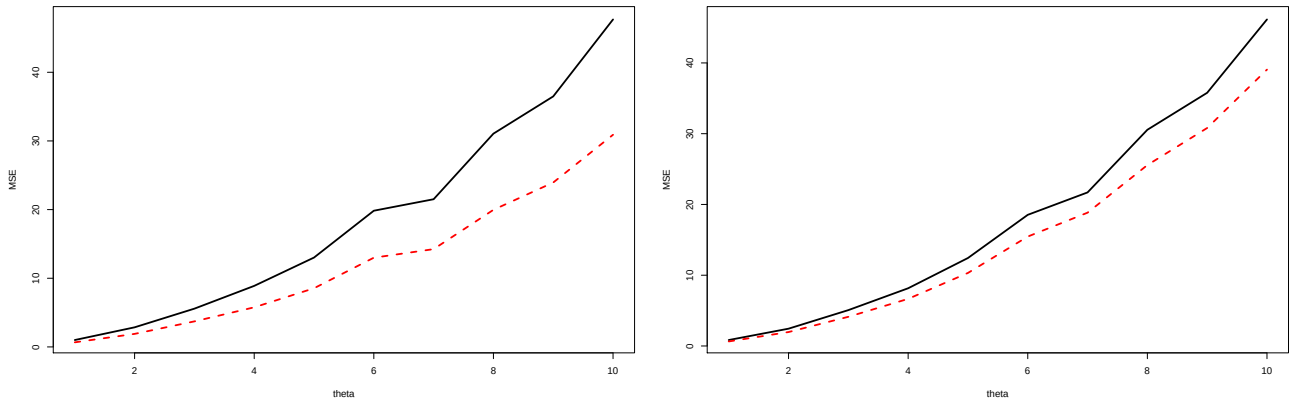

Figure 2: MSE of Watterson's estimate (left panel) and Tajima's π (right panel) in dependance of θ for pooled samples. Pool size $n = 10$, expected coverage $\lambda = 20$, minimum minor allele frequency $d = 2$. Solid line bias corrected estimate $\hat{\theta}^{(d)*}$, dashed line shrinkage estimate $\hat{\theta}_s^{(d)*}$

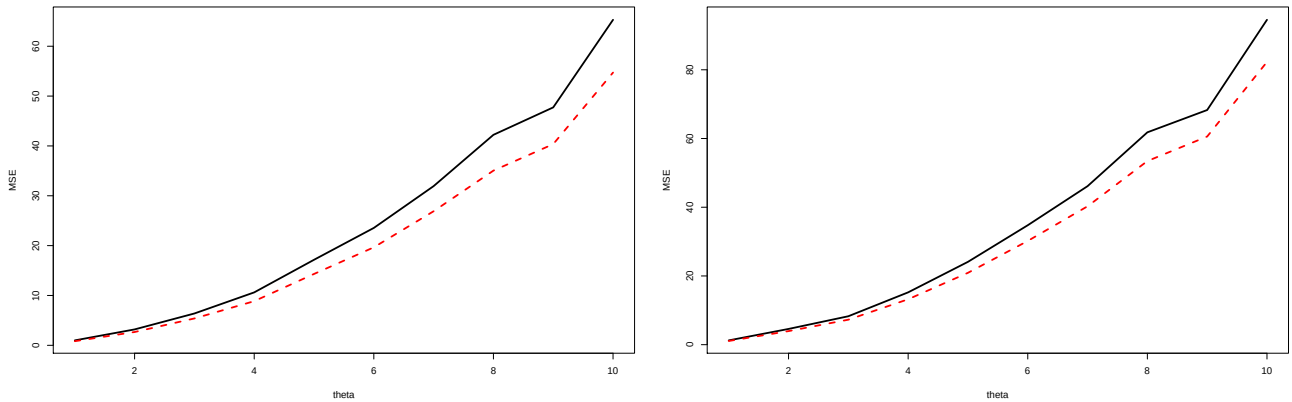

Figure 3: MSE of Watterson's estimate (left panel) and Tajima's π (right panel) in dependance of θ for pooled samples. Pool size $n = 50$, expected coverage $\lambda = 50$, minimum minor allele frequency $d = 3$. Solid line bias corrected estimate $\hat{\theta}^{(d)*}$, dashed line shrinkage estimate $\hat{\theta}_s^{(d)*}$

REFERENCES (RÉFERENCES)

Durrett, R. Probability Models for DNA Sequence Evolution. *Springer, New York, 2008.*

Ewens, W. J., 1974 A note on the sampling theory for infinite alleles and infinite sites models. *Theoretical Population Biology, 6, 143-148.*

Fay, J.C. and Wu C.-I. (2000) Hitchhiking under positive Darwinian selection. *Genetics, 155:1405-1413.*

Fay, Y.X. and Li W.H. (1993) Statistical tests of neutrality of mutations. *Genetics, 133:693-709.*

Futschik A. and Gach F. (2008) On the inadmissibility of Watterson's estimator. *Theoretical Population Biology, 73, 212-221.*

Futschik, A., & Schlötterer, Ch. (2010) Massively Parallel Sequencing of Pooled DNA Samples–The Next Generation of Molecular Markers. *Genetics 186, 207-218*

Johnstone, I.M. (1984) Admissibility, difference equations and recurrence in estimating a Poisson mean. *Ann. Statist., 12, 1173-1198.*

Kofler, R., Orozco-ter Wengel, P., De Maio, N., Pandey, R.V., Nolte V., Futschik, A., Kosiol, C., Schlötterer, Ch. (2011) PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE 6(1): e15925.*

Kolaczkowski, B., Kern, A.D., Holloway, A.K. & Begun, D.J. (2011) Genomic Differentiation Between Temperate and Tropical Australian Populations of Drosophila melanogaster. *Genetics 187, 245-260.*
Li, H., Ruan, J.; & Durbin, R. (2008). Mapping Short DNA Sequencing Reads and Calling Variants using Mapping Quality Scores. *Genome Research, 18, 1851-185.*
Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics 105, 437-460.*
Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics 123, 585-595.*
Watterson, G.A., (1975). On the number of segregating sites. *Theoretical Population Biology 7, 256-276.*
Zeng, K., Fu, Y.-X., Shi, S., and Wu, C.-I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics, 174: 1431-1439.*

RÉSUMÉ (ABSTRACT)

Estimating population genetic parameters is challenging, as common estimators often exhibit a high variance. Under these circumstances it is surprising that some of the commonly used estimators are inadmissible with respect to the mean squared error and can be uniformly improved. In this article, we first review previous work on this subject and also provide improved versions of estimators for which shrinkage has not yet been investigated. This work is based on classical Sanger sequencing data.

As new high throughput sequencing techniques are becoming more and more popular, we then investigate population genetic inference for such next generation sequencing data. While new sequencing techniques provide sequence information at a fraction of cost of Sanger sequencing, individual sequencing of a even moderate samples can still be quite cost intensive. Especially for non-model organisms, it is therefore quite popular to sequence entire pools of individuals simultaneously. Inference based on such pooled samples raises methodological challenges, as sequencing errors cannot be readily identified. We derive uniformly improved estimates under such an experimental design.