Published in final edited form as:
J Comput Biol. 2016 September ; 23(9): 756–768. doi:10.1089/cmb.2016.0039.

Improved Versions of Common Estimators of the Recombination Rate

Kerstin Gärtner and
Vienna Graduate School of Population Genetics, Institut für Populationsgenetik, Vetmeduni Vienna, 1210 Vienna, Austria, Phone: +43 1 25077 4336, Fax: +43 1 25077 4390

Andreas Futschik*
Department of Applied Statistics, Johannes Kepler University, 4040 Linz, Austria, Phone: +43 732 2468 6822, Fax: +43 732 2468 6800

Kerstin Gärtner: Kerstin.Gaertner@vetmeduni.ac.at

Abstract

The scaled recombination parameter ρ is one of the key parameters, turning up frequently in population genetic models. Accurate estimates of ρ are difficult to obtain, as recombination events do not always leave traces in the data. One of the most widely used approaches is composite likelihood. Here we show that popular implementations of composite likelihood estimators can often be uniformly improved by optimizing the trade-off between bias and variance. The amount of possible improvement depends on parameters such as the sequence length, the sample size, and the mutation rate, and can be considerable in some cases. It turns out that ABC, with composite likelihood as a summary statistic, also leads to improved estimates, but now in terms of the posterior risk. Finally, we demonstrate a practical application on real data from *Drosophila*.

1 Introduction

In diploid organisms, homologous chromosomes are paired during meiosis. In this process, pieces of DNA are frequently exchanged between the chromosomes, leading to a mixture of maternal and paternal genetic information. This process is called recombination. By producing new combinations of alleles and breaking up the linkage between genes, recombination increases the variation in a population, making it an important evolutionary force. Recombination rates vary between species and across the genome. Knowing the respective recombination rates is of great importance in several situations. It is e.g. necessary for understanding the process of recombination itself. At the population level, knowing the population recombination rate $\rho = 4N_e r$, with N_e being the effective population size and r being the recombination rate per base pair (bp), is important for the analysis of population genetic data. For instance, as recombination reduces the amount of linkage disequilibrium (LD) between segregating sites (Hill and Robertson, 1986) and positive selection tends to

*andreas.futschik@jku.at.

Author Disclosure Statement
The authors confirm that no competing financial interests exist.

produce areas of high LD, recombination helps to localize signals of selection in DNA sequence data, see e.g. Sabeti et al. (2002, 2006); O'Reilly et al. (2008).

However, obtaining accurate estimates of the recombination rate is challenging as not all historical recombination events leave traces in a corresponding sample of DNA sequences. Even the best estimation methods available provide estimates that exhibit a considerable amount of uncertainty. The literature suggests several different methods for estimating ρ, including the computation of lower bounds on the number of recombination events (Hudson and Kaplan, 1985; Wiuf, 2002; Myers and Griffiths, 2003), the calculation of moments or other summary statistics (Hudson, 1987; Wall, 2000; Batorsky et al., 2011), and regression based methods (Lin et al., 2013). Approaches based on maximum likelihood are used commonly as well.

Due to the high computational effort, these methods are often either approximate likelihood methods (Hey and Wakeley, 1997; Hudson, 2001; Fearnhead and Donnelly, 2002; McVean et al., 2002; Li and Stephens, 2003; Wall, 2004) or, if they are full likelihood methods, they still approximate the likelihood e.g. via importance sampling or Markov chain Monte Carlo (MCMC) algorithms (Griffiths and Marjoram, 1996; Kuhner et al., 2000; Fearnhead and Donnelly, 2001).

Hobolth and Jensen (2014) describe a method to estimate the recombination rate based on Markov approximations to the tree building process of the ancestral recombination graph (McVean and Cardin, 2005; Marjoram and Wall, 2006). Other methods for estimating recombination rates use approximate Bayesian computation (ABC), which is a Bayesian method that avoids the calculation of a likelihood function, e.g. Lopes et al. (2014); Arenas et al. (2015). For an overview on ABC see for instance Beaumont et al. (2002).

Some of the methods are implemented as software packages such as *LDhat* (McVean and Auton, 2007) or *LDhelmet* (Chan et al., 2012). These programs use the composite likelihood method of Hudson (2001) or, more precisely, a modification of the latter by McVean et al. (2002) implementing a finite-sites mutation model. A good overview on composite likelihood methods is provided by Varin et al. (2011). *LDhat* and *LDhelmet* also permit to estimate recombination rates that vary across the genome by combining composite likelihood with a Bayesian approach using a reversible jump Markov chain Monte Carlo (rjMCMC) algorithm (Green, 1995).

In this paper, we investigate whether there is room for improving composite likelihood estimators. As a measure of performance for an estimator $\tilde{\rho}$ of ρ, we focus on the mean squared error

$$\mathrm{MSE}_{\rho}(\tilde{\rho}) := \mathbb{E}(\tilde{\rho} - \rho)^2 .$$

The MSE provides the expected squared distance between true parameter and its estimate and may be decomposed into the sum of variance and squared bias:

$$\mathrm{MSE}_{\rho}(\tilde{\rho}) = \mathrm{Var}_{\rho}(\tilde{\rho}) + \mathrm{Bias}_{\rho}(\tilde{\rho})^2 \quad (1)$$

Estimators that can be uniformly improved with respect to the MSE are called inadmissible in the statistical literature, see e.g. Berger (2013). In classical statistics, shrinkage sometimes leads to such a uniform improvement, see e.g. Gruber (1998). In a population genetic context, Futschik and Gach (2008) showed that Watterson's estimator of the scaled mutation parameter θ is inadmissible, and provided a uniformly better estimator by shrinkage, i.e. multiplying the original estimator with a suitable constant $c < 1$. In subsequent sections, we show that such uniform improvements are often also possible for composite likelihood estimators of ρ.

For our practical computations, we will use the composite likelihood estimator implemented in *LDhelmet*, and also consider an older estimator provided by *LDhat*. Our focus is on stretches of DNA with constant recombination. For recombination landscapes the approach would need to be applied separately on each segment with distinct recombination rate. We do this when we apply our method to real data in section 5.

The remainder of this paper is structured as follows: The composite likelihood method of McVean et al. (2002) for estimating ρ and an alternative version implemented in *LDhelmet* are explained in section 2, as well as our method of improvement. In section 3, we explore the improvement of two implementations of the composite likelihood method by *LDhat* and *LDhelmet* and present simulations and results. We briefly discuss an alternative approach for improving the estimation of ρ based on ABC in section 4. An example for the application of our method to real data is shown in section 5. Section 6 concludes this paper by a discussion and an outlook.

2 Estimating the population recombination rate by composite likelihood and possible improvements

In this section we explain how composite likelihood has been implemented for estimating ρ. Further, we explain our approach to improve composite likelihood estimates.

2.1 A composite likelihood estimate of ρ

The composite likelihood method of McVean et al. (2002) extends the method of Hudson (2001) by permitting repeated mutations to occur at a site during the history of a sample. However, these (reversible) mutations are assumed to lead to no more than two alleles segregating. The estimation process is carried out in four steps: At first the population mutation rate θ per site is estimated. Hereby an approximate finite-sites version of Watterson's estimator is used. The second step is to classify every pair of segregating sites into sets of equivalent configurations. In the next step, the likelihood of each of these sets is estimated under the value of θ from step 1 and a range of values for ρ using the importance sampling method of Fearnhead and Donnelly (2001). At last, ρ is estimated for the whole sequence by combining the likelihoods from all pairs of segregating sites. The estimated ρ is the value with the highest composite log likelihood (McVean et al., 2002).

The described method is implemented in the software packages *LDhat* and *LDhelmet*, with the latter package implementing some more accurate approximations (Chan et al., 2012). The improvement in accuracy results for instance from solving a system of recursion equations for the computation of the pairwise likelihoods instead of applying importance sampling, and the implementation of a quadra-allelic mutation model instead of a biallelic one.

In the following, we denote the estimator provided by the function *pairwise* in *LDhat* by $\hat{\rho}$. Furthermore $\breve{\rho}$ signifies the estimator implemented as *max_lk* in *LDhelmet*. *LDhelmet* can only estimate the crossing over type of recombination (see e.g. Cromie and Smith (2007)), while *LDhat* contains also an option to estimate the rate of gene conversion. Here, our focus is on the estimation of the rate of crossing over.

2.2 Improved estimation of ρ

In order to improve the estimators of ρ introduced in subsection 2.1 we will optimize the trade-off between bias and variance. This is related to the statistical concept of shrinkage, see Gruber (1998), and Bayesian statistics. With shrinkage, bias is introduced for the sake of reducing the variance. If the gain in variance is larger than the loss due to additional bias, this leads to an improvement in terms of the MSE. A uniform improvement over the whole parameter range, however, can be achieved only under certain circumstances. A famous example is the James-Stein estimator of the mean of a multivariate normal distribution (Stein, 1956).

Bayes estimators on the other hand are constructed to minimize the weighed (with respect to the prior distribution) integral of an error measure such as the MSE.

It will turn out that our considered estimators are biased already. In order to optimize the trade-off between bias and variance, the required correction may therefore either lead to a decrease or an increase in bias, depending on the relative magnitudes of the two sources of error.

As no explicit formulas are available for the bias and the variance of composite likelihood estimators of ρ, we model bias and variance using regression based on simulated data. As will be shown in section 3.1, the following general model captures bias and variance of both $\hat{\rho}$ and $\breve{\rho}$ very accurately.

$$\mathrm{Bias}_{\rho}(\tilde{\rho}) = \gamma_1 \cdot \rho^2 + \beta_1 \cdot \rho + \alpha_1 \quad (2)$$

$$\mathrm{Var}_{\rho}(\tilde{\rho}) = \gamma_2 \cdot \rho^2 + \beta_2 \cdot \rho + \alpha_2 \quad (3)$$

We now investigate a generic rescaled estimator $\tilde{\rho}^* := c \cdot \tilde{\rho}$ with a positive constant c. Straightforward calculations lead to

$$\mathrm{Bias}_\rho(\tilde{\rho}^*) = c\left(\gamma_1\rho^2 + (\beta_1 + 1)\rho + \alpha_1\right) - \rho\,. \quad (4)$$

and

$$\mathrm{Var}_\rho(\tilde{\rho}^*) = c^2\left(\gamma_2\rho^2 + \beta_2\rho + \alpha_2\right). \quad (5)$$

Hence

$$\begin{aligned}\mathrm{MSE}_\rho(\tilde{\rho}^*) = {} & c^2\left(\gamma_1^2\rho^4 + 2\gamma_1(\beta_1 + 1)\rho^3 + (\gamma_2 + 2\gamma_1\alpha_1 + (b_1 + 1)^2)\rho^2 + (\beta_2 + 2(\beta_1 + 1)\alpha_1)\rho \right.\\ & \left. +(\alpha_2 + \alpha_1^2)\right) - 2c\left(\gamma_1\rho^3 + (b_1 + 1)\rho^2 + \alpha_1\rho\right) + \rho^2\,.\end{aligned} \quad (6)$$

In order to obtain an estimator that improves $\tilde{\rho}$, we minimize $\mathrm{MSE}_\rho(c \cdot \tilde{\rho})$ in c. This leads to

$$\begin{aligned} & c(\rho) \\ & = \frac{\gamma_1\rho^3 + (\beta_1 + 1)\rho^2 + \alpha_1\rho}{\gamma_1^2\rho^4 + 2\gamma_1(\beta_1 + 1)\rho^3 + \left(\gamma_2 + 2\gamma_1\alpha_1 + (\beta_1 + 1)^2\right)\rho^2 + (\beta_2 + 2(\beta_1 + 1)\alpha_1)\rho + \left(\alpha_2 + \alpha_1^2\right)}\,.\end{aligned} \quad (7)$$

This constant cannot directly be used for improving $\tilde{\rho}$ as it depends on the unknown ρ. One possible strategy would be to insert $\tilde{\rho}$ instead of ρ in (7). This approach worked reasonably well for Watterson's estimator of θ in Futschik and Gach (2008), but did not always lead to a uniformly improved estimator.

Alternatively, with S denoting the set of possible values of ρ (i.e. the parameter space), take $c^* = c(\rho^*)$ satisfying

$$\left|1 - c\left(\rho^*\right)\right| = \inf_{\rho \in S} |1 - c(\rho)|$$

as modifying constant with $\tilde{\rho}^*$. This will lead to a uniform improvement, if either $\sup_{\rho \in S} [c(\rho)] < 1$ or $\inf_{\rho \in S} [c(\rho)] > 1$. Otherwise we get $c(\rho^*) = 1$, and the original estimator remains unchanged, i.e. $\tilde{\rho}^* = \tilde{\rho}$.

3 Application to $\hat{\rho}$ and $\breve{\rho}$

In the following, we explore bias and variance of $\hat{\rho}$ and $\breve{\rho}$ and compare these estimators in terms of the MSE. Then we apply our method of improvement.

For the simulations concerning $\hat{\rho}$ and $\breve{\rho}$, we used the following simulation setup: For specified values of ρ, DNA sequence data was generated by the program *msms* (Ewing and Hermisson, 2010). The output of *msms* was transformed into fasta files via *ms2dna* (Haubold and Pfaffelhuber, 2013). For each of these fasta files, ρ was estimated by $\breve{\rho}$ or $\hat{\rho}$. Our analysis was then performed in *R* (R-Core-Team, 2013).

3.1 Variance, bias, and MSE of $\hat{\rho}$ and $\breve{\rho}$

Using extensive simulation runs, we explored variance and bias of $\hat{\rho}$ and $\breve{\rho}$. Figure 1 provides a typical example.

Using our simulations, figures 2 (a) and (b) show the MSE of $\hat{\rho}$ and $\breve{\rho}$ as functions of the true recombination rate ρ for various combinations of sample size (n), sequence length (l) and mutation rate (θ). For (a), with n=10, l=15001 bp, and θ=0.005/bp, $\breve{\rho}$ performs uniformly better than $\hat{\rho}$ in terms of the MSE, while under (b), where n=12, l=5001 bp, θ=0.005/bp, $\hat{\rho}$ outperforms $\breve{\rho}$ for almost all considered values of ρ.

Over a large range of configurations of the parameters n, l and θ, figure 3 provides an overall picture of the relative performance of $\hat{\rho}$ and $\breve{\rho}$ in terms of the MSE. For each scenario we considered 15 different true ρ values. The color coded score shows for how many of these 15 values the MSE of $\breve{\rho}$ is smaller than the MSE of $\hat{\rho}$. Apart from the situations where the sequence length is very short and at the same time θ is small, the MSE of $\breve{\rho}$ is smaller than the MSE of $\hat{\rho}$ for most or sometimes all considered values of ρ. Thus, for the scenarios we consider, $\breve{\rho}$ outperforms $\hat{\rho}$ in the majority of cases.

An estimated value $\hat{\theta}$ of the population mutation rate θ needs to be provided with $\hat{\rho}$ and $\breve{\rho}$. According to our observation, inaccuracies in $\hat{\theta}$ affect the estimators of ρ only slightly. Indeed, the differences in $MSE(\breve{\rho})$ and $MSE(\hat{\rho})$ tend to be negligible when using Watterson's estimator, compared to the improved version proposed in Futschik and Gach (2008).

3.2 Improving $\breve{\rho}$

Using regression with our simulated data, we estimated bias and variance of $\breve{\rho}$. As some of the estimated coefficients did not turn out to be significantly different from zero, we dropped the corresponding terms and simplified our models (4)–(6). This led to

$$\mathrm{Bias}_{\rho}(\breve{\rho}) = b_3 \cdot \rho + a_3 \quad (8)$$

$$\mathrm{Var}_{\rho}(\breve{\rho}) = c_4 \cdot \rho^2 \quad (9)$$

$$\mathrm{MSE}_{\rho}(\breve{\rho}) = \left(c_4 + b_3^2\right)\rho^2 + \left(2b_3a_3\right)\rho + a_3^2 \quad (10)$$

We first corrected for the constant bias by substracting the intercept a_3, resulting in the estimator $\breve{\rho}_2 = \breve{\rho} - a_3$. The optimal modifying constant for $\breve{\rho}_2$ turns then out to be

$$c_m = \frac{1 + b_3}{c_4 + \left(1 + b_3\right)^2}, \qquad (11)$$

which is independent of ρ. The approximate computation of c_m uses estimates for the regression coefficients in (8) and (9).

As an example, consider a model with θ=0.02/bp, n=10, l=15001 bp. Figure 4 (a) plots the MSE of the original estimator $\breve{\rho}$, as well as the improved version $\breve{\rho}*$. The improvement as percentage of ρ (shown in figure 4 (b)) is noticeable under this scenario.

With θ=0.02/bp, n=10, l=15001 bp, the improved MSE results from a large bias reduction. The variance increases, but to a smaller extent, see figure 5 (a). Here $c_m = 1.289$. For the parameters θ=0.005/bp, n=12, l=3001 bp $c_m = 0.816$, and the MSE is improved due to a reduction in the variance, while the bias increases, see figure 5 (b).

Figure 6 shows c_m (color coded) depending on the model parameters. Overall, the constant increases with θ and the sequence length l, and decreases with the sample size n.

The corresponding average improvement (over all considered values of ρ) achieved relative to the true value of ρ is presented in figure 7 (a). Figure 7 (b) shows the maximum relative improvement over ρ. In some cases the achieved gains are large. We observed such cases in particular when θ and the sequence length are large and the sample size is small.

Under some parameter combinations, the estimated shrinkage constants are nearly one, and there is not much room for uniform improvement. The original $\breve{\rho}$ and $\breve{\rho}*$ are then nearly identical, and the noise in the estimated regression coefficients may occasionally even lead to a marginal worsening. This could be avoided by setting $c_m = 1$, if its estimated value differs by less than ϵ from one, with ϵ denoting a bound on the simulation noise.

As it would be tedious to carry out a large amount of simulations to obtain modifying constants for each new model configuration, we fitted a regression model in order to quantify the dependence of the optimal modifying constant c_m and the bias correction term a_3 on the parameters sample size n, sequence length l, and mutation rate θ. By exploiting smoothness, this formula often (but not always) provides slightly more accurate estimates than those we obtained from individual simulations under single parameter combinations. This is since the smoothing reduces the random noise in the estimated coefficients. The following model provides a good fit.

$$\begin{aligned} c_m(\theta, n, l) = {} & 15.42\theta + 1.08 \cdot 10^{-1} n - 1.84 \cdot 10^{-3} n^2 + 8.52 \frac{1}{n} \\ & + 3.21 \cdot 10^{-5} l - 615.26 \frac{1}{l} - 1.41 \cdot 10^{-6} nl - 2.71 \cdot 10^{-1} n\theta \\ & - 3.74 \cdot 10^{-4} l\theta - 8.12 \cdot 10^{-1} \end{aligned} \qquad (12)$$

For a_3, we got

$$\begin{aligned} a_3(\theta, n, l) = {} & 2.93 \cdot 10^{-4} n - 6.31 \cdot 10^{-6} n^2 + 2.11 \cdot 10^{-2} \frac{1}{n} \\ & + 3.66 \cdot 10^{-8} l - 2.50 \cdot 10^{-6} l\theta - 4.86 \cdot 10^{-3} . \end{aligned} \qquad (13)$$

3.3 Improving $\hat{\rho}$

As with $\breve{\rho}$, there is also room for improving $\hat{\rho}$. Our simulated data suggest the following formulas, describing the dependence of bias, variance, and MSE on ρ.

$$\text{Bias}_\rho(\hat{\rho}) = c_1 \cdot \rho^2 + b_1 \cdot \rho \qquad (14)$$

$$\text{Var}_\rho(\hat{\rho}) = c_2 \cdot \rho^2 + b_2 \cdot \rho \qquad (15)$$

$$\text{MSE}_\rho(\hat{\rho}) = c_1^2 \rho^4 + 2c_1 b_1 \rho^3 + \left(b_1^2 + c_2\right)\rho^2 + b_2 \rho \qquad (16)$$

Since the nonzero coefficients differ for $\hat{\rho}$ and $\breve{\rho}$, the modifying constant has a different structure now:

$$c_m(\rho) = \frac{c_1 \rho^2 + \left(b_1 + 1\right)\rho}{c_1^2 \rho^3 + 2c_1\left(b_1 + 1\right)\rho^2 + \left(c_2 + b_1^2 + 2b_1 + 1\right)\rho + b_2} . \qquad (17)$$

Depending on ρ, $c_m(\rho)$ may take values both smaller and larger than one under some scenarios. In such situations we work with modifying constants $c_m(\hat{\rho})$. However, for small values of l this approach works less satisfactory.

We first consider again the scenario $\theta = 0.02$/bp, $n = 10$, and $l = 15001$ bp, using the same simulated data as with $\breve{\rho}$. In Figure 8, the MSE is shown both for $\hat{\rho}$ as well as for $c_m(\hat{\rho})\hat{\rho}$. Except for the smallest values of ρ, $MSE(c_m(\hat{\rho})\hat{\rho}) < MSE(\hat{\rho})$.

Not unexpectedly, the errors $MSE(c_m(\rho)\hat{\rho})$ would be even smaller with the theoretically optimal $c_m(\rho)$. But this does not help in practice, as the true ρ will be unknown.

Under the scenario $\theta = 0.008$/bp, $n = 7$, $l = 15001$ bp, the optimal modifying constant is monotonically increasing in ρ and always larger than one. When using the minimum of $c_m(\rho)$ over the considered range of ρ, we obtain a uniformly improved MSE. Figure 9 (a) displays $c_m(\rho)$ depending on ρ, and figure 9 (b) shows the MSE depending on ρ for the original and the improved estimator with $c_m = 1.158$, the optimal modifying constant for $\rho = 0.002$/bp.

4 Approximate Bayesian computation

Approximate Bayesian Computation (ABC) is a method to approximate the posterior distribution of one or more parameters of interest when no closed form expression is available for the likelihood. According to Bayes' rule it holds that

$$\mathbb{P}(\rho|D) = \frac{\mathbb{P}(D|\rho)\mathbb{P}(\rho)}{\mathbb{P}(D)}, \quad (18)$$

where $\mathbb{P}(\rho|D)$ is the posterior probability of the parameter ρ given the data D, $\mathbb{P}(D|\rho)$ is the likelihood, $\mathbb{P}(\rho)$ the prior, and $\mathbb{P}(D) = \int \mathbb{P}(D|\rho)\mathbb{P}(\rho)\, d\rho$. With ABC, a sample from an approximate posterior is simulated without directly using the likelihood. Instead, a sample is simulated under parameters randomly drawn from the prior distribution.

Parameter values that lead to simulated data close to the observed data D are taken as sample of the posterior distribution. The comparison of the simulated data sets with the observed one is carried out in terms of low dimensional but informative summary statistics. For our calculations we used the rejection algorithm of Pritchard et al. (1999), as well as the regression algorithm of Beaumont et al. (2002). Both algorithms are provided in the R-software package *abc* (Csillery et al., 2012). While the rejection algorithm is the most basic version of ABC, a regression correction of the accepted parameter values usually gives a better approximation to the posterior.

4.1 Our application of ABC

ABC is often used with easy to compute summary statistics for the unknown parameters. In Lopes et al. (2014) for instance, ρ (as well as θ and the non-synonymous synonymous rate ratio) is inferred from summary statistics like the number of segregating sites, moments of the heterozygosity, and several other measures. Here, we used only $\breve{\rho}$ as a single but sophisticated summary statistic. In a different context, the combination of ABC with a composite likelihood approach has been investigated by Ruli et al. (2015).

Bayesian estimators are known to minimize the posterior risk, which is in our case the integrated MSE weighted with the prior. Being an approximate approach, ABC may be expected to lead to estimators that are not too far from optimizing the posterior risk.

We noticed that the performance of ABC was slightly better when we used an equidistant grid of values of ρ instead of a sample from the (uniform) prior. This effect has been observed also in the context of Quasi - Monte Carlo methods, see e.g. Caflisch (1998). In this spirit, we took parameter values uniformly on a narrow equidistant grid, and generated data under these parameter values. We then used $\breve{\rho}$ on each data set to obtain simulated summary statistics. We used the same simulated data as in in section 3.1. In particular, we used 100 collections of fasta files for 141 equidistant values of ρ between 0.002/bp and 0.03/bp. As with cross-validation, each fasta file was once considered as the observed data set while the remaining fasta files were treated as a sample from the prior distribution. By iterating over all possible “real data” sets, we estimated bias and variance of the ABC posterior mean and median. Missing values were removed which led to slightly fewer than 100 simulations for some values of ρ. For our computations, we used the package *abc* in *R*.

4.2 Results for ABC

The regression algorithm outperformed the rejection algorithm (not shown here). After testing different tolerance levels, we decided on a tolerance level of 40 %, i.e. 40 % of the parameter values sampled from the prior have been accepted for the posterior. Figure 10 (a) shows the MSE depending on the true ρ for the original $\breve{\rho}$ estimator as well as for the ABC based estimator. The MSE as a proportion of the true value of ρ, i.e. the MSE divided by ρ^2, is displayed in figure 10 (b). With ABC, we obtain considerably improved MSE values when the true recombination rate is larger than approximately 0.015/bp, while for smaller recombination rates the MSE increases by a small amount. When measured as a proportion of ρ, this increase can be quite large, however, for small recombination rates. As ABC estimators, both posterior mean and posterior median gave quite similar results.

5 Example on real data

For ten haploid sequenced individuals of a *Drosophila melanogaster* population from Raleigh we looked at sequence data from the X chromosome. The data is available at http://pooldata.genetics.wisc.edu/dgrp_sequences.tar.bz2, http://johnpool.net/genomes.html. We considered sequentially 1000 pieces of 10Kb length and used $\breve{\rho}$ to estimate for each piece a constant recombination rate. For θ we used 0.008/bp, as in Chan et al. (2012), where ρ was estimated for the same *Drosophila* population.

We calculated the optimal modifying constant c_m and the constant term of the bias a_3 for the underlying parameters according to (12) and (13) and obtained $c_m = 1.13$, $a_3 = -2.83 \cdot 10^{-4}$. We substracted a_3 from each estimate and multiplied the result by c_m. As c_m is larger than 1, we increased the estimates by our method. In figure 11 (a) we show the original and the modified estimates for a range of values of ρ.

For understanding which accuracy can be expected, we show in figure 11 (b) the MSE plotted against the true value of ρ for θ=0.08/bp, n=10 and l=10 Kb.

The population recombination rate ρ is a parameter often needed in population genetic inference. More accurate estimates of ρ can therefore influence also the quality of estimation of other population genetic parameters, and may be beneficial for detecting signs of selection in population genetic data.

6 Discussion

We proposed an approach for improving composite likelihood estimators of ρ. In particular, we looked at versions of the composite likelihood method of Hudson (2001), as implemented in the software packages *LDhat* and *LDhelmet* ($\hat{\rho}$ and $\breve{\rho}$). As our simulations show, even these sophisticated widely used estimators still exhibit room for improvement with relatively little effort.

Although the rescaling factors used are not exact but estimated from simulations, our approach usually led to improved estimators, often considerably. Under some parameter configurations however, the original and the modified estimators were nearly identical. In such cases, the estimated rescaling constants was very close to one, and the estimation noise influenced whether a marginal improvement was seen or not.

In some cases the optimal rescaling factor c_m for $\hat{\rho}$ turned out to be both larger and smaller than one, depending on the unknown value of ρ. In such cases, we inserted $\hat{\rho}$ instead of ρ in the formula for c_m. Apart from very small values of ρ, this approach also led to improved estimators.

In order to apply our proposed rescaled estimator without having to carry out simulations, we present a formula for computing the modifying constant over a wide range of sample sizes, mutation rates and sequence lengths. We make such a formula also available for a sometimes helpful bias correction by an additive constant. Additionally we provide an *R* package on http://www.jku.at/ifas/content/e98868/employee_groups_wiss98976/employees144622/subdocs237646/content296458/ModifyMaxLkAndPairwise.zip where these formulas as well as $c_m(\hat{\rho})$ are implemented.

Notice that the MSE of the modified version of $\hat{\rho}$ is larger than that of the rescaled $\breve{\rho}$ in most cases. Averaged over the 15 different values of ρ, the rescaled $\breve{\rho}$ estimator outperformed the modified $\hat{\rho}$ estimator in 98.7 % of the scenarios. Thus, in general, we recommend the use of the rescaled $\breve{\rho}$ estimator $\breve{\rho}*$.

Additionally we presented a Bayesian approach based on ABC with $\breve{\rho}$ as summary statistic. The resulting estimator showed a reduced posterior risk with respect to the MSE.

We also applied our method to real data from a *Drosophila melanogaster* population (DGRP from Raleigh). To fit possible local variation in ρ, we divided the sequence into smaller intervals of equal length for which we estimated ρ separately.

In future work, we plan to derive a method to identify segments of constant recombination rates. There might be not only room for improving the estimators themselves, but also for improving the partitions.

Acknowledgments

We thank Christian Schlötterer and Claus Vogl for helpful comments.

The work has been carried out at the Vienna Graduate School of Population Genetics funded by the Austrian Science Fund (FWF): DK W1225-B20.

References

Arenas M, Lopes JS, Beaumont MA, et al. Codabc: A computational framework to coestimate recombination, substitution and molecular adaptation rates by approximate bayesian computation. Molecular biology and evolution. 2015; 32(4):1109–1112. [PubMed: 25577191]

Batorsky R, Kearney MF, Palmer SE, et al. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proceedings of the National Academy of Sciences. 2011; 108(14):5661–5666.

Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002; 162(4):2025–2035. [PubMed: 12524368]

BergerJO. Statistical decision theory and Bayesian analysisSpringer Science & Business Media; 2013

Caflisch RE. Monte carlo and quasi-monte carlo methods. Acta numerica. 1998; 7:1–49.

Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genet. 2012; 8(12):e1003090. [PubMed: 23284288]

Cromie GA, Smith GR. Branching out: meiotic recombination and its regulation. Trends in Cell Biology. 2007; 17(9):448–455. [PubMed: 17719784]

Csillery K, Francois O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC): *R package: abc*. Methods in Ecology and Evolution. 2012; 3(3):475–479.

Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010; 26(16):2064–2065. [PubMed: 20591904]

Fearnhead P, Donnelly P. Estimating recombination rates from population genetic data. Genetics. 2001; 159(3):1299–1318. [PubMed: 11729171]

Fearnhead P, Donnelly P. Approximate likelihood methods for estimating local recombination rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64(4):657–680.

Futschik A, Gach F. On the inadmissibility of Watterson's estimator. Theoretical Population Biology. 2008; 73(2):212–221. [PubMed: 18215409]

Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995; 82(4):711–732.

Griffiths RC, Marjoram P. Ancestral inference from samples of DNA sequences with recombination. Journal of Computational Biology. 1996; 3(4):479–502. [PubMed: 9018600]

GruberM. Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression EstimatorsVol. 156. CRC Press; 1998

HauboldB, , PfaffelhuberP. ms2dna, v. 1.16: Convert simulated haplotype data to DNA sequences. 2013Available at: http://guanine.evolbio.mpg.de/bioBox/

Hey J, Wakeley J. A coalescent estimator of the population recombination rate. Genetics. 1997; 145(3): 833–846. [PubMed: 9055092]

Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics. 1986; 38(6):226–231.

Hobolth A, Jensen JL. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theoretical Population Biology. 2014; 98:48–58. [PubMed: 24486389]

Hudson RR. Estimating the recombination parameter of a finite population model without selection. Genetical Research. 1987; 50(03):245–250. [PubMed: 3443297]

Hudson RR. Two-locus sampling distributions and their application. Genetics. 2001; 159(4):1805–1817. [PubMed: 11779816]

Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985; 111(1):147–164. [PubMed: 4029609]

Kuhner MK, Yamato J, Felsenstein J. Maximum likelihood estimation of recombination rates from population data. Genetics. 2000; 156(3):1393–1401. [PubMed: 11063710]

Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003; 165(4):2213–2233. [PubMed: 14704198]

Lin K, Futschik A, Li H. A fast estimate for the population recombination rate based on regression. Genetics. 2013; 194(2):473–484. [PubMed: 23589457]

Lopes JS, Arenas M, Posada D, et al. Coestimation of recombination, substitution and molecular adaptation rates by approximate Bayesian computation. Heredity. 2014; 112(3):255–264. [PubMed: 24149652]

Marjoram P, Wall JD. Fast ’coalescent’ simulation. BMC genetics. 2006; 7(1):16. [PubMed: 16539698]

McVeanG, , AutonA. LDhat 2.1: a package for the population genetic analysis of recombination. 2007Available at: http://www.stats.ox.ac.uk/˜mcvean/LDhat/manual.pdf

McVean G, Awadalla P, Fearnhead P. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics. 2002; 160(3):1231–1241. [PubMed: 11901136]

McVean GAT, Cardin NJ. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005; 360(1459):1387–1393.

Myers SR, Griffiths RC. Bounds on the minimum number of recombination events in a sample history. Genetics. 2003; 163(1):375–394. [PubMed: 12586723]

O’Reilly PF, Birney E, Balding DJ. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Research. 2008; 18(8):1304–1313. [PubMed: 18617692]

Pritchard JK, Seielstad MT, Perez-Lezaun A, et al. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution. 1999; 16(12):1791–1798. [PubMed: 10605120]

R-Core-TeamR: A Language and Environment for Statistical Computing2013

Ruli E, Sartori N, Ventura L. Approximate Bayesian computation with composite score functions. Statistics and Computing. 2015

Sabeti PC, Reich DE, Higgins JM, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002; 419(6909):832–837. [PubMed: 12397357]

Sabeti PC, Schaffner SF, Fry B, et al. Positive natural selection in the human lineage. Science. 2006; 312(5780):1614–1620. [PubMed: 16778047]

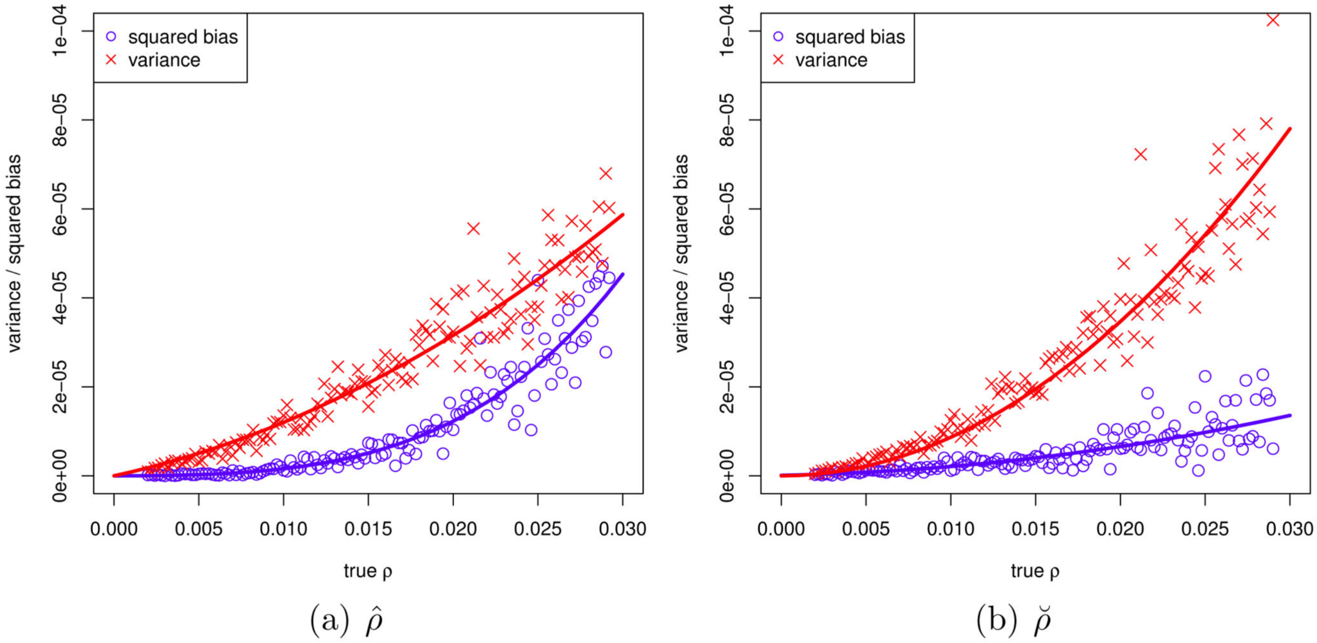
Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley symposium on mathematical statistics and probability. 1956; 1:197–206.

Varin C, Reid N, Firth D. An overview of composite likelihood methods. Statistica Sinica. 2011; 21(1):5–42.

Wall JD. A comparison of estimators of the population recombination rate. Molecular Biology and Evolution. 2000; 17(1):156–163. [PubMed: 10666715]

Wall JD. Estimating recombination rates using three-site likelihoods. Genetics. 2004; 167(3):1461–1473. [PubMed: 15280255]

Wiuf C. On the minimum number of topologies explaining a sample of DNA sequences. Theoretical Population Biology. 2002; 62(4):357–363. [PubMed: 12427459]

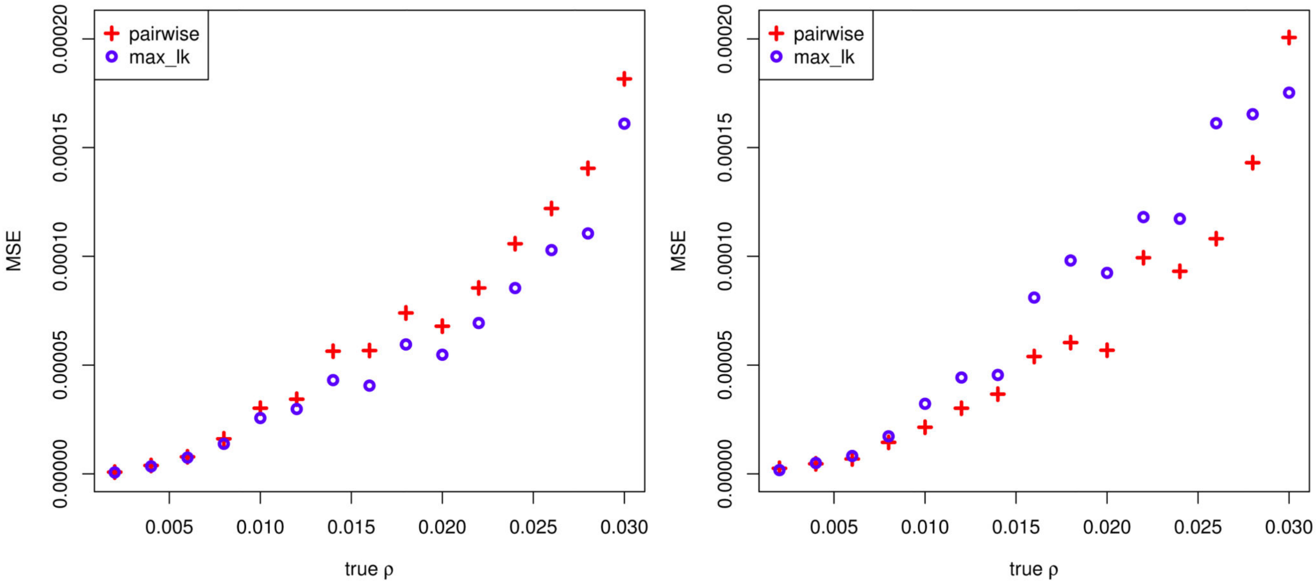


(a) $\hat{\rho}$

(b) $\breve{\rho}$

Figure 1.

Squared bias and variance of $\hat{\rho}$ and $\breve{\rho}$ in 1/bp^2; true ρ in 1/bp. Each plot symbol is based on 100 independent simulation runs (missing values that occured were removed), the curves display the resulting estimated regression relationships. Model parameters: θ=0.01/bp, n=20, l=5001 bp. Estimated regression coefficients (see equations (8), (9), (14), (15)): $b_1 = -7.67 \cdot 10^{-2}$, $c_1 = -4.92$, $b_2 = 8.24 \cdot 10^{-4}$, $c_2 = 3.78 \cdot 10^{-2}$; $a_3 = -3.31 \cdot 10^{-4}$, $b_3 = -1.12 \cdot 10^{-1}$, $c_4 = 8.67 \cdot 10^{-2}$.



(a) n=10, l=15001 bp, θ=0.005/bp

(b) n=12, l=5001 bp, θ=0.005/bp

Figure 2.

MSE of $\hat{\rho}$ (*pairwise*) and $\breve{\rho}$ (*max_lk*) for different values of ρ with different values of the parameters sample size (n), sequence length (l) and θ in (a) and (b); calculation of MSE from 50 independent simulations per value of ρ.

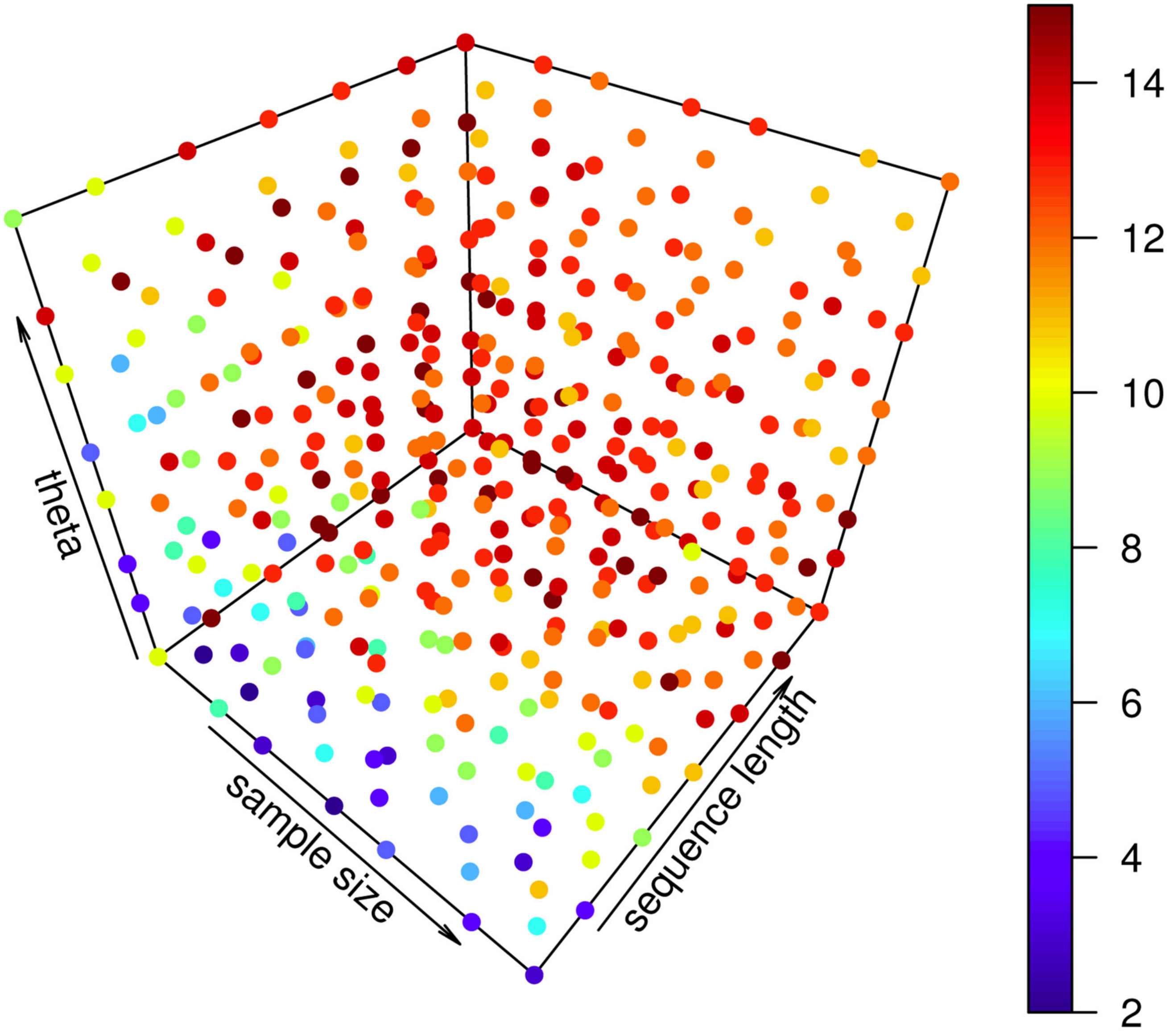
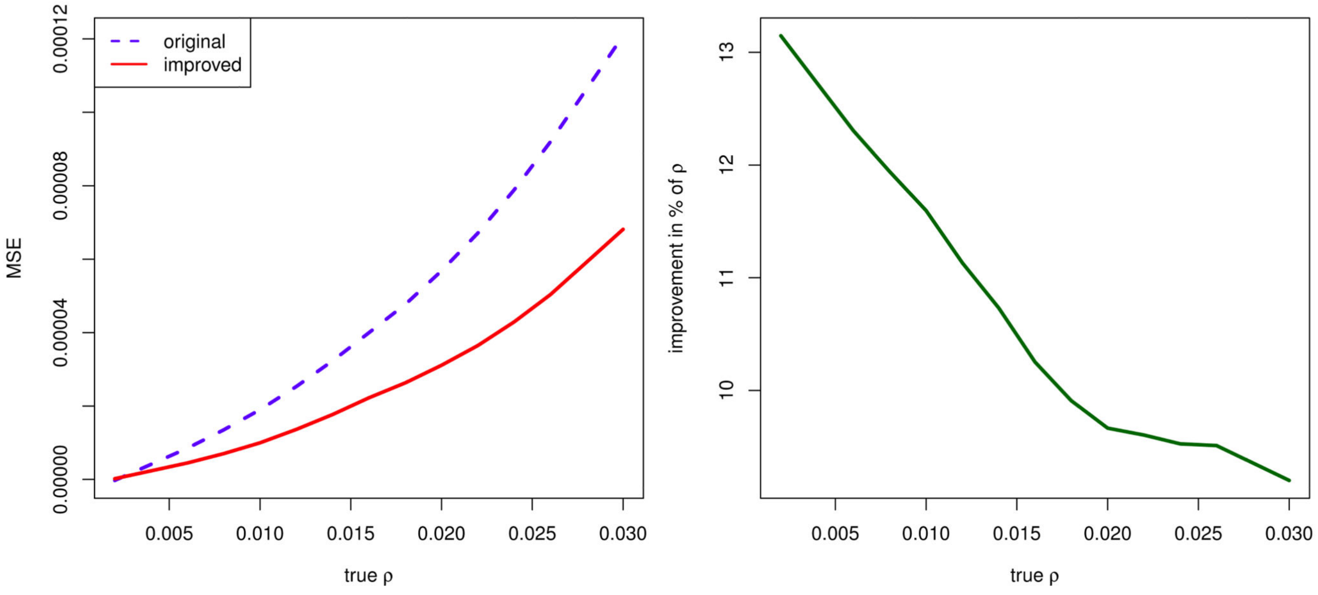


Figure 3.

Each dot displays the number of cases out of 15 values of $\rho \in [0.002, 0.03]$/bp, for which the MSE of $\breve{\rho}$ is smaller than that of $\hat{\rho}$. Parameter ranges: θ: (0.005/bp - 0.023/bp), n: (7 - 22), l: (3001 bp - 17501 bp); MSE estimated from 47 independent simulations per value of ρ.

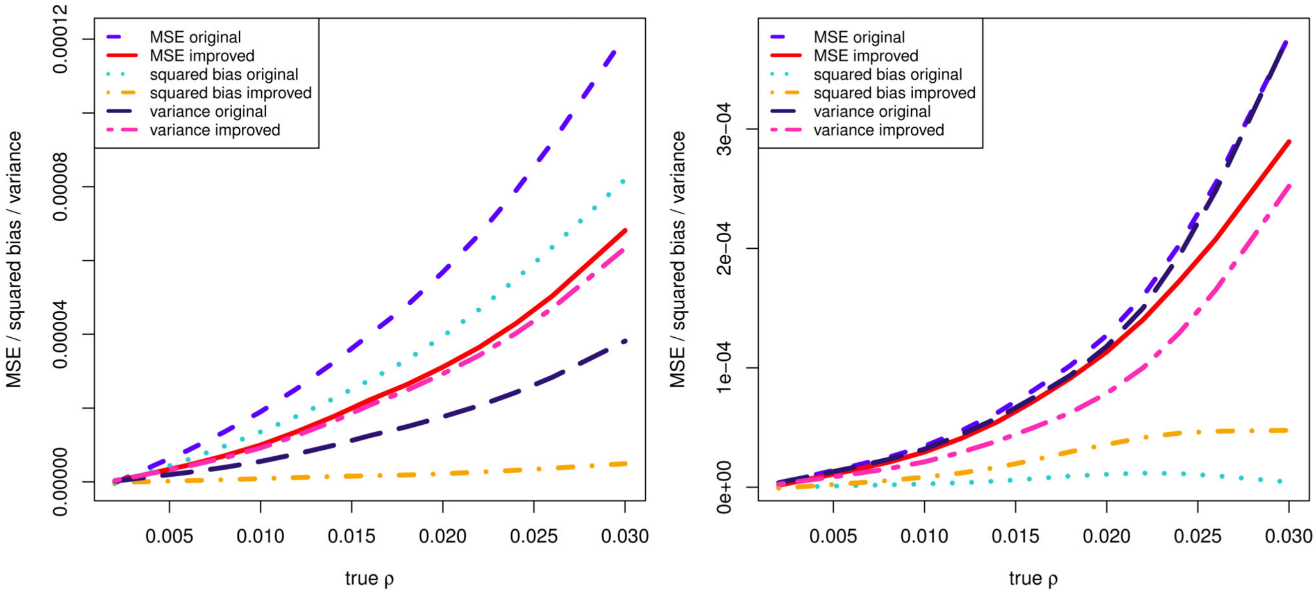


(a) MSE of $\breve{\rho}$ and its improved version

(b) Improvement as percentage of ρ

Figure 4.

(a): MSE of $\breve{\rho}$ (original and improved); (b) improvement as percentage of ρ. Parameters: θ=0.02/bp, n=10, l=15001 bp; results are based on 75 simulations per value of ρ for estimating c_m, and 75 independent simulations per value of ρ to obtain the MSE.



(a) θ=0.02/bp, n=10, l=15001 bp

(b) θ=0.005, n=12, l=3001 bp

Figure 5.

MSE, variance and squared bias in $(1/\text{bp})^2$ of original and improved estimators for different scenarios. True ρ in 1/bp. Calculation of modifying constant based on 75 simulations per value of true ρ, calculation of MSE, variance, and bias based on 75 different simulations per value of true ρ. (a) θ=0.02/bp, n=10, l=15001 bp. (b) θ=0.005/bp, n=12, l=3001 bp.

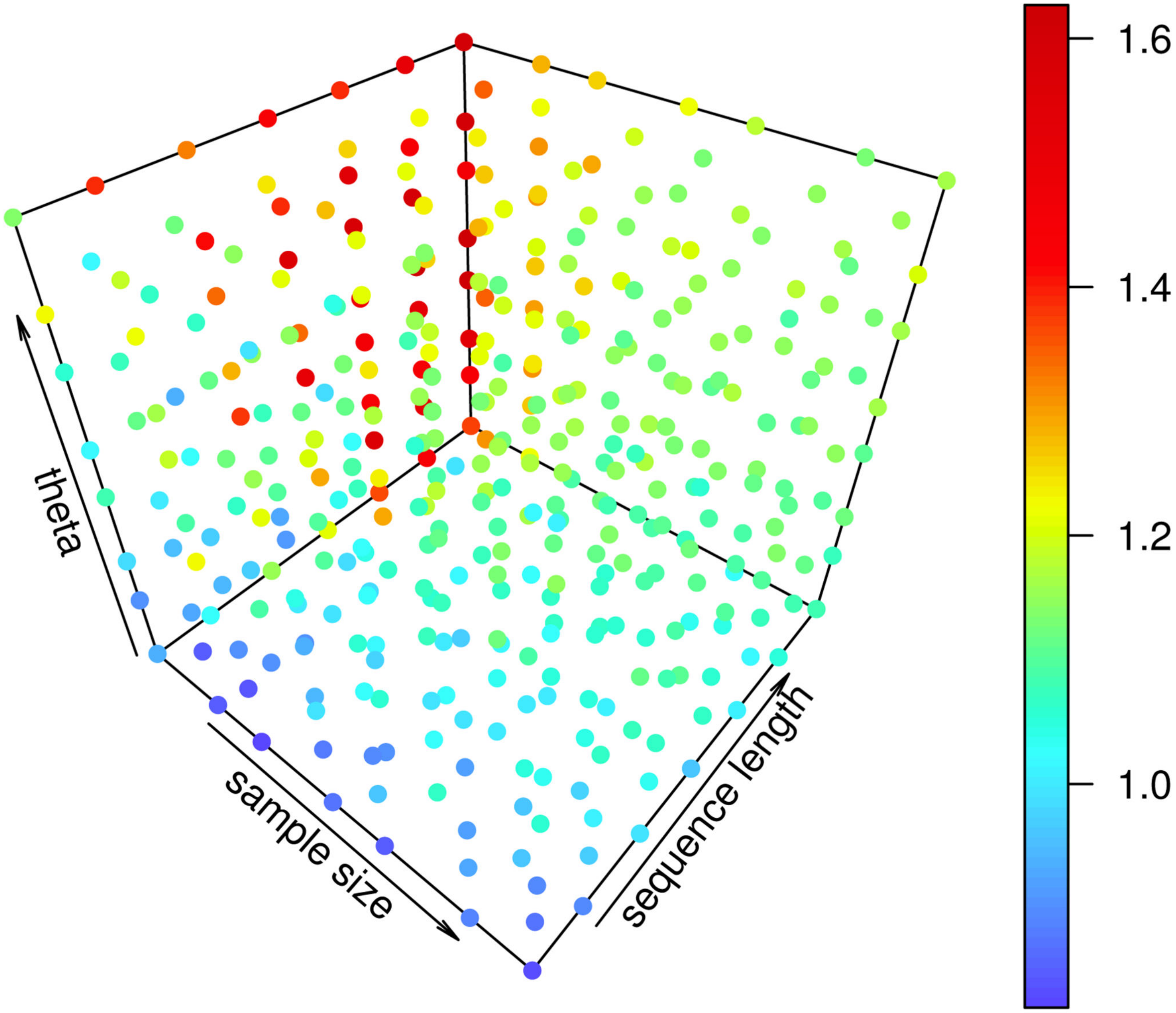
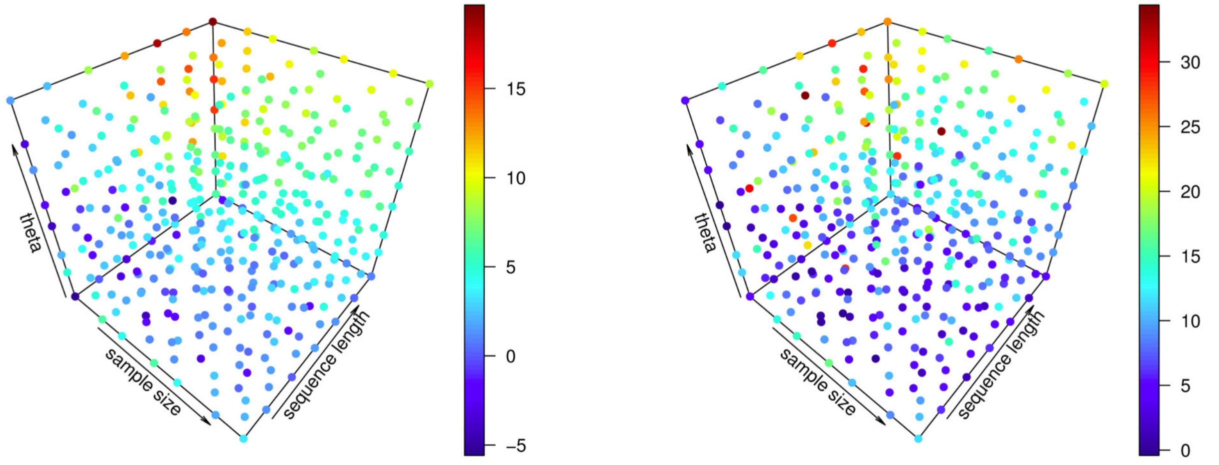


Figure 6.

Dependence of the optimal modifying constant (color coded) on the parameters θ (0.005/bp - 0.023/bp), n (7 - 22) and l (3001 bp - 17501 bp); calculation of MSE from 47 independent simulations per value of ρ.



(a) Amount of average relative improvement

(b) Amount of maximum relative improvement (over ρ)

Figure 7.

Amount of relative improvement averaged over ρ (a), and maximum relative improvement (b) of $\breve{\rho}$ in percent (color coded). The parameter ranges θ: (0.005/bp - 0.023/bp), n: (7 - 22), and l: (3001 bp - 17501 bp) were considered. Simulation effort: 24 simulations per value of ρ for calculating c_m, 23 simulations per value of ρ for the MSE estimates.

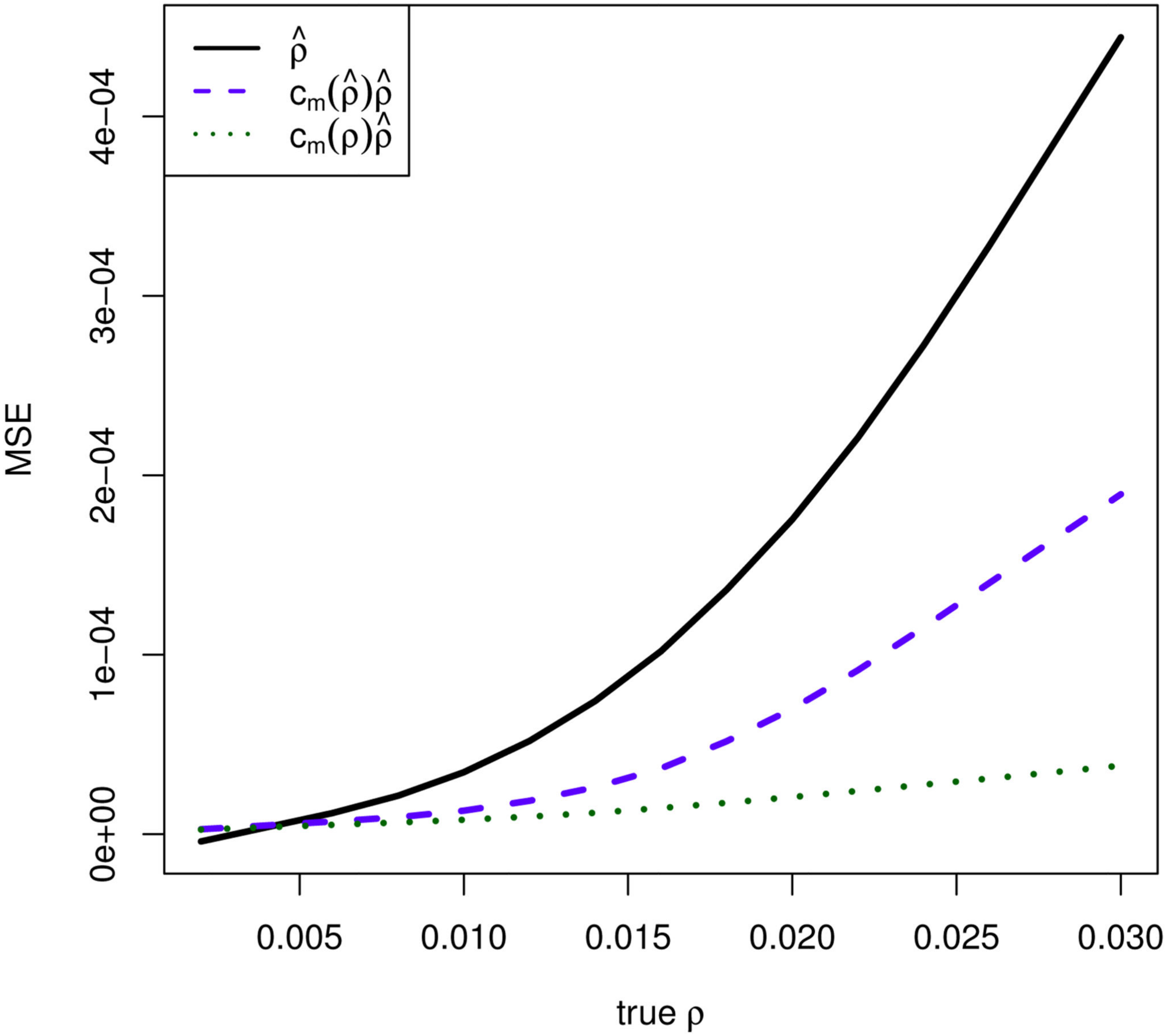
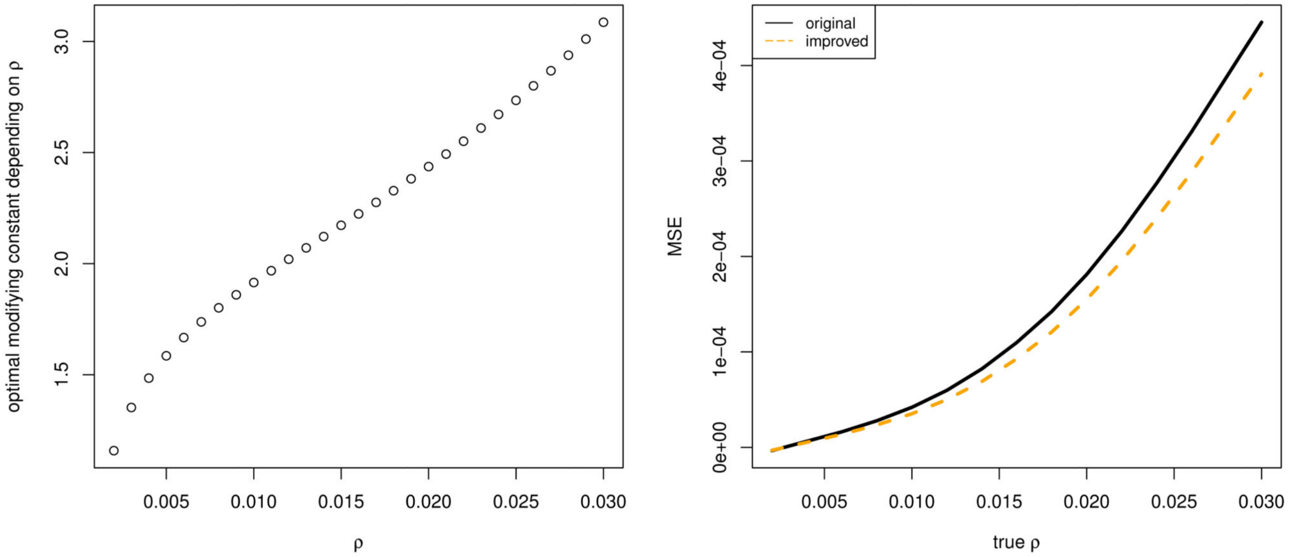


Figure 8.

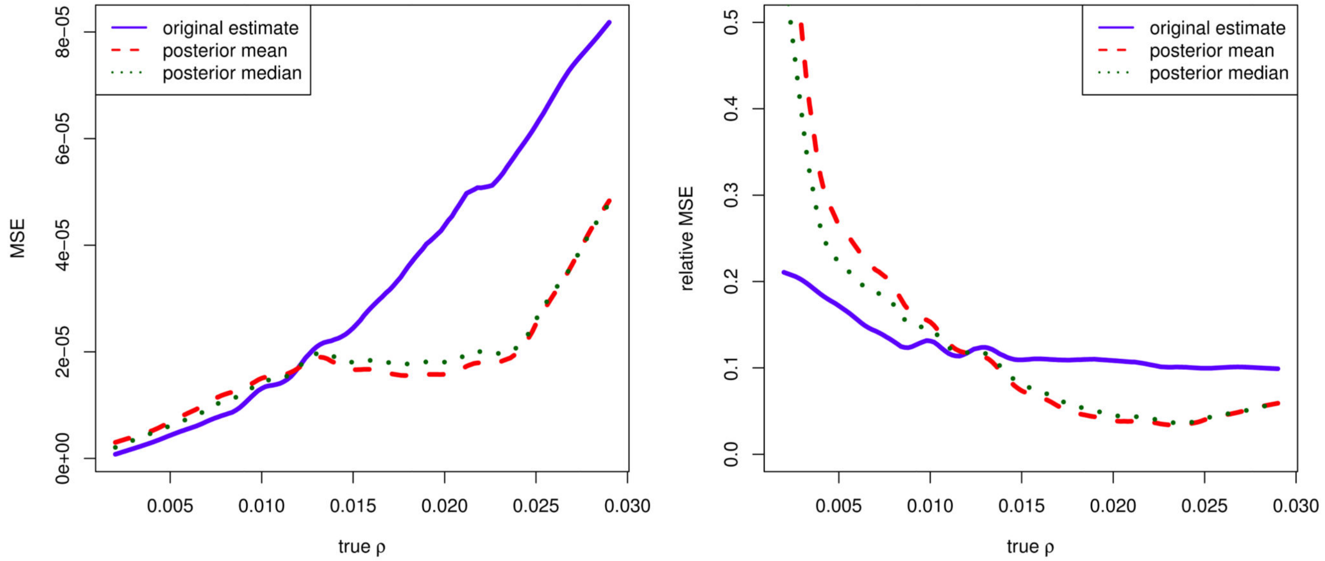
Dependence of MSE on ρ for $\hat{\rho}$, $c_m(\hat{\rho})\hat{\rho}$, and $c_m(\rho)\hat{\rho}$; $\theta = 0.02$/bp, $n = 10$, $l = 15001$ bp; 25 independent simulations per value of ρ for calculating the optimal modifying constant, 25 different independent simulations per value of ρ for calculation of the MSE.



(a) modifying constant depending on ρ (b) MSE of original and improved estimator

Figure 9.

$\theta = 0.008$/bp, $n = 7$, $l = 15001$ bp; 25 simulations per value of ρ for calculation of the modifying constant, 25 simulations per value of ρ for calculation of the MSEs. (a) Optimal modifying constant depending on ρ; ρ in 1/bp. (b) MSE in $(1/\text{bp})^2$ of original and improved $\hat{\rho}$ estimator with $c_m = 1.158$ for ρ in 1/bp.



(a) MSE

(b) MSE as proportion of ρ

Figure 10.

MSE of original and improved estimates in $(1/\text{bp})^2$ (a) and MSE divided by ρ^2 (b) for ρ in 1/bp. Calculation based on 100 simulations per value of ρ, tolerance of 40 % in ABC.

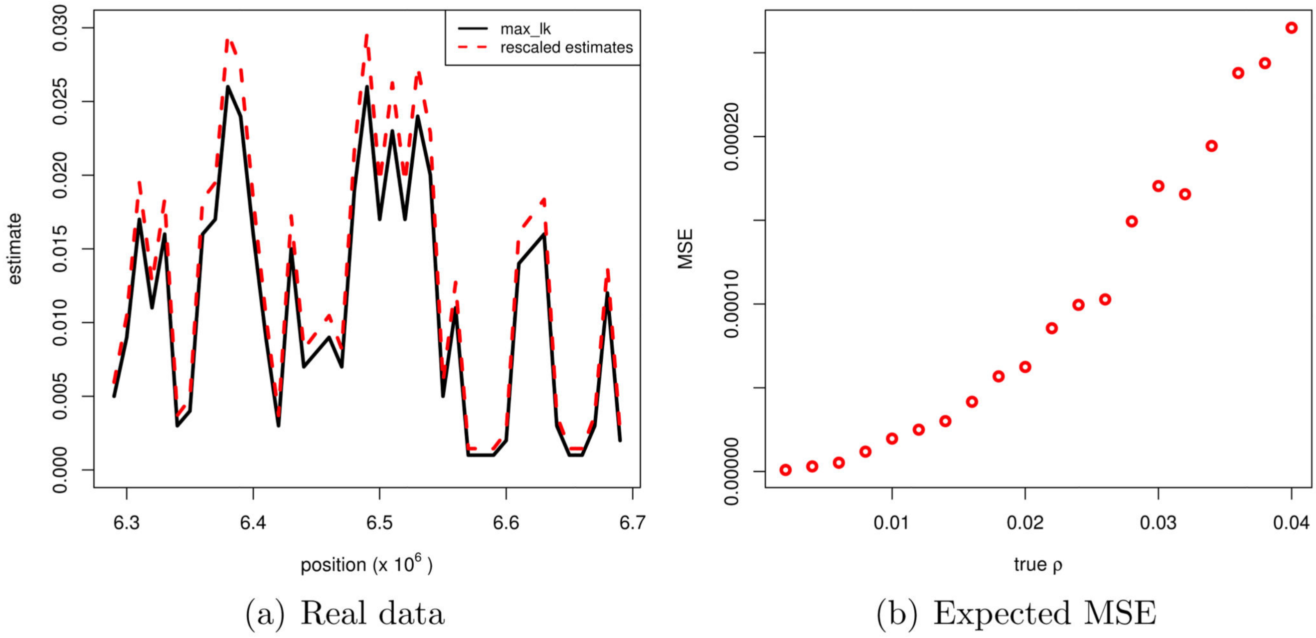


Figure 11.

(a) Original and rescaled estimates for a certain range of values of ρ; pieces of 10Kb length for sequence data of the X chromosome from a *Drosophila melanogaster* population (DGRP from Raleigh); 10 haploid individuals, θ=0.008/bp. (b) MSE in $(1/bp)^2$ against ρ in 1/bp; n=10, l=10000 bp, θ=0.008/bp, calculation based on 100 simulated values per value of true ρ.