Approximate Inference for High Dimensional Population Genetic Models using Stochastic Gradient Methods

Johanna Bertl[1,2] Greg Ewing[1] Andreas Futschik[1,2]

[1]University of Vienna

[2]Vienna Graduate School of Population Genetics

Coalescent parameter estimation

In the coalescent, the likelihood of a parameter vector Θ given data D_{obs}, can be written as

$$L(\Theta; D_{obs}) = P(D_{obs} \mid \Theta) = \sum_{T \in \mathcal{T}} P(D_{obs} \mid T, \Theta) P(T \mid \Theta),$$

where $\mathcal{T}$ is the space of all possible coalescent trees (Stephens, 2001).

$\Rightarrow$ approximate inference methods (MCMC, IS, ABC)

Approximate maximum likelihood estimation

Approximate maximum likelihood estimation in coalescent models:

- Estimating the likelihood with MCMC through tree space and IS (Kuhner, Yamato and Felsenstein, 1995)
- IS in the Griffith-Tavaré recursions (Griffiths and Tavaré, 1994)

Our approach: Approximate maximum likelihood estimation by a stochastic gradient algorithm

Gradient algorithms

Maximization problem: Let $L : \mathbb{R}^p \mapsto \mathbb{R}$ be a function and ∇L its gradient. Find

$$\arg\max_{\Theta} L(\Theta).$$

Gradient algorithm:

$$\Theta_k = \Theta_{k-1} + a_k \nabla L(\Theta_{k-1})$$

Gradient algorithms

Maximization problem: Let $L : \mathbb{R}^p \mapsto \mathbb{R}$ be a function and ∇L its gradient. Find

$$\arg\max_{\Theta} L(\Theta).$$

Problem: L and ∇L are unknown, but L can be estimated for any $\Theta \in \mathbb{R}^p$

Gradient algorithms

Maximization problem: Let $L : \mathbb{R}^p \mapsto \mathbb{R}$ be a function and ∇L its gradient. Find

$$\arg\max_{\Theta} L(\Theta).$$

Problem: L and ∇L are unknown, but L can be estimated for any $\Theta \in \mathbb{R}^p$

Stochastic gradient algorithm (Kiefer and Wolfowitz, 1952):

$$\Theta_k = \Theta_{k-1} + a_k \hat{\nabla}_{c_k} \hat{L}(\Theta_{k-1})$$

with

$$\left(\hat{\nabla}_{c_k} \hat{L}(\Theta_k)\right)^{(i)} = \frac{\hat{L}(\Theta_k + c_k \mathbb{1}_i) - \hat{L}(\Theta_k - c_k \mathbb{1}_i)}{2c_k}$$

for $i = 1, \ldots, p$.

Estimation of $L(\Theta; D_{obs})$ in the coalescent

- Use $L(\Theta; S_{obs})$ instead of $L(\Theta; D_{obs})$.

Estimation of $L(\Theta; D_{obs})$ in the coalescent

- Use $L(\Theta; S_{obs})$ instead of $L(\Theta; D_{obs})$.
- Estimation of $L(\Theta; S_{obs})$ with simulations and kernel density estimation:
  1. Simulate m datasets D_j^*, $j = 1, \ldots, m$ from the coalescent with parameter Θ.
  2. Compute the summary statistics S_j^* from D_j^*, $j = 1, \ldots, m$.
  3. Estimate $L(\Theta; s) = p(s \mid \Theta)$ with multivariate kernel density estimation:

$$\hat{L}(\Theta; s) = \hat{p}(s \mid \Theta) = \frac{1}{m\sqrt{\det(H)}} \sum_{j=1}^{m} \kappa\left(H^{-1/2}(s - S_j)\right)$$

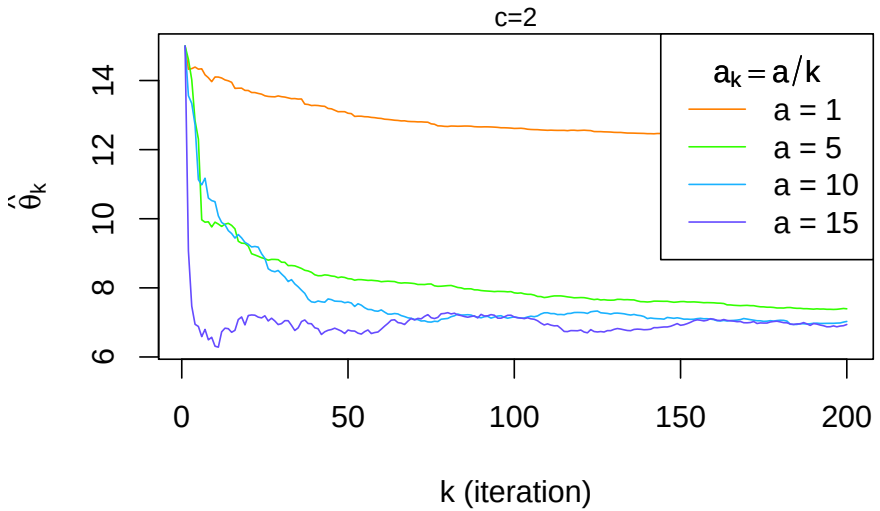
  4. Evaluate $\hat{L}(\Theta; s)$ at S_{obs}.

Illustration: neutral coalescent without demography

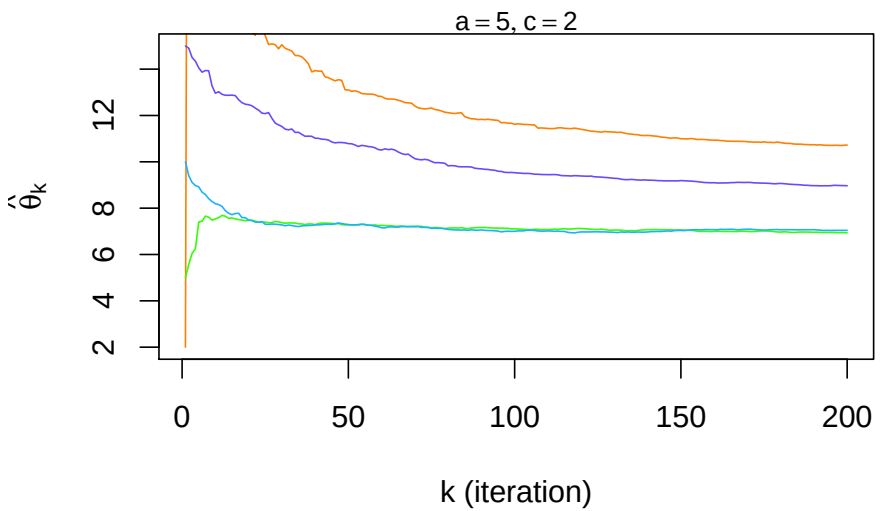
- Model parameter: scaled mutation rate $\theta = 5$
- Constant population size
- Sample size: $n = 20$
- Infinite sites model
- Summary statistic: number of segregating sites S

$\theta = 5, S = 23, n = 20, \hat{\theta}_W = 6.48$

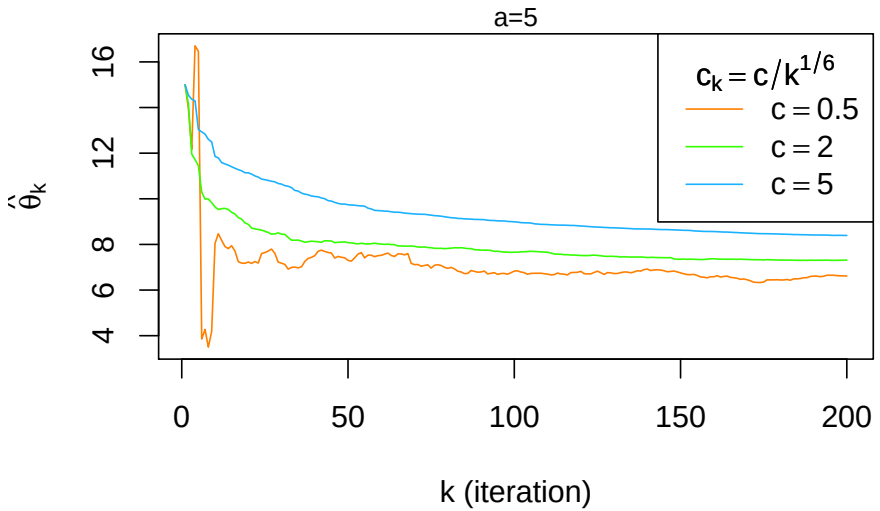
Step size (a_k)



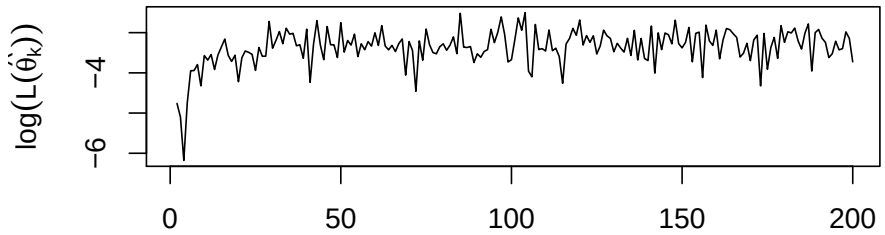
Different starting points



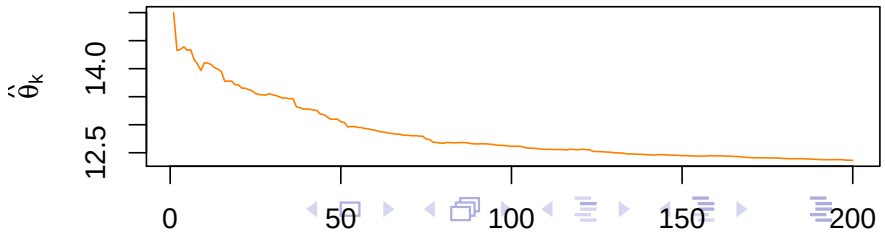
Finite differences (c_k)



Simulated log−likelihood



Estimation sequence



Gain sequences a_k and c_k

Step size	**Finite differences**
$a_k = \frac{a}{(k+A)^\alpha}$	$c_k = \frac{c}{k^\gamma}$

Heuristic (Spall, 2003)

- Set $\alpha = 1$ and $\gamma = 1/6$.
- Set $A \approx 0.1K$.
- Set $c = \widehat{sd}(\log \hat{L}(\Theta_0; S_{obs}))$.
- Set $b_i, i = 1, \ldots, p$ as the desired step size in step 1.
  1. Set
  $$a_{temp,i} = \frac{b_i(A+1)^\alpha}{\left(\hat{\nabla}_{c_0} \log \hat{L}(\Theta_0; S_{obs})\right)^{(i)}}.$$
  2. Set $a = \min\{a_{temp,1}, \ldots, a_{temp,p}\}$.

Good starting values

- G random starting points $\Theta_{0,1}, \ldots, \Theta_{0,G}$
- Compute improved starting points $\Theta^*_{0,1}, \ldots, \Theta^*_{0,G}$ by a simplified stochastic gradient algorithm that approximates

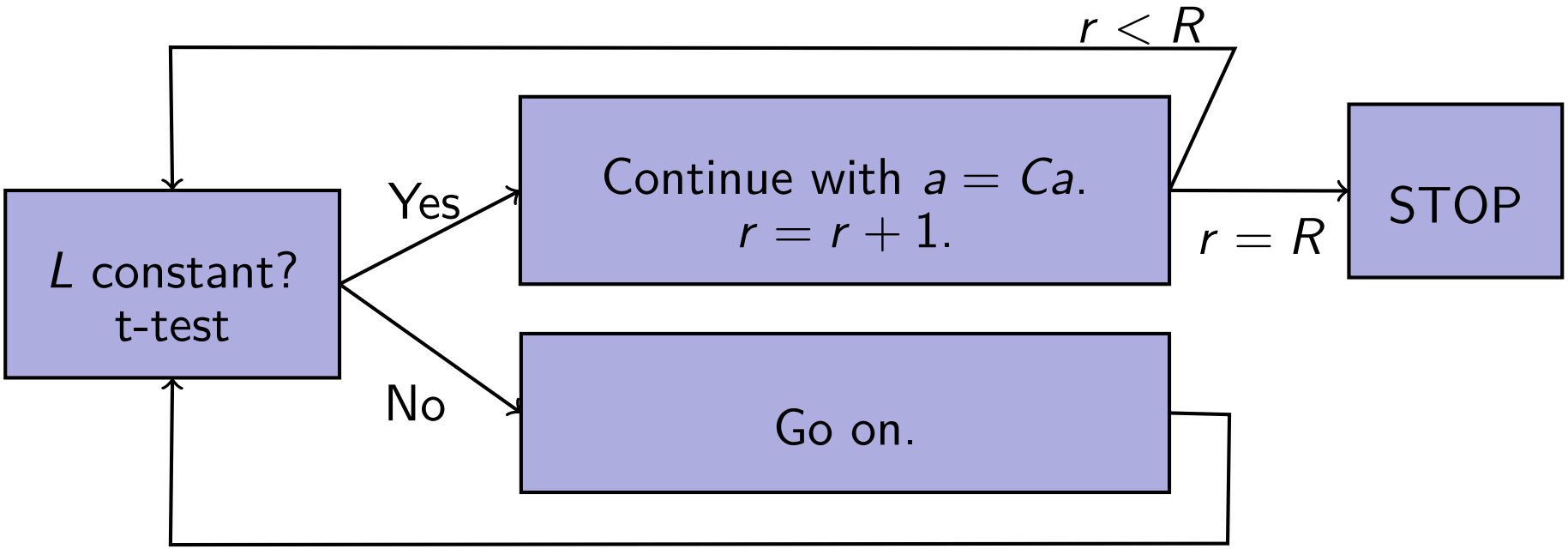
$$\arg\min_{\Theta} E\left(\|S(\Theta) - S_{obs}\|^2 \mid \Theta \right)$$

- Selection of the g best starting values $\Theta_{0,J_1}, \ldots, \Theta_{0,J_g}$ by simulating the likelihood

Convergence diagnostics

Compare $\hat{L}(\Theta_{K-l})$ and $\hat{L}(\Theta_K)$ to find out if the likelihood is constant.
Problem: Likelihood can be seemingly constant because of too small a.
$\Rightarrow$ Try different values of a.

Start with $r = 1$.



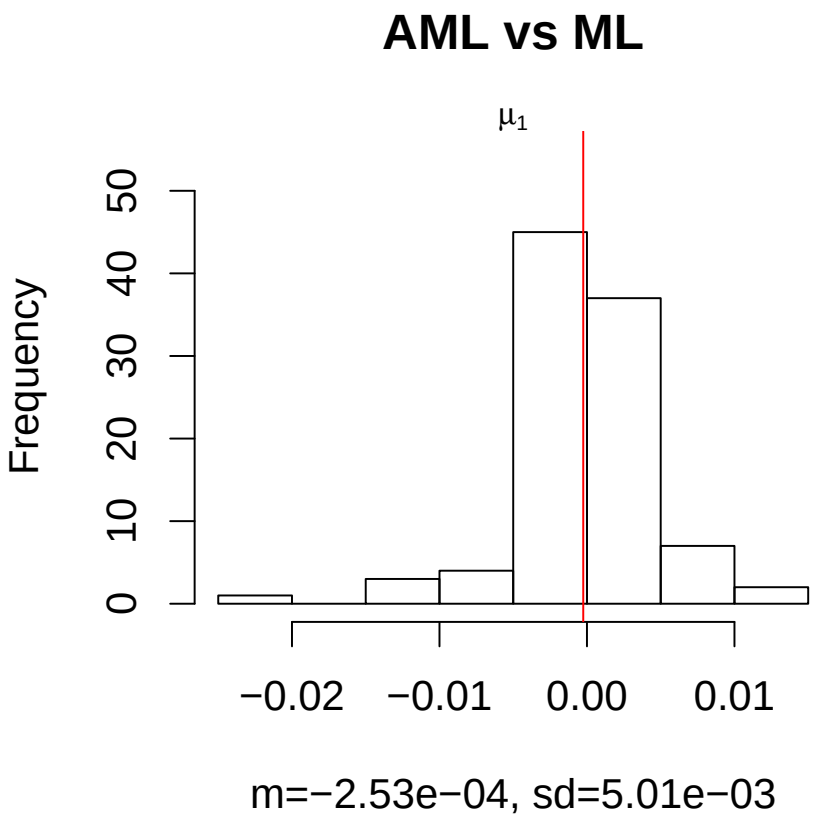
Model: 10-dimensional normal distribution

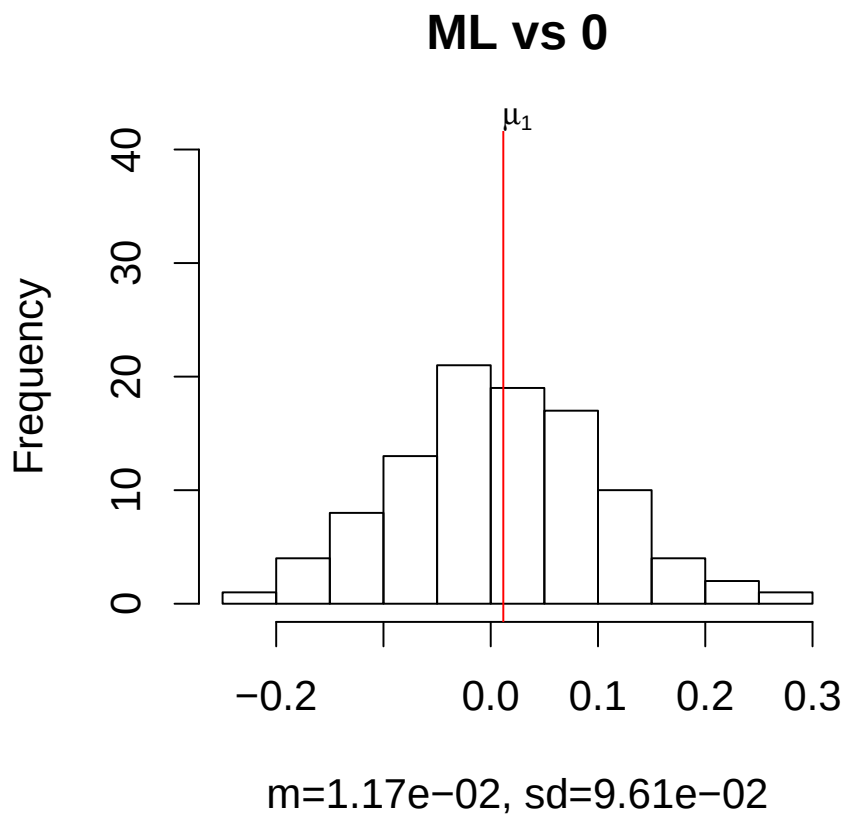
- Independent standard normal distributed random variables $X_1, \ldots, X_{10}$ with $n = 100$ realisations each.
- Estimated parameters: $\mu_1, \ldots, \mu_{10}$ (the variances $\sigma_1^2 = \ldots = \sigma_{10}^2 = 1$ are fixed).
- Summary statistics ($d = 10$):

$$\begin{array}{cccccc} \bar{X}_1 & \bar{X}_2 + \bar{X}_3 & \bar{X}_4 + \bar{X}_5 & \bar{X}_7 & \bar{X}_9 + \bar{X}_{10} \\ & \bar{X}_2 - \bar{X}_3 & \bar{X}_5 + \bar{X}_6 & \bar{X}_7 + \bar{X}_8 & \bar{X}_9 \cdot \bar{X}_{10} \\ & & \bar{X}_6 + \bar{X}_4 & & \end{array}$$

Simulation: 10-dimensional normal distribution

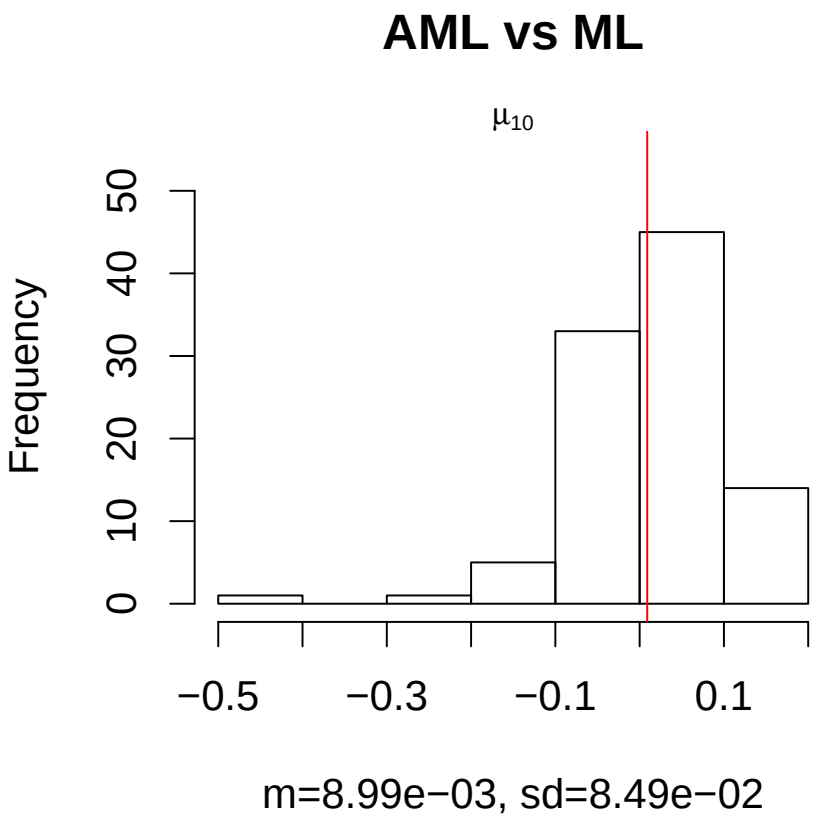
Results for μ_1: Comparison between AML and ML estimator

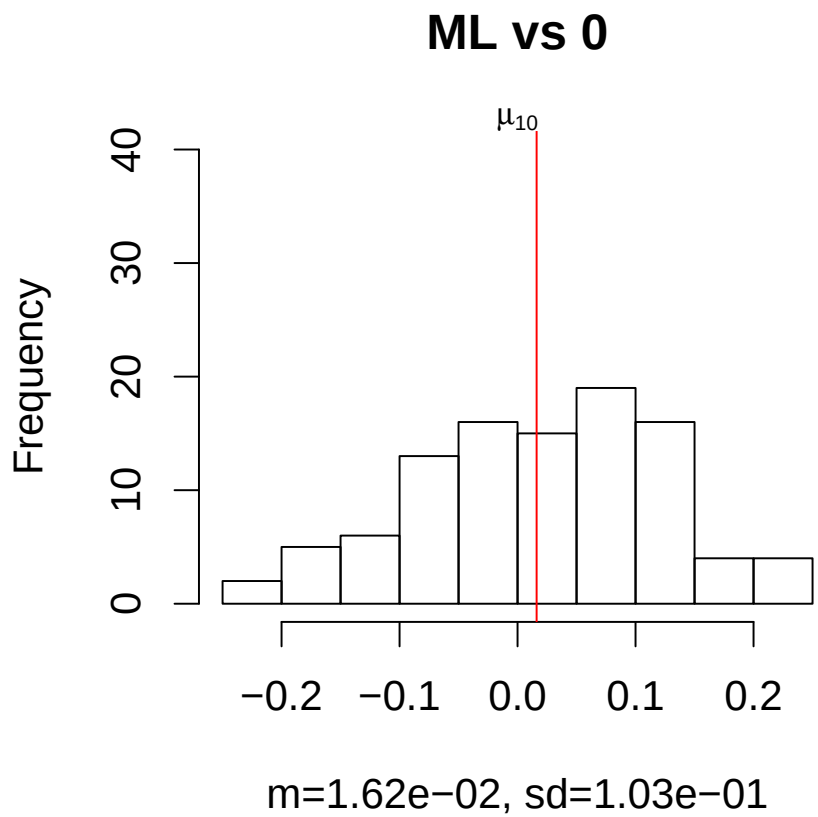




Simulation: 10-dimensional normal distribution

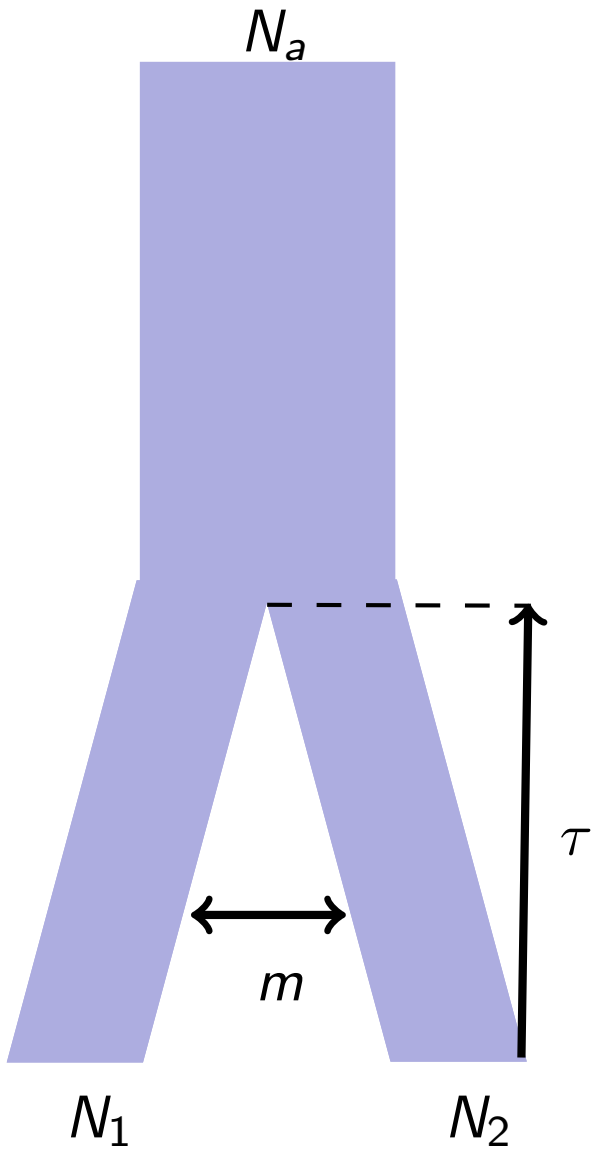
Results for μ_{10}: Comparison between AML and ML estimator



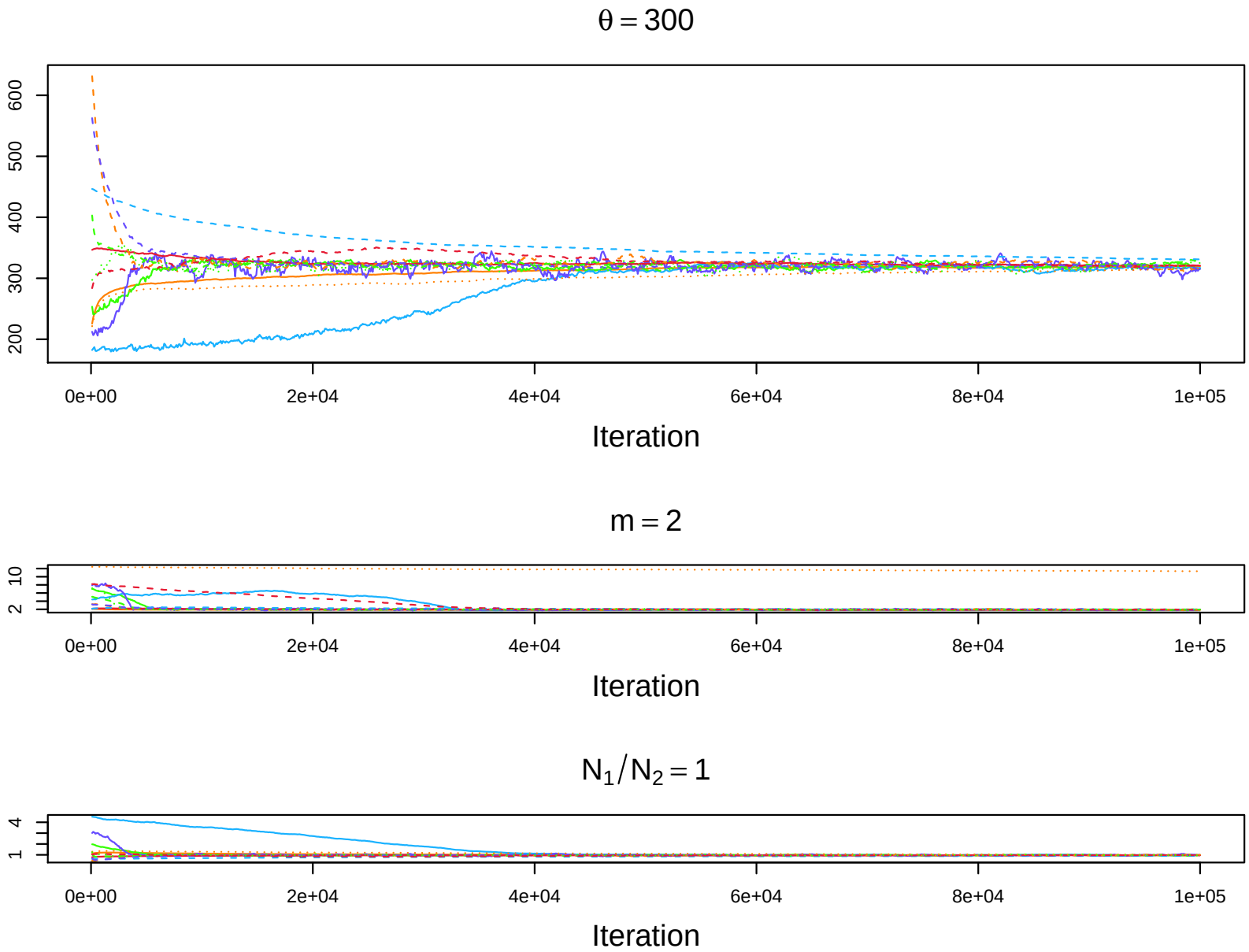


Model: Isolation-Migration coalescent model

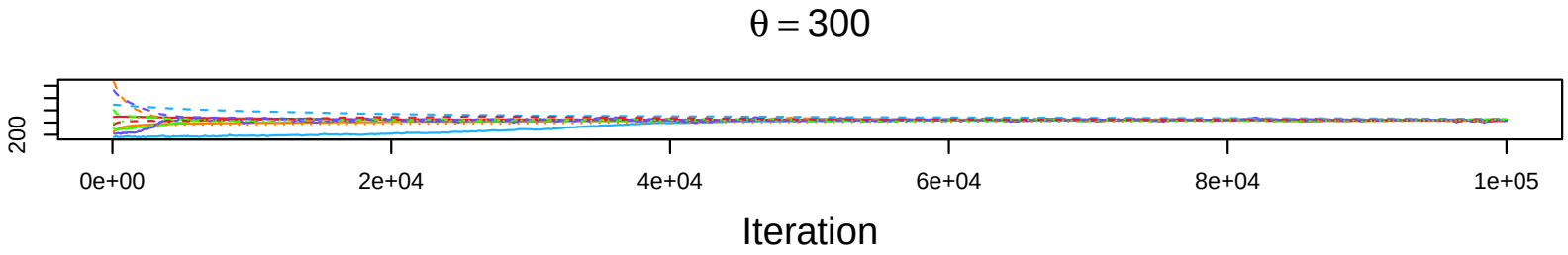
- Model parameters: $\theta = 300$, $m_1 = m_2 = 2$ and $N_1/N_2 = 1$
- Time τ since population split is not estimated
- Constant population size in both demes
- 25 samples per deme
- Summary statistics: $\hat{\theta}_W$, $\hat{\theta}_{W,1}$, $\hat{\theta}_{W,2}$ and F_{ST}
- Infinite sites model
- Coalescent simulations: `msms` (Ewing and Hermisson, 2010)

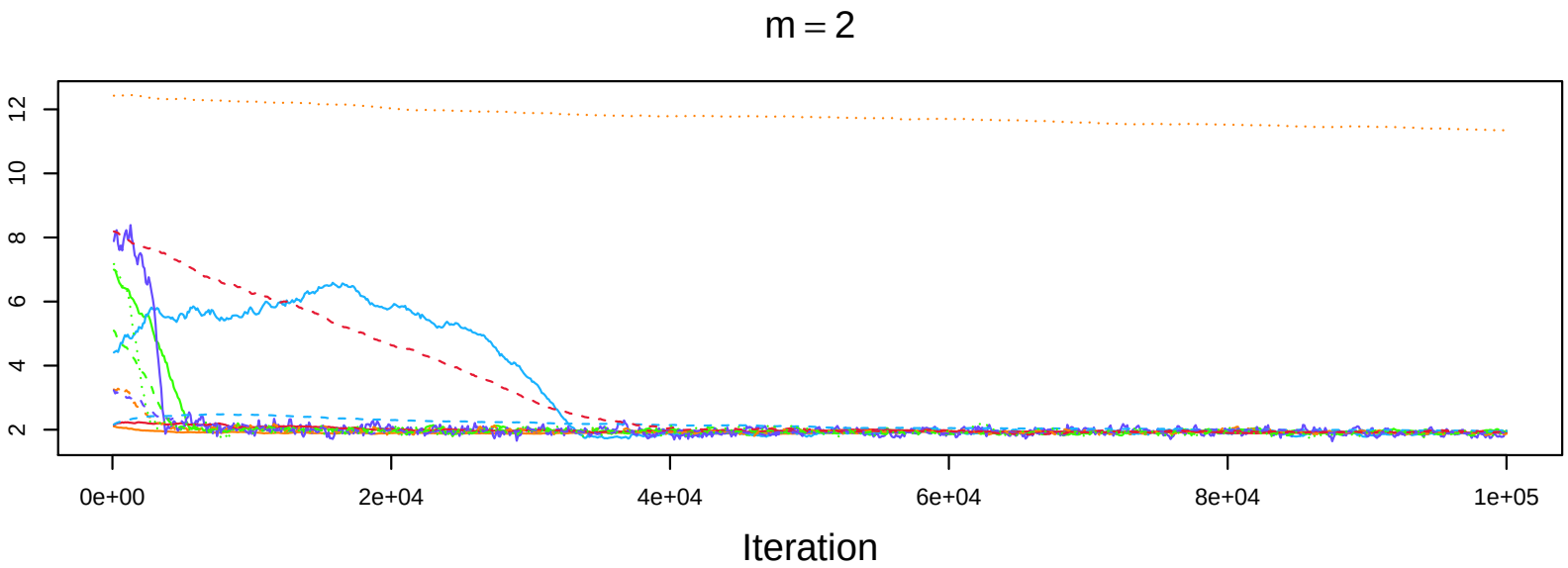


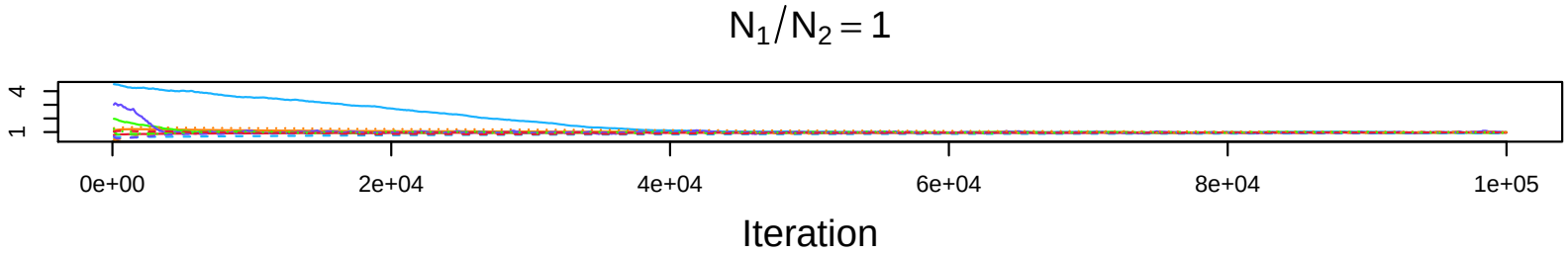
Example sequences



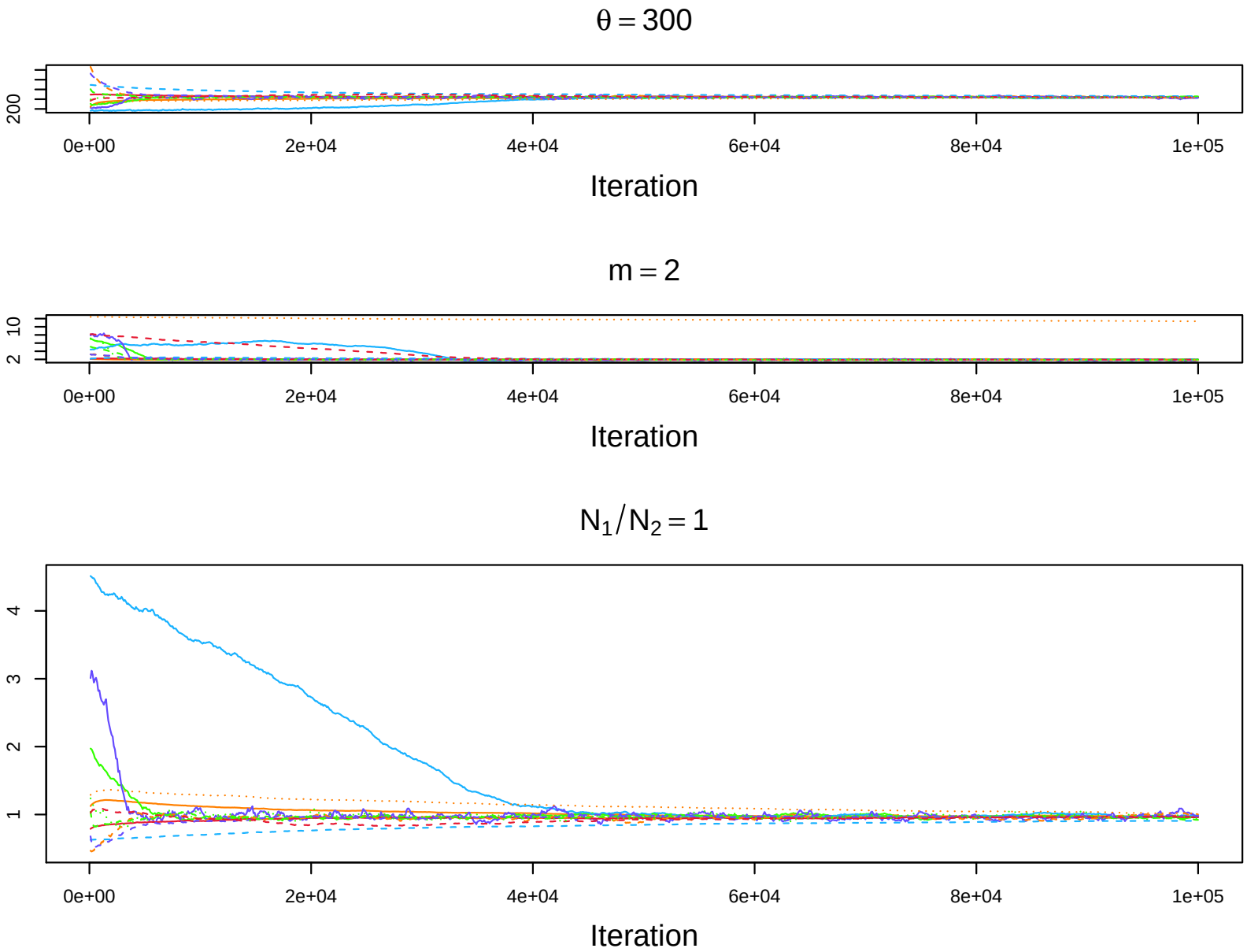
Example sequences







Example sequences



Conclusions and outlook

Conclusions:

- Quick convergence to ML estimator –
  BUT good tuning is important!
- Tuning guidelines and convergence diagnostics $\Rightarrow$ automatic algorithm

Outlook:

- Apply to higher dimensional coalescent models (simulations and real data)
- Confidence intervals (Bootstrap)
- (Joint) Site Frequency Spectrum as summary statistics