9.5 Estimation of pairwise coalescence times

Regev Schweiger and Richard Durbin (private communication) recently observed that the posterior mean of TMRCA in PSMC (Li and Durbin, 2011) is biased, and that the extent of bias depends on the true TMRCA as well as mutation and recombination rates. In what follows, we try to explain this phenomenon theoretically. For simplicity, we do not consider a full hidden Markov model, but the analysis carried out below captures the essence of the issue. Indeed, the theoretical predictions (Figure 9.7–Figure 9.9) derived below agree well with the empirical results that Schweiger and Durbin obtained using PSMC.

9.5.1 Preliminaries

Assume a constant-size, panmictic population and the infinite-sites model where mutations arrive according to a Poisson Point Process with rate $\theta/2$. Let T denote the TMRCA for a sample of size 2 and let M denote the number of mutations on the 2-leaved coalescent tree. Then, we have

$$\mathbb{P}[M = m \mid T = t] = \frac{1}{m!}(\theta t)^m e^{-\theta t}, \tag{9.5}$$

while the posterior density of T given $M = m \in \mathbb{N}_0$ is (see Proposition 3.5)

$$f(t \mid M = m) = \frac{1}{m!}(1+\theta)^{m+1} t^m e^{-(1+\theta)t}. \tag{9.6}$$

Hence, if θ is known and we are given the observation $M = m$, then we can construct the following estimators of T:

1. **Maximum likelihood**: By maximizing (9.5), we get

$$\widehat{T}_{\mathrm{ML}} = \frac{m}{\theta}.$$

2. **Maximum a posteriori**: By maximizing (9.6), we get

$$\widehat{T}_{\mathrm{MAP}} = \frac{m}{1+\theta}.$$

3. **Posterior mean**: By taking the expectation with respect to (9.6), we get

$$\widehat{T}_{\mathrm{PM}} = \frac{m+1}{1+\theta}.$$

Now, we note that the parameter θ is in fact a random variable and that it depends on the size (or width) W of the genomic region that supports the coalescent tree with TMRCA T. More precisely,

$$\theta = \theta_0 W,$$

where $\theta_0 = 4N_e\mu$ with μ being the per-base per-generation mutation rate. The distribution of W depends on the TMRCA T and the recombination rate. Assume that the recombination rate is constant over the region of interest and define $\rho = 4N_e r$, where r corresponds to the per-generation recombination rate between two consecutive bases. Then,

$$\mathbb{P}[W = w \mid T = t] = [1 - p(\rho, t)]^{w-1} p(\rho, t),$$

where $p(\rho, t)$ denotes the success probability of changing the TMRCA in one-base transition step. Under SMC′ (Marjoram and Wall, 2006), a generalized version of the sequentially Markov coalescent, $p(\rho, t)$ can be found as (Hobolth and Jensen, 2014)

$$p(\rho, t) = [e^{tQ}]_{1,1}, \quad \text{where } Q = \begin{pmatrix} -\rho & \rho & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{9.7}$$

9.5.2 Bias

Putting everything together, the conditional expectation of an estimator $\widehat{T}$ given $T = t$ can be found as

$$\begin{aligned} \mathbb{E}[\widehat{T}(M, W) \mid T = t] &= \sum_{w=1}^{\infty} \sum_{m=0}^{\infty} \widehat{T}(m, w) \mathbb{P}[M = m \mid T = t, W = w] \mathbb{P}[W = w \mid T = t] \\ &= \sum_{w=1}^{\infty} \sum_{m=0}^{\infty} \widehat{T}(m, w) \left[\frac{1}{m!} (\theta_0 w t)^m e^{-\theta_0 w t} \right] [1 - p(\rho, t)]^{w-1} p(\rho, t). \end{aligned} \tag{9.8}$$

For $\widehat{T} = \widehat{T}_{\mathrm{ML}}$, the above computation can be carried out in closed-form and one obtains

$$\mathbb{E}[\widehat{T}_{\mathrm{ML}}(M, W) \mid T = t] = t. \tag{9.9}$$

For $\widehat{T} = \widehat{T}_{\mathrm{MAP}}$, we obtain

$$\mathbb{E}[\widehat{T}_{\mathrm{MAP}}(M, W) \mid T = t] = \frac{\theta_0 t}{1 + \theta_0} \, {}_2F_1\Big(1, \frac{1}{\theta_0}; 2 + \frac{1}{\theta_0}; 1 - p\Big), \tag{9.10}$$

where ${}_2F_1$ denotes the ordinary hypergeometric function. Finally, for $\widehat{T} = \widehat{T}_{\mathrm{PM}}$, we get

$$\mathbb{E}[\widehat{T}_{\mathrm{PM}}(M, W) \mid T = t] = \frac{\theta_0 t}{1 + \theta_0} \, {}_2F_1\Big(1, \frac{1}{\theta_0}; 2 + \frac{1}{\theta_0}; 1 - p\Big) + \frac{p}{\theta_0} \Phi\Big(1 - p, 1, 1 + \frac{1}{\theta_0}\Big), \tag{9.11}$$

where Φ denotes the Hurwitz-Lerch transcendent.

We see from (9.9) that $\widehat{T}_{\mathrm{ML}}$ is an unbiased estimator of T, whereas (9.10) and (9.11) imply that $\widehat{T}_{\mathrm{MAP}}$ and $\widehat{T}_{\mathrm{PM}}$ are biased. Figures 9.7 and 9.8 respectively illustrate how $\mathbb{E}[\widehat{T}_{\mathrm{MAP}}(M, W) \mid T = t]$ and $\mathbb{E}[\widehat{T}_{\mathrm{PM}}(M, W) \mid T = t]$ deviate from t, for various values of r and μ. In general, $\widehat{T}_{\mathrm{PM}}$ appears less biased than $\widehat{T}_{\mathrm{MAP}}$.

9.5.3 Mean squared error

We saw in the previous section that the maximum likelihood estimator $\widehat{T}_{\mathrm{ML}}$ is unbiased, while $\widehat{T}_{\mathrm{MAP}}$ and $\widehat{T}_{\mathrm{PM}}$ are biased. Does this imply that $\widehat{T}_{\mathrm{ML}}$ is a better estimator than are

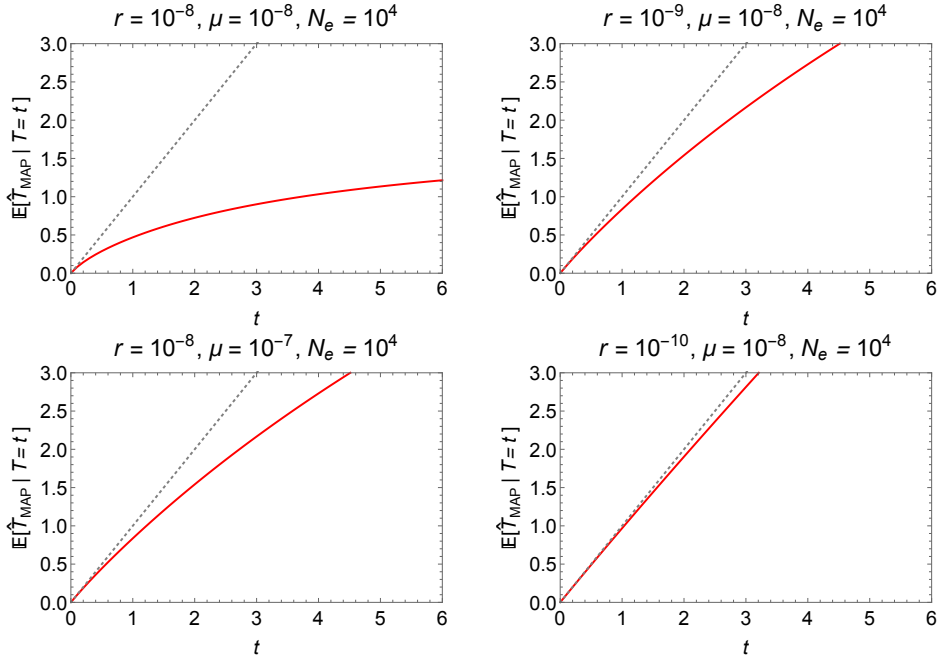

Fig. 9.7: $\mathbb{E}[\widehat{T}_{\mathrm{MAP}}(M,W) \mid T = t]$ as a function of t for various values of r and μ, with $N_e = 10^4$. Dotted lines correspond to $y = x$.

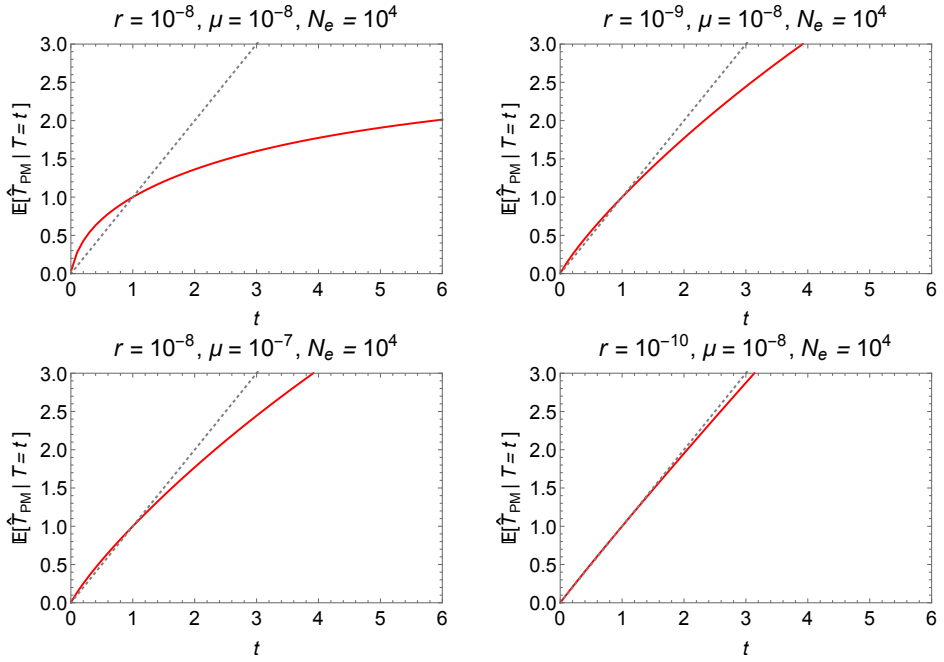

Fig. 9.8: $\mathbb{E}[\widehat{T}_{\mathrm{PM}}(M,W) \mid T = t]$ as a function of t for various values of r and μ, with $N_e = 10^4$. Dotted lines correspond to $y = x$.

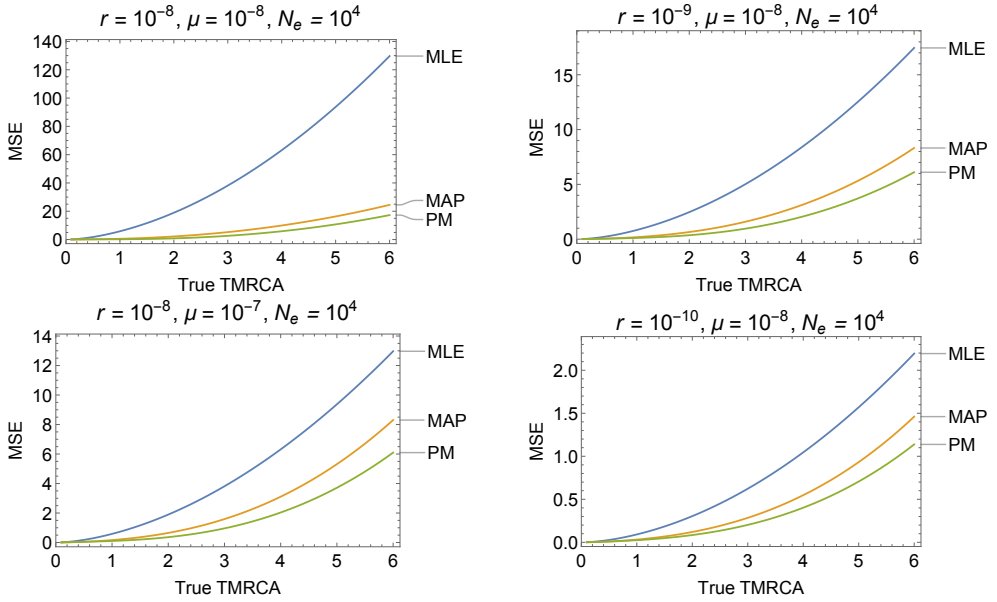

Fig. 9.9: Mean squared error (MSE) as a function of t for various values of r and μ, with $N_e = 10^4$.

$\widehat{T}_{\text{MAP}}$ and $\widehat{T}_{\text{PM}}$? Not necessarily! To see why, consider a quadratic loss function, in which case the risk $R(t, \widehat{T})$ corresponds to the mean squared error (MSE), which can be decomposed as

$$R(t, \widehat{T}) = \text{MSE}[\widehat{T}] = \mathbb{E}[(\widehat{T} - t)^2] = \text{Var}(\widehat{T}) + [\text{bias}(\widehat{T})]^2,$$

where $\text{bias}(\widehat{T}) = \mathbb{E}[\widehat{T}(M, W) \mid T = t] - t$. So, the key question is how $\text{Var}(\widehat{T}_{\text{ML}})$ compares with $\text{Var}(\widehat{T}_{\text{MAP}})$ and $\text{Var}(\widehat{T}_{\text{PM}})$. We can compute $\mathbb{E}[(\widehat{T} - t)^2 \mid T = t]$ in a similar vein as in (9.8), yielding

$$\begin{aligned}
R(t, \widehat{T}_{\text{ML}}) &= \frac{-tp \log(p)}{(1-p)\theta_0}, \\
R(t, \widehat{T}_{\text{MAP}}) &= \frac{pt}{\theta_0^2} \left[\theta_0 \Phi\Big(1 - p, 1, 1 + \frac{1}{\theta_0}\Big) + (t-1)\Phi\Big(1 - p, 2, 1 + \frac{1}{\theta_0}\Big) \right], \\
R(t, \widehat{T}_{\text{PM}}) &= \frac{p}{\theta_0^2} \left[\theta_0 t \Phi\Big(1 - p, 1, 1 + \frac{1}{\theta_0}\Big) + [1 + t(t-3)]\Phi\Big(1 - p, 2, 1 + \frac{1}{\theta_0}\Big) \right],
\end{aligned}$$

where p depends on the population-scaled recombination rate ρ and the true TMRCA t as shown in (9.7). As Figure 9.9 shows, it turns out that $\text{Var}(\widehat{T}_{\text{ML}})$ can be substantially larger than $\text{Var}(\widehat{T}_{\text{MAP}})$ and $\text{Var}(\widehat{T}_{\text{PM}})$ if $t \gg 0$, thereby leading to much larger expected loss $R(t, \widehat{T}_{\text{ML}})$ compared to $R(t, \widehat{T}_{\text{MAP}})$ and $R(t, \widehat{T}_{\text{PM}})$.

References

Griffiths RC (1981) Neutral two-locus multiple allele models with recombination. Theoretical Population Biology 19:169–186

Griffiths RC (1991) The two-locus ancestral graph. Selected Proceedings of the Sheffield Symposium on Applied Probability IMS Lecture Notes–Monograph Series 18:100–117

Griffiths RC, Marjoram P (1997) An ancestral recombination graph. In: Donnelly P, Tavaré S (eds) Progress in population genetics and human evolution, vol 87, Springer-Verlag, Berlin, pp 257–270

Hobolth A, Jensen JL (2014) Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theoretical population biology 98:48–58

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23:183–201

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496

Marjoram P, Wall JD (2006) Fast "coalescent" simulation. BMC Genet 7:16

Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. Nature Reviews Genetics 11(3):221–233

Sasaki M, Lange J, Keeney S (2010) Genome destabilization by homologous recombination in the germ line. Nature Reviews Molecular Cell Biology 11(3):182–195

Wiuf C (2000) A coalescence approach to gene conversion. Theoretical Population Biology 57:357–367