On the inadmissibility of common population genetic estimates and their improvement by using shrinkage

Andreas Futschik

Dept. of Applied Statistics
JK University Linz

06/2016 MCEB

Contents in more detail

- *The scaled mutation parameter* θ and Watterson's estimate of θ
- *Inadmissability of Watterson's estimate* and a uniformly better estimate.
- *Estimating the scaled recombination parameter* — and improvements

Neutral Evolution and the Wright–Fisher Model

The Wright–Fisher assumptions

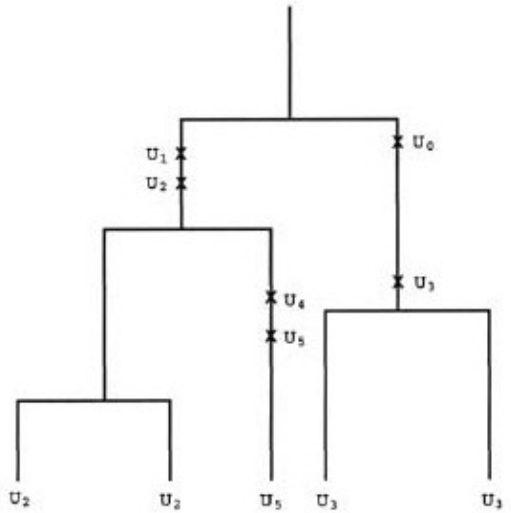
- Population of genes or chromosome segments of size N
- Constant population size N over time
- Random inheritance mechanism as described above

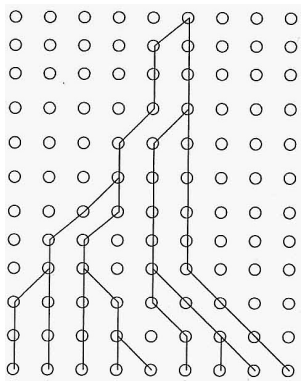
lead to the basic population genetic null model:
neutral evolution without demographic effects.

Coalescent Process

- Model the random genealogical history of a sample of n sequences.
- Look backward in time: Going one unit up in the coalescent corresponds to going back N generations in time.
- By letting $N \to \infty$, discrete Wright-Fisher genealogies can be approximated by time continuous coalescent process.

Figure 4.5. A coalescent tree for $n = 5$ with mutations





Basic Properties of the Coalescent under Wright–Fisher Model

- Time scale: $t = \text{generations}/N$.
- Let L denote the total length of the coalescent. Then
  - $P(L \leq t) = (1 - e^{-t/2})^{n-1}$
  - $E(L) = 2\sum_{j=1}^{n-1} \frac{1}{j} \approx 2\log(n) + 2C$.
  - $\text{Var}(L) = 4\sum_{j=1}^{n-1} \frac{1}{j^2} \approx \frac{2\pi^2}{3}$.

Estimating the scaled mutation parameter θ
Assumptions

- Assume that n nucleotide sequences are sampled from a population under *coalescent model*.
- We consider *infinite sites model*, i.e. each mutation takes place at a new location and therefore generates a new segregating site.

Estimating the scaled mutation parameter θ

- Want to estimate *scaled mutation parameter*

$$\theta = 2N\mu,$$

where μ denotes the individual mutation rate per generation.
(Notice: $\theta = 4N\mu$ if N defined as number of diploid individuals.)

- θ is twice the expected number of mutations occurring during one time unit on the coalescent tree.
- Typically $\mu \approx 10^{-9} \times b$, with b being the sequence length (number of bases).
- For Drosophila $N \approx 10^6$.

Estimation problem from statistical point of view

- We observe Poisson process with rate $\theta/2$ on time interval $[0, L]$.
- Only probability distribution but not the value of L is known.
- We observe S events and want to estimate θ.

Watterson's estimate

- Watterson (1974), Ewens (1973)
- Let S denote total number of mutations that occurred on the coalescent
- Watterson's estimate:

$$\hat{\theta}_W = 2S/\,\mathrm{E}(L).$$

- Watterson (1974) cited $>$ 1250 times.

Properties of Watterson's estimate

- Watterson's estimate is unbiased:

$$\mathrm{E}_\theta \, \hat{\theta}_W = \theta$$

- $\mathrm{MSE}_\theta(\hat{\theta}_W) = \mathrm{Var}_\theta(\hat{\theta}_W) = \frac{2}{\mathrm{E}(L)} \, \theta + \frac{\mathrm{Var}(L)}{[\mathrm{E}(L)]^2} \, \theta^2.$

Further properties of Watterson's estimate

- $\hat{\theta}_W$ is consistent.
- $\hat{\theta}_W$ is asymptotically normal distributed (Klein et al. (1999))
- As $n \to \infty$, $\hat{\theta}_W$ achieves asymptotic variance of the maximum likelihood estimator (MLE) calculated in the idealized situation where the number of mutations is known for each branch of the coalescent tree. (Fu and Li (1993).)

But Watterson's estimate can be improved using shrinkage

Theorem

- Watterson's estimator $\hat{\theta}_W$ is inadmissable
- There is a shrinkage estimate whose MSE is below that of $\hat{\theta}_W$ for all $0 < \theta < \infty$.
- This shrinkage estimate cannot be uniformly improved by any other estimate linear in S (i.e. of the form $a_n S + b_n$).

(F & Gach (2008).)

Classical example of improvement by shrinkage: *James* and *Stein* (1961) estimate.

How to obtain a uniformly better estimator?

- Minimize the MSE with respect to normalization factor c:

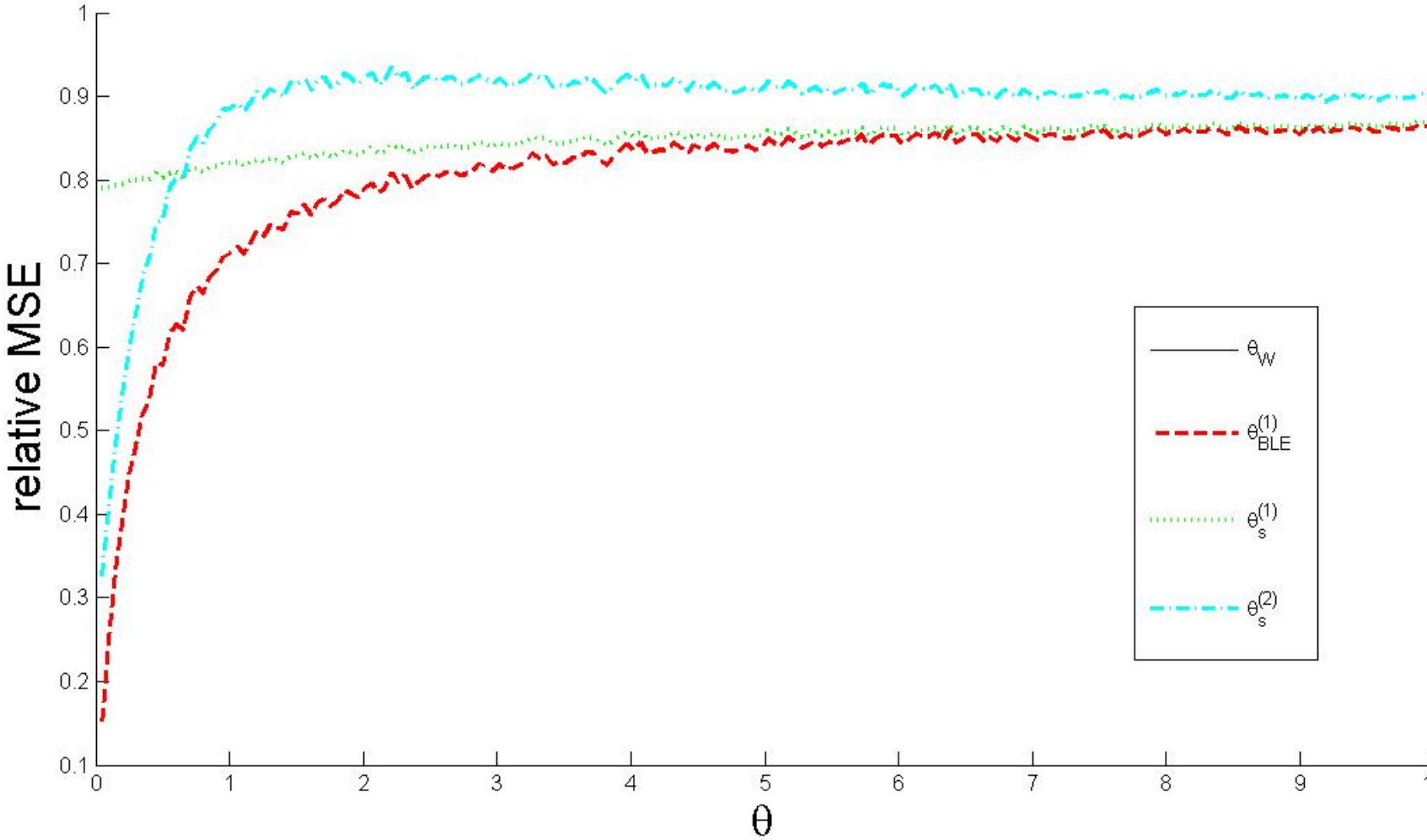
$$\min_c \mathrm{MSE}_\theta(c\hat{\theta}_W) = \min_c \mathrm{E}_\theta \left(c\frac{2S}{\mathrm{E}(L)} - \theta \right)^2$$

- Solution depends on θ:

$$\tilde{\theta} = \frac{2S}{\mathrm{E}(L) + \frac{\mathrm{Var}(L)}{\mathrm{E}(L)} + \frac{2}{\theta}}$$

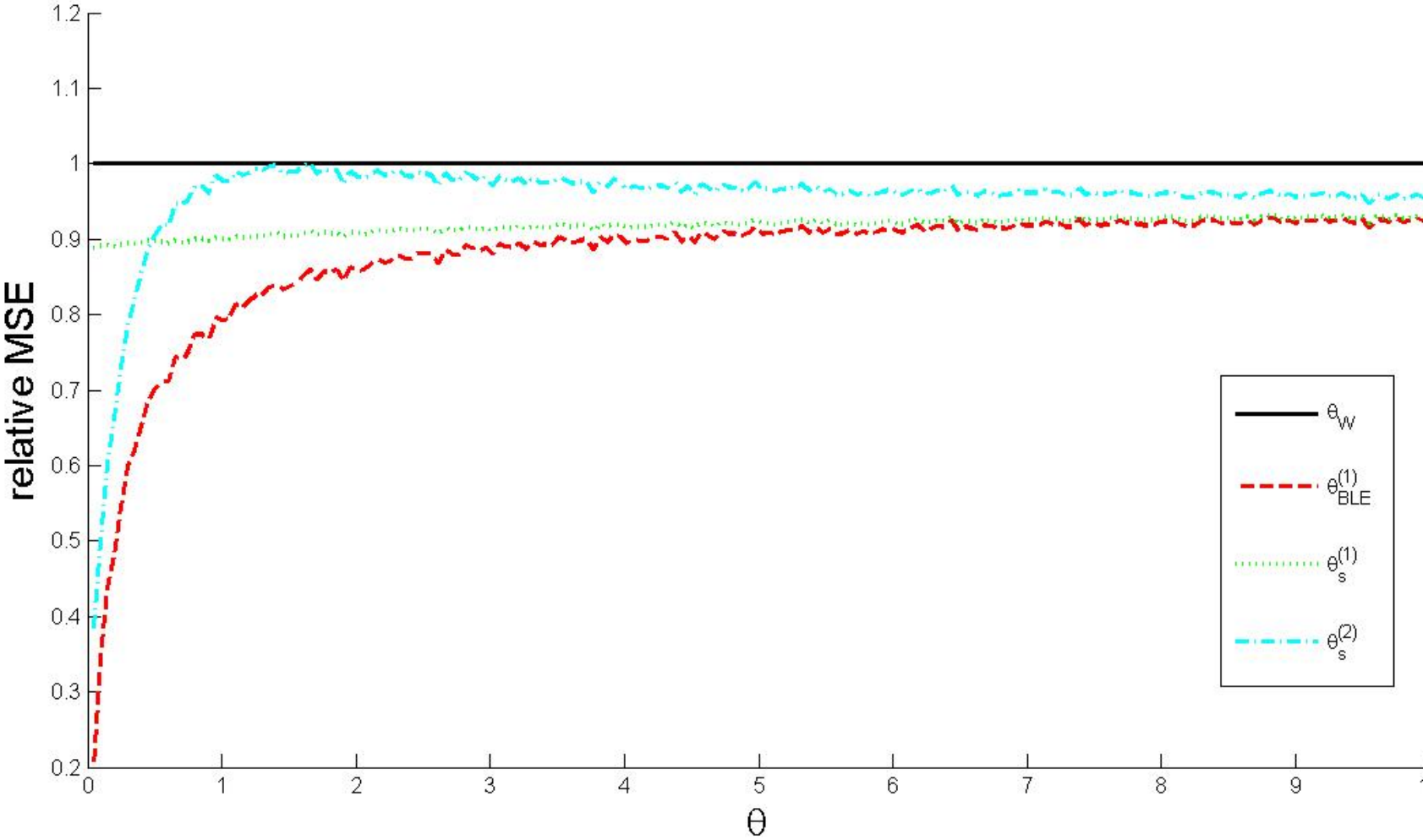
- Two possible approaches:
  - $\hat{\theta}_s^{(1)} := \frac{2S}{\mathrm{E}(L) + \frac{\mathrm{Var}(L)}{\mathrm{E}(L)}}$
  - $\hat{\theta}_s^{(2)} := \frac{2S}{\mathrm{E}(L_n) + \frac{\mathrm{Var}(L_n)}{\mathrm{E}(L_n)} + \frac{2}{\hat{\theta}_W}}$.

How large is the gain? (MSE relative to Watterson)



$n = 20$

MSE relative to Watterson



$n = 100$

Does this approach work more generally?

Lemma

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

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

with $a, b \geq 0$.

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

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

  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

$$\hat{\theta}_s := \frac{\hat{\theta}}{b+1}.$$

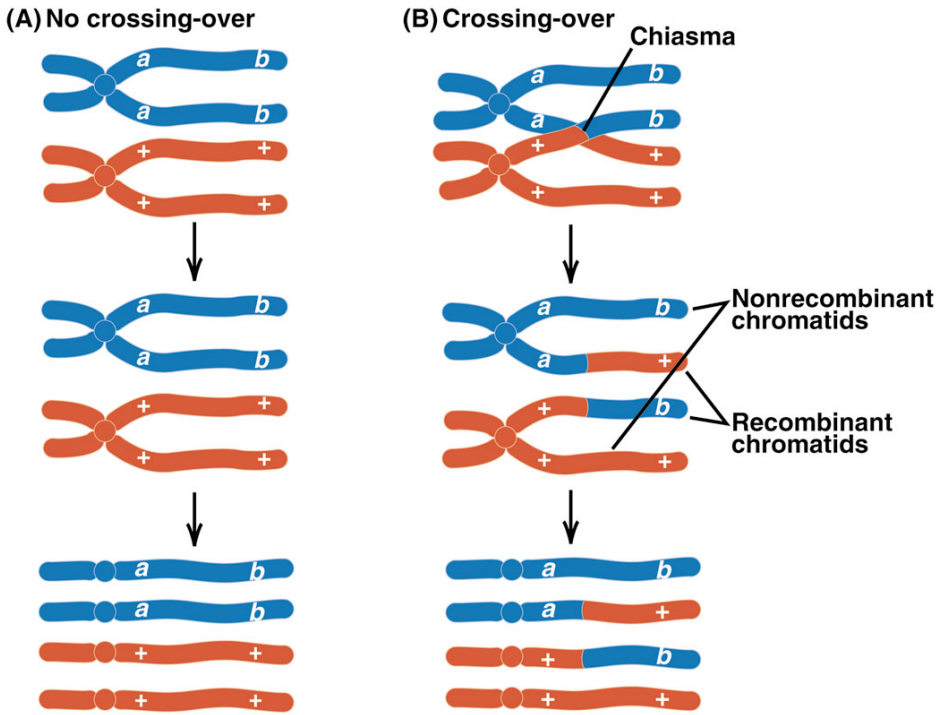
Application to other Estimators

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)	$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2\eta_i$	$\theta + \frac{2[36n^2(2n+1)\sum_{i=1}^{n} i^{-2} - 116n^3 + 9n^2 + 2n - 3]}{9n(n-1)^2}\theta^2$

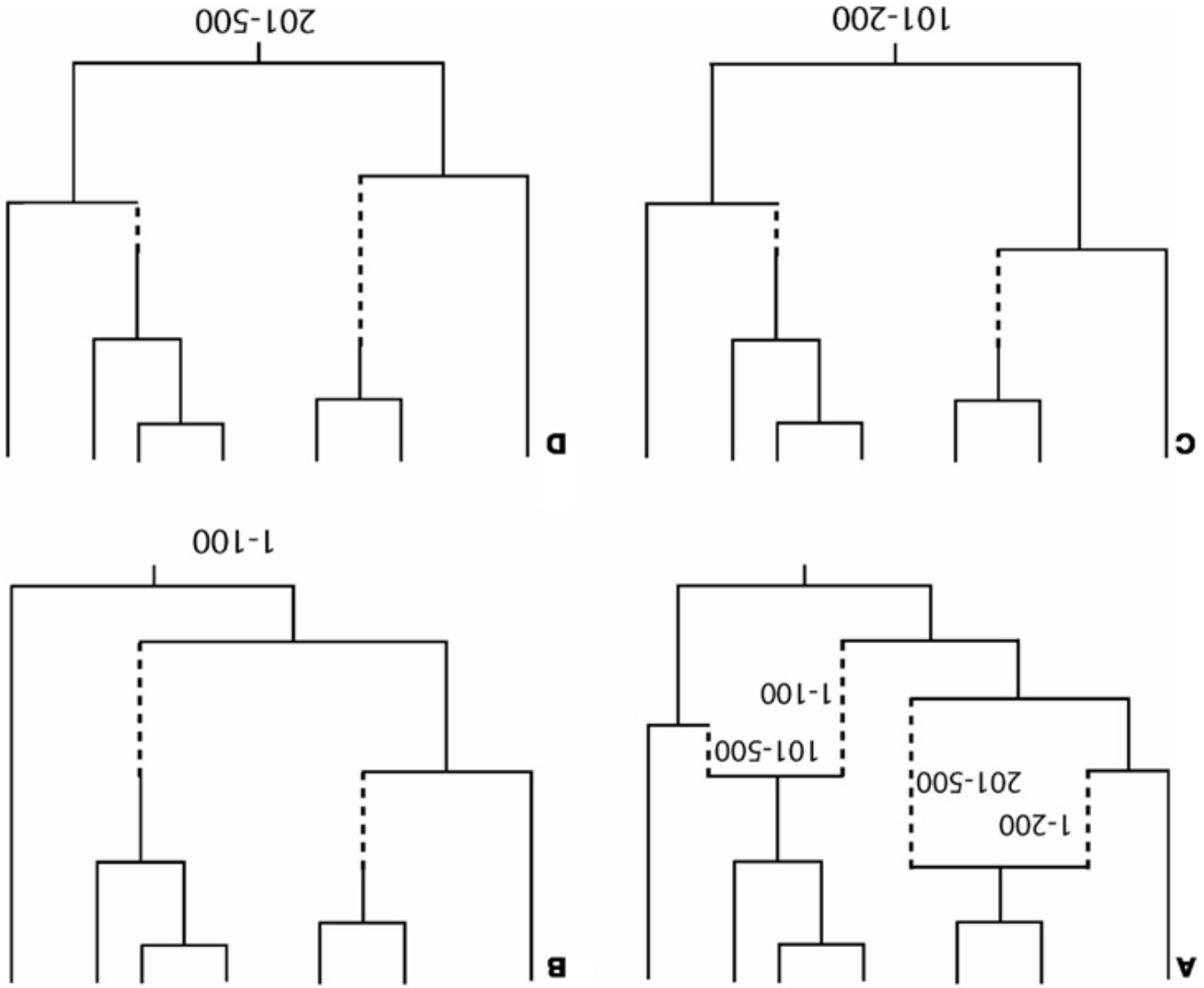
Here

- η_i . . . number of sites where the mutant allele is present i times in sample of size n
- $c_n := \sum_{i=1}^{n-1} i^{-1}$

Recombination



Ancestral Recombination Graph



Why recombination is more complicated

- Recombination is more difficult to estimate than scaled mutation parameter
- Not all recombination events leave traces in data
- 4-gamete test: two loci A, B with two genotypes G_1, G_2:

	locus	
genotype	L_1	L_2
G_1	n_{11}	n_{12}
G_2	n_{21}	n_{22}

- Sufficient but not necessary condition for recombination: $n_{ij} > 0, \forall i, j$.

Composite Likelihood Approach

- Summarize: $D_{ij} = \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$
- Composite likelihood:

$$\prod_{i<j} p(D_{ij})$$

- Maximum composite likelihood estimate of Hudson (2001)
- Software implementations (incuding generalizations & improvements):
  - LDhelmet (Chan, Jenkins, Song), 2012)
  - LDhat (McVean and Auton, 2007)

Does shrinkage also work for recombination?

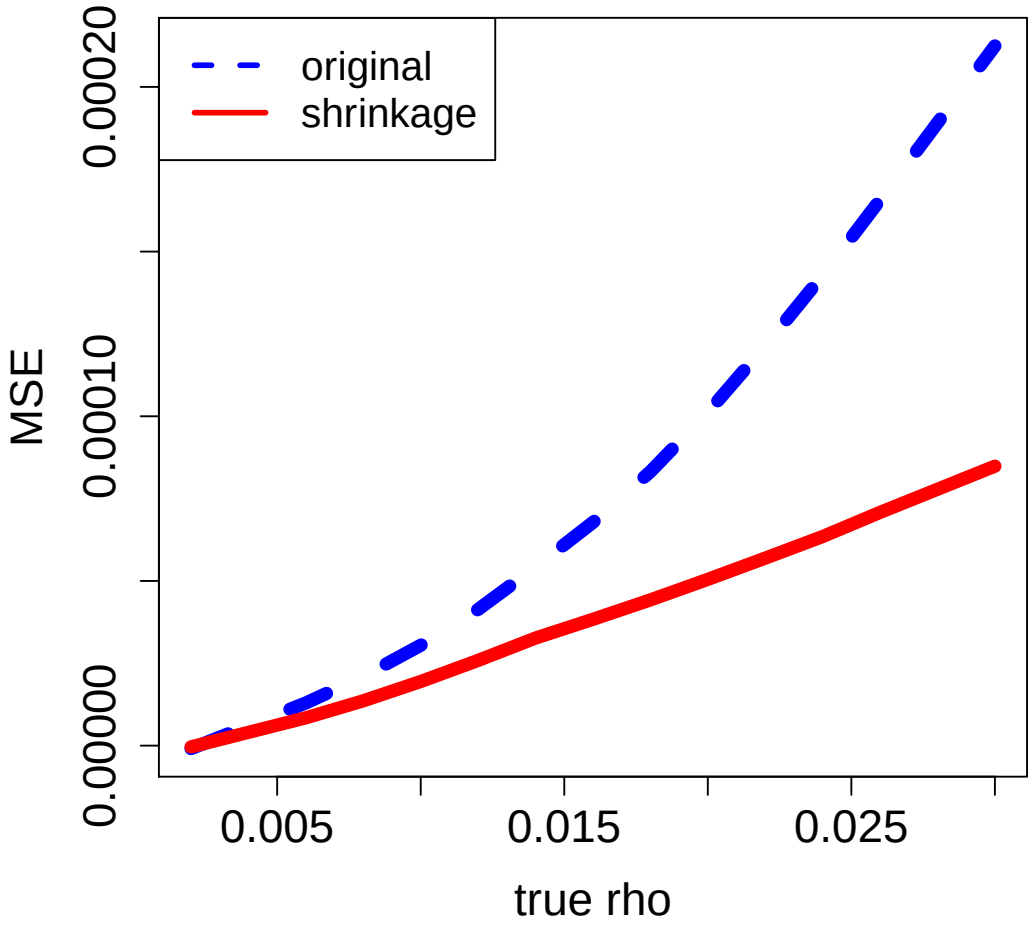
- Problem: No explicit formulas for bias and variance available in recombination context.
- Estimate relationship by using regression

$$\text{Bias}(\hat{\rho}) = a_1 + b_1\rho$$

$$\text{Var}(\hat{\rho}) = a_2 + b_2\rho + b_3\rho^2$$

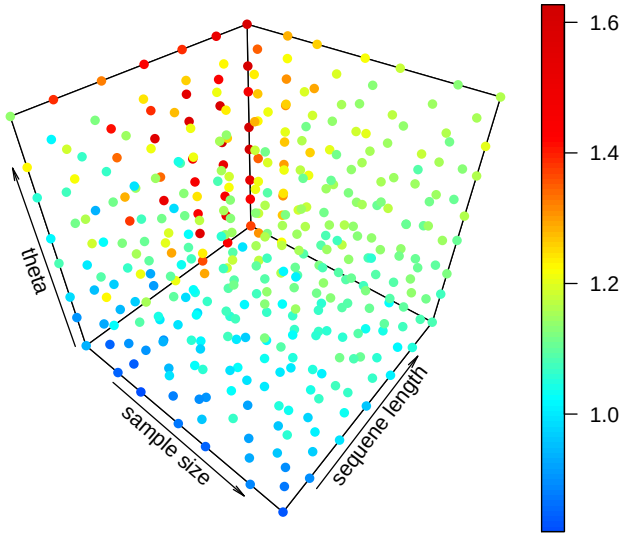
(Gärtner & F 2016)

Gain compared to LDhelmet



$n = 20$, rho per base pair.

Dependence of Optimum Shrinkage Constants on Model Parameters



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.

Summary

- Even nowadays, frequently used estimators are sometimes inadmissible!
- Good news: improvement is possible and fairly straightforward.

Acknowledgement