supplemental material

MLEs using Moore-Penrose inverse

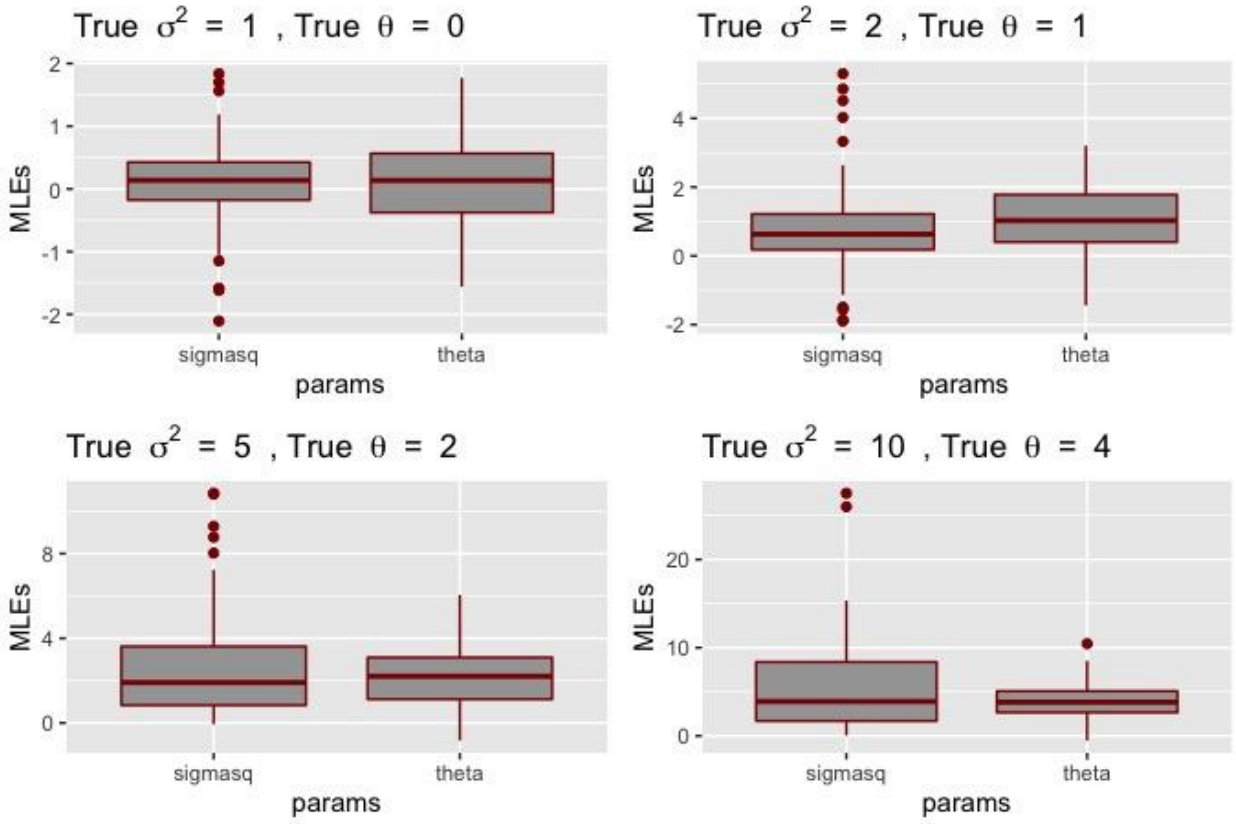

Figure 14. Unstable MLEs estimate of σ^2 using the Moore-Penrose inverse method (Chelosky's method fails) to compute the inverse of ill-condition $\boldsymbol{C}$ matrix. The pseudoinverse is computed using R package `MASS` function `ginv`. The tree size is 5 and traits are simulated given the true parameters and tree. Then MLEs are computed using the trait and tree. For each parameter set, 100 replicates are simulated and the values of MLEs are reported using boxplots. Four true parameters for $\sigma^2 = 1, 2, 5, 10$ are assessed, the boxplots in the four panels suggested that the MLEs for σ^2 produces large bias and the parameter estimates are not reliable.

Lemmas and their proofs

Lemma 1. *The shortest tip length of an ultrametric phylogenetic tree is the smallest eigenvalue of* $\boldsymbol{C}$ *. i.e.* $\min_\lambda\{\det(\boldsymbol{C}-\lambda\boldsymbol{I})=0\}=b$ *where* b *is the smallest tip length and* $\boldsymbol{I}$ *is an* n *by* n *identity matrix.*

Proof: Given an ultrametric bifurcated tree T of n tips, there exists a unique strictly ultrametric matrix (Nabben and Varga, 1994) $\boldsymbol{C}$ for representing the relatedness among the group of species. Let b be the smallest tip length, by the property of the structure of the ultrametric tree, $\boldsymbol{C}-b\boldsymbol{I}$ has at two identical

columns as well as two identical rows. This implies that $\det(\boldsymbol{C}-b\boldsymbol{I})=0$. Therefore, b is an eigenvalue of $\boldsymbol{C}$.

The next step is to show that b is the smallest eigenvalue in the eigenvalue set of $\boldsymbol{C}$. We claim that for all $\lambda_0 < b$, then λ_0 is not an eigenvalue. Consider the matrix $\boldsymbol{C}_0 = \boldsymbol{C} - \lambda_0 \boldsymbol{I}$, then $\boldsymbol{C}_0$ is still a strictly ultrametric matrix which is always invertible (see Nabben and Varga (1994), and Corollary 6.2.27 in Horn and Johnson (1986)). Then we have $\det(\boldsymbol{C}_0) \neq 0$ which implies $\det(\boldsymbol{C}-\lambda_0\boldsymbol{I}) \neq 0$. This consequence indicates that λ_0 is not an eigenvalue of $\boldsymbol{C}$. Therefore, b is the smallest eigenvalue of $\boldsymbol{C}$. +

Lemma 2. *Let $\boldsymbol{C}$ be the n by n strictly ultrametric matrix from the tree and κ be the condition number of $\boldsymbol{C}$. Let $\boldsymbol{C}_1$ be the matrix obtained by dropping the shortest tip from the tree and κ_1 be the condition number of $\boldsymbol{C}_1$. Then $\kappa \geq \kappa_1$.*

Proof: Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of $\boldsymbol{C}$. Since $\boldsymbol{C}_1$ is still a strictly ultrametic matrix of size $(n-1)$ by $(n-1)$, we can assume that $\boldsymbol{C}_1$ has eigenvalues $0 < \tau_1 \leq \tau_2 \leq \cdots \leq \tau_{n-1}$. By a special case of the Cauchy's interlacing theorem (see ch. 10.1 in Parlett (1987)), we have $0 < \lambda_1 \leq \tau_1 \leq \lambda_2 \leq \tau_2 \leq \cdots \leq \lambda_{n-1} \leq \tau_{n-1} \leq \lambda_n$. The condition number, defined as the ratio of the largest eigenvalue to the smallest eigenvalue are computed as $\kappa = \lambda_n / \lambda_1$ and $\kappa_1 = \tau_{n-1}/\tau_1$ for $\boldsymbol{C}$ and $\boldsymbol{C}_1$, respectively. From direct algebraic calculation, we have $\kappa - \kappa_1 = \frac{\lambda_n}{\lambda_1} - \frac{\tau_{n-1}}{\tau_1} = \frac{\tau_1\lambda_n - \tau_{n-1}\lambda_1}{\lambda_1\tau_1} \geq \frac{\tau_1\tau_{n-1} - \tau_{n-1}\lambda_1}{\lambda_1\tau_1} = \frac{\tau_{n-1}(\tau_1-\lambda_1)}{\lambda_1\tau_1} \geq 0$. This concludes that $\kappa \geq \kappa_1$. +

Lemma 3. *The set $\{\tau_i\}_{i=1}^{d}$ where τ_i is the length between the i^{th} and the $(i+1)^{\text{th}}$ speciation event has d elements if and only if the number of distinct elements in $\boldsymbol{C}$ of an ultrametric tree is $d+1$.*

Proof: Let $\tau_1, \tau_2, \cdots, \tau_d$ be the lengths between the i^{th} and the $(i+1)^{\text{th}}$ speciation events. Define $c_{i+1} = \tau_i + c_i, i = 1, 2, \cdots, d$ (i.e. $c_2 = \tau_1 + c_1 = \tau_1, c_3 = \tau_2 + c_2 = \tau_2 + \tau_1$, $c_4 = \tau_3 + c_3 = \tau_3 + \tau_2 + \tau_1, \cdots, c_{d+1} = \tau_d + c_d = \tau_d + \tau_{d-1} + \cdots + \tau_1$), then each c_i represents the node height of the i^{th} speciation events and c_i is stacked up by concatenating the branch segment between two successive speciation events.

On the other hand, without loss of generality, suppose $\{c_{d+1} > c_d > \cdots > c_1\}$ are $d+1$ distinct elements in $\boldsymbol{C}$. Since each element in $\boldsymbol{C}$ measures the affinity between a pair of species, each $c_i, i = 1, 2, \cdots, d+1$ is equal to the node height of the tree from the root. Define $\tau_i = c_{i+1} - c_i, i = 1, \cdots, d$, then each τ_i represents a value of taking the difference between two successive node heights, τ_i is equivalent to the branch segment between the i^{th} and the $(i+1)^{\text{th}}$ speciation so $\{\tau_i\}_{i=1}^{d}$ has exactly d elements. +

Lemma 4. *Under shrinkage method with $\delta = 1$, $\boldsymbol{S}_\delta = \boldsymbol{I}$ an identity matrix. Assume an ultrametric tree, the RMSD for θ under Brownian Motion model with rate parameter σ has an upper bound $\sqrt{\sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}/n^2}$ when trait $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)^t$ is instead analyzed under identical independent*

distributed model where n *is taxa size and* $c_{ij}, i, j = 1, 2, \cdots, n$ *is an element in the phylogenetic covariance matrix* $\boldsymbol{C}$ *.*

proof: Let θ and σ^2 be the true parameter of the Brownian motion model so that $E[y_i] = \theta$ and $E[y_i y_j] = \mathrm{Cov}[y_i, y_j] + E[y_i]E[y_j] = \sigma^2 c_{ij} + \theta^2$. The RMSD of θ can be algebraically computed as $\mathrm{RMSD}_\theta^2 = E[(\theta - \hat{\theta})^2] = E[\theta - \bar{y}]^2 = E\left[(\theta - \sum_{i=1}^n y_i / n)^2\right] = E\left[\theta^2 - 2\theta \sum_{i=1}^n y_i / n + (\sum_{i=1}^n y_i / n)^2\right] =$ $\theta^2 - 2\theta^2 + \sum_{i=1}^n \sum_{j=1}^n E[y_i y_j] / n^2 = -\theta^2 + \sum_{i=1}^n \sum_{j=1}^n (\sigma^2 c_{ij} + \theta^2) / n^2 = \sum_{i=1}^n \sum_{j=1}^n \sigma^2 c_{ij} / n^2$. In particular, when the root to tips tree height of ultrametic tree is 1 (i.e. $0 \le c_{ij} \le 1$ for all $i, j = 1, 2, \cdots, n$, so $\sum_{i=1}^n \sum_{j=1}^n c_{ij} \le n^2$), the $\mathrm{RMSD}_\theta = \sqrt{\sigma^2 \sum_{i=1}^n \sum_{j=1}^n c_{ij} / n^2}$ has a natural upper bound σ .

Lemma 5. *Under shrinkage method with* $\delta = 1$, $\boldsymbol{S}_\delta = \boldsymbol{I}$ *an identity matrix. Assume an ultrametric tree, the RMSD*^{2} *for* σ^2 *under Brownian Motion model has a lower bound*

$$E[(\sigma^2 - \hat{\sigma}^2)^2] \ge (\sigma^2)^2 \left(2 \sum_{i=1}^n \sum_{j=1}^n c_{ij} / n^2 + \left(\sum_{i=1}^n c_{ii} / n \right)^2 - 2 \right) + Var\left[\left(\sum_{i=1}^n \frac{(y_i - \bar{y})}{n} \right)^2 \right], \quad (10)$$

when trait $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)^t$ *is instead analyzed under identical independent distributed model where* n *is taxa size and* $c_{ij}, i, j = 1, 2, \cdots, n$ *is an element in the phylogenetic covariance matrix* $\boldsymbol{C}$ *.*

proof: The MLE for Brownian motion model given tree transformed under the shrinkage method with $\delta = 1$ is $\hat{\sigma}^2 = \sum_i (y_i - \bar{y})^2 / n$. The RMSD_σ^2 for σ^2 is expressed as following

$$E[(\sigma^2 - \hat{\sigma}^2)^2] = E[(\sigma^2 - \sum_{i=1}^n (y_i - \bar{y})^2 / n)^2] = (\sigma^2)^2 - 2\frac{\sigma^2}{n} E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] + \sigma^2 E\left[\left(\sum (y_i - \bar{y})^2 / n\right)^2\right] = (\sigma^2)^2 + ⓐ + ⓑ$$

.

For ⓐ, direct computation yield to $\frac{-2\sigma^2}{n} E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] = \frac{-2\sigma^2}{n} (\sum_{i=1}^n E[y_i^2] - nE[\bar{y}^2])$. Since $E[y_i y_j] = \sigma^2 (c_{ij} + \theta^2)$, $ⓐ = \frac{-2\sigma^2}{n} E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] = \frac{-2(\sigma^2)^2}{n} \sum_{i=1}^n (c_{ii} + \theta^2) + \frac{2(\sigma^2)^2}{n^2} \sum_{i=1}^n \sum_{j=1}^n (c_{ij} + \theta^2)$ $= \frac{-2(\sigma^2)^2}{n} \sum_{i=1}^n c_{ii} + \frac{2(\sigma^2)^2}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{ij} \ge -2(\sigma^2)^2 + 2(\sigma^2)^2 \sum_{i=1}^n \sum_{j=1}^n c_{ij} / n^2$.

For ⓑ, by $E[X^2] = \mathrm{var}[X] + (E[X])^2$,

$$ⓑ = E\left[\left(\sum_{i=1}^{n}(y_i-\bar{y})^2/n\right)^2\right] = \mathrm{var}\left[\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n}\right] + \frac{1}{n^2}\left(E\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\right)^2.$$ And

$$\frac{1}{n^2}\left(E\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\right)^2 = \frac{(\sigma^2)^2}{n^2}\left(\sum_{i=1}^{n}c_{ii} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij}\right)^2$$

$$\geq (\sigma^2)^2\left(\sum_{i=1}^{n}c_{ii}/n-1\right)^2 = (\sigma^2)^2 + (\sigma^2)^2\left(\sum_{i=1}^{n}c_{ii}/n\right)^2 - \frac{2\sum_{i=1}^{n}c_{ii}}{n} \geq (\sigma^2)^2\left(\sum_{i=1}^{n}c_{ii}/n\right)^2 - (\sigma^2)^2.$$

Combining above three terms $(\sigma^2)^2 + ⓐ + ⓑ$, a lower bound for $\mathrm{RMSD}^2_{\sigma^2}$ is shown in Eq. (10). In particular, when the root to tips tree height of ultrametic tree is 1 (i.e. $0 \leq c_{ij} \leq 1$, $\sum_{i=1}^{n}c_{ii}/n = 1$ and $\sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij} \leq 1$), $\mathrm{RMSD}^2_{\sigma^2} \geq (\sigma^2)^2 + \mathrm{Var}\left[\left(\sum_{i=1}^{n}\frac{(y_i-\bar{y})}{n}\right)^2\right]$. +

Scripts and relevant files

The following links connect to the relevant files and the R scripts for reproducing the tables and the figures from simulations in this work. All files can be accessed through the online folder https://tonyjhwueng.info/KappaPCM.

1. Figure 1 and Figure 4. https://tonyjhwueng.info/KappaPCM/Fig1.r.
2. Figure 2. https://tonyjhwueng.info/KappaPCM/Fig2.r.
3. Figure 3. https://tonyjhwueng.info/KappaPCM/Fig3.r.
4. Figure 5. https://tonyjhwueng.info/KappaPCM/UpperBoundKappa.R.
5. Figure 6. https://tonyjhwueng.info/KappaPCM/compshrunktree.r.
6. Figure 7. https://tonyjhwueng.info/KappaPCM/droptipkappa.R.
7. Figure 8. https://tonyjhwueng.info/KappaPCM/3taxaXYZ.pptx.
8. Figure 9. https://tonyjhwueng.info/KappaPCM/GraphicalAbstractV2.R.
9. Figure 10. https://tonyjhwueng.info/KappaPCM/Fig10.R.
10. Figure 11 and Figure 12. https://tonyjhwueng.info/KappaPCM/AssessmentMethodsSummaryCombineReplicate.R.
11. Figure 13. https://tonyjhwueng.info/KappaPCM/AssemRepSummaryShrinkPlot.R.
12. Figure 14. https://tonyjhwueng.info/KappaPCM/inaccest.R.
13. Table 1. https://tonyjhwueng.info/KappaPCM/MuMinCoef.R.
14. solvefail.pdf: https://tonyjhwueng.info/KappaPCM/solvefail.pdf.
15. MPfail.pdf: https://tonyjhwueng.info/KappaPCM/MPfail.pdf.
16. pmmfelprunzerobranch.pdf: https://tonyjhwueng.info/KappaPCM/pmmfelprunzerobranch.pdf.