Multi-task Learning Improves Ancestral State Reconstruction

Lam Si Tung Ho*

Department of Mathematics and Statistics
Dalhousie University, Halifax, Nova Scotia, Canada

Vu Dinh

Department of Mathematical Sciences, University of Delaware, USA

Cuong V. Nguyen

Department of Engineering, University of Cambridge, UK

Abstract

We consider the ancestral state reconstruction problem where we need to infer phenotypes of ancestors using observations from present-day species. For this problem, we propose a multi-task learning method that uses regularized maximum likelihood to estimate the ancestral states of various traits simultaneously. We then show both theoretically and by simulation that this method improves the estimates of the ancestral states compared to the maximum likelihood method. The result also indicates that for the problem of ancestral state reconstruction under the Brownian motion model, the maximum likelihood method can be improved.

Keywords: ancestral state reconstruction, multi-task learning, maximum likelihood estimator

1. Introduction

Inferring phenotypes (values or states of a trait) of ancestral species using observations from present-day species is an important problem that lies at the

*Corresponding author
Email address: lam.ho@dal.ca (Lam Si Tung Ho)

heart of evolutionary biology. This problem, usually called ancestral state reconstruction, has many modern applications including inferring the origin of the HIV-1 pandemic in Central Africa in the 1920s (Faria et al., 2014; Gill et al., 2017), understanding the global circulation patterns of influenza A/H1N1 and B viruses (Bedford et al., 2015), and testing between two popular competing hypotheses (Anatolia and steppe) for the origin of the Indo-European languages (Bouckaert et al., 2012).

One of the most popular models for ancestral state reconstruction is to assume a trait (or character) evolves along the branches of a phylogenetic tree according to a stochastic process. The observations at the leaves of this tree are the trait values of the present species while the ancestral state is the trait value at the root. In this model, a well-known approach for reconstructing the ancestral state is the maximum likelihood method, where we maximize the likelihood of the observed trait values with respect to parameters that depict the ancestral state.

In this paper, we are interested in reconstructing the ancestral states for multiple continuous traits concurrently. For continuous traits, the stochastic process that characterizes the traits' evolution is usually assumed to follow a Brownian motion model (Felsenstein, 2004). If the maximum likelihood method is applied to each trait of the problem separately, we can construct the ancestral states of the traits independently. However, in this work, we theoretically show that simultaneously reconstructing the ancestral states of several continuous traits can be improved by multi-task learning using the regularized maximum likelihood method.

Multi-task learning is an important machine learning framework that aims to improve the learning performance by combining data from many tasks. It has been applied successfully in many areas including natural language processing (Dong et al., 2015; Lu et al., 2016), computer vision (Li et al., 2010; Zhang et al., 2012), and feature selection (Argyriou et al., 2006; Zhang et al., 2006). Among the methods for multi-task learning, regularization techniques are perhaps the simplest and most popular (Evgeniou & Pontil, 2004; Feldman et al., 2014; Lu

et al., 2016). The idea behind this technique is to use a penalty term to pull the learned models closer to each other. In this work, we show this regularization method improves the ancestral state reconstruction for evolutionary data.

In summary, our work makes the following novel contributions to the ancestral state reconstruction problem. First, we propose a regularized maximum likelihood method to simultaneously reconstruct the ancestral states of several continuous traits from observations. In essence, this method pulls the information from different traits together using an ℓ_2-penalty term. We then prove theoretically that the proposed method helps to improve the accuracy of the ancestral states' estimates for both traits that belong to the same set of species or to different sets of species. Our simulation on real phylogenetic trees also confirms the theoretical findings in the paper. The results indicates that for the problem of ancestral state reconstruction under the Brownian motion model, the maximum likelihood method can be improved.

2. Ancestral State Reconstruction under the Brownian Motion Model

In evolutionary biology, living species are related to each other and share descendants from a common ancestor. This relatedness is depicted by a phylogenetic tree whose leaves represent the species at the present time and whose root represents the common ancestor of these species. In this tree, each internal node corresponds to a speciation event at which a population splits into two distinct populations and edge lengths of the tree measure the evolutionary time between speciation events. In practice, researchers reconstruct phylogenetic trees from DNA sequences and calibrate these trees (i.e., translating edge lengths into absolute time) using fossils and geological events. Figure 1 visualizes a calibrated 4507-species mammal tree from Bininda-Emonds et al. (2007) that is constructed from molecular data.

Ancestral state reconstruction is the problem of estimating the trait value of the common ancestor from the trait values of present-day species. This is a useful task for understanding the evolutionary history of living organisms.

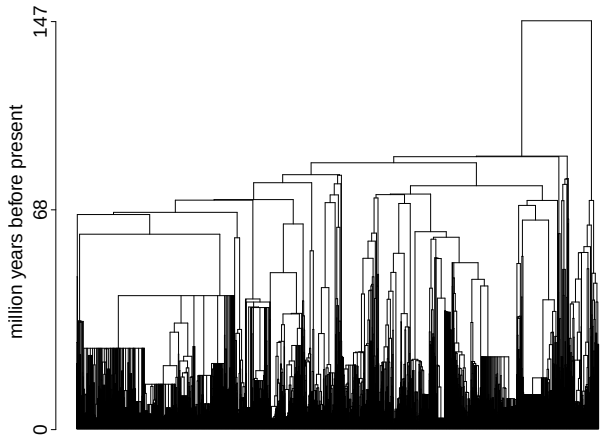


Figure 1: The calibrated 4507-species mammal tree from Bininda-Emonds et al. (2007).

For example, Gill et al. (2017) considered geographical traits (longitude and latitude) to study the spread of HIV-1 in Central Africa and by reconstructing the ancestral states of these traits, they are able to infer the origin of this pandemic.

One mathematical approach for the ancestral state reconstruction problem is to model the evolution of a trait along a phylogenetic tree by a stochastic process. In this paper, we focus on ancestral state reconstruction of continuous traits and for this setting, the Brownian motion (BM) model is one of the most commonly used approaches. This model assumes a trait evolves along each branch of the phylogenetic tree according to a BM. At each speciation event (i.e., at each node of the phylogenetic tree), the BM splits into several processes which evolve independently along descendant edges (see Ané (2008) for more details).

Under this BM model, the trait value at the root of the phylogenetic tree (that is, the ancestral state) is the starting value μ of the BM. The observed trait values $\mathbf{Y} \in \mathbb{R}^n$ at the leaves, where n is the number of leaves or species, follow the Normal distribution $\mathcal{N}(\mu\mathbf{1}, \sigma^2\mathbf{V})$ where $\mathbf{1}$ is an all-ones vector of length n, σ^2 is the variance of the BM, and $\mathbf{V} = [v_{ij}]_{1\leq i,j\leq n}$ is the phylogenetic

correlation matrix between species. Here, v_{ij} is the distance (i.e., total edge lengths) from the root to the most recent common ancestor of species i and j. Applications of the BM model include modeling flower size of Euphorbiaceae species (Davis et al., 2007), body mass of mammals (Cooper & Purvis, 2010), and chromosome number of primates (Baum et al., 2016).

A popular method for estimating the ancestral state μ of this trait is the maximum likelihood (ML) approach, which estimates μ and σ using the following formulae:

$$\begin{aligned}(\hat{\mu}^{\mathrm{ML}}, \hat{\sigma}^{\mathrm{ML}}) &= \operatorname*{argmax}_{\mu,\sigma} \log \mathbb{P}(\mathbf{Y} \mid \mu, \sigma^2) \\ &= \operatorname*{argmax}_{\mu,\sigma} \left\{ -\frac{(\mu\mathbf{1} - \mathbf{Y})^\top \mathbf{V}^{-1}(\mu\mathbf{1} - \mathbf{Y})}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) \right\}. \end{aligned} \tag{2.1}$$

For all σ, the above optimization problem can be solved in closed form. That is,

$$\hat{\mu}^{\mathrm{ML}} = (\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{Y}), \tag{2.2}$$

$$\hat{\sigma}^{\mathrm{ML}} = \sqrt{\frac{1}{n}(\hat{\mu}^{\mathrm{ML}}\mathbf{1} - \mathbf{Y})^\top \mathbf{V}^{-1}(\hat{\mu}^{\mathrm{ML}}\mathbf{1} - \mathbf{Y})}. \tag{2.3}$$

Note that $\hat{\mu}^{\mathrm{ML}}$ does not depend on σ. To measure the quality of an estimator, we often use the mean squared error (MSE). The MSE of an estimator $\hat{\mu}$ is defined as $\mathrm{MSE}(\hat{\mu}) = \mathbb{E}(\hat{\mu} - \mu)^2$. For the ML estimator above, its MSE is:

$$\mathrm{MSE}(\hat{\mu}^{\mathrm{ML}}) = \frac{\sigma^2}{\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}} \tag{2.4}$$

where σ^2 is the true variance of the BM model.

3. Multi-task Ancestral State Reconstruction

In this paper, we consider the problem of reconstructing the ancestral states of m continuous traits simultaneously under the BM model from m vectors of trait values $\{\mathbf{Y}_i\}_{i=1}^m$. We refer to this problem as the *multi-task ancestral state reconstruction problem*. A naive approach to this problem would apply the ML method above for each trait independently or attempt to estimate the ancestral

states of multiple traits jointly under the multivariate BM model using ML method. However, we note that the joint ML estimators are the same as the ML estimators when we estimate the ancestral state of each trait separately. Indeed, let $\mathbf{X}$ be the $n \times m$ matrix of trait values for n species and m traits (that is, the i-th column is the trait values $\mathbf{Y}_i$ of the trait i-th), then Revell & Harmon (2008) pointed out that the ML estimators of $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$ under the multivariate BM is $\hat{\mu}^{\mathrm{ML}} = (\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{X})$. Hence, $\hat{\mu}_i^{\mathrm{ML}} = (\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{Y}_i)$, which is exactly the ML estimator for μ_i when we estimate it separately.

In this work, we propose a method to estimate all the m ancestral states simultaneously using a regularized maximum likelihood objective. We will also prove that our method can improve the estimators of the ancestral states compared to the naive ML method. More specifically, we propose the following multi-task estimator for the problem that estimates the ancestral states by:

$$(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m) = \underset{\mu_1,\mu_2,\ldots,\mu_m}{\operatorname{argmax}} \sum_{i=1}^{m} \log \mathbb{P}(\mathbf{Y}_i \mid \mu_i, 1) - \frac{\lambda}{2}\left[\sum_{1\le k,l\le m} (\mu_k - \mu_l)^2\right], \tag{3.1}$$

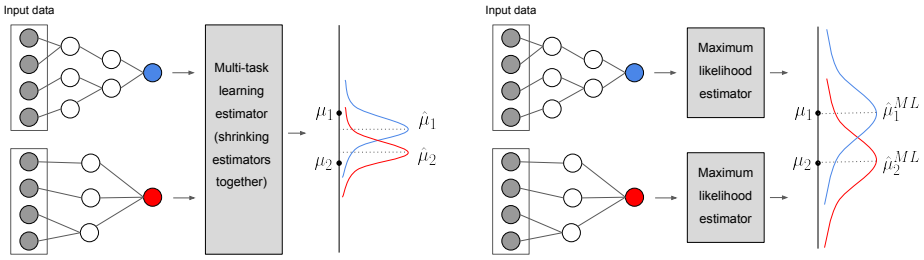
where μ_i is the parameter representing the ancestral state of trait i and λ is a non-negative parameter that balances the importance of the regularizer term $\sum_{1\le k,l\le m} (\mu_k - \mu_l)^2$. We call λ the regularizer parameter. Note that for simplicity, we have assumed the BMs for all traits have unit variance, i.e., $\sigma_i^2 = 1$ for all $i = 1, 2, \ldots, m$. In practice, if $\{\sigma_i\}_{i=1}^m$ are known, this assumption can be satisfied by standardizing the data using:

$$\mathbf{Y}'_i = \frac{\mathbf{Y}_i}{\sigma_i}, \quad \forall i = 1, 2, \ldots, m. \tag{3.2}$$

If $\{\sigma_i\}_{i=1}^m$ are unknown, we can standardize the data using any consistent estimator of $\{\sigma_i\}_{i=1}^m$, e.g., the ML estimators $\{\hat{\sigma}_i^{\mathrm{ML}}\}_{i=1}^m$.

Objective functions similar to (3.1) were also used by Feldman et al. (2014) for independent Gaussian data and by Lu et al. (2016) for natural language data. The idea of using a regularized maximum likelihood objective to estimate the parameters of different models jointly is commonly used in machine learning

for the multi-task learning problem (Evgeniou & Pontil, 2004). However, these multi-task learning algorithms are usually applied to highly complex models that render their theoretical analysis difficult. Our work, on the other hand, is able to provide theoretical guarantees for the estimators under the BM model. The main idea of the additional regularizer term is to shrink the estimators together. As a result, the estimators are slightly biased but can have smaller variances and MSE compared to the ML estimators (Figure 2).



(a) Multi-task learning method (b) Maximum likelihood method

Figure 2: Schematic figure illustrating the distinction between the proposed multi-task learning method (a) and the maximum likelihood (ML) methods (b). We want to estimate the ancestral states μ_1 and μ_2 of two traits at the roots (blue and red nodes) of two or more phylogenetic trees (which could be the same or different). The ML estimator of each trait is unbiased and follows Gaussian distributions. Our multi-task learning method shrinks the estimators together, which make them slightly biased but can reduce the mean squared error (bias-variance tradeoff).

In the following, we shall prove that the estimators obtained from (3.1) are better than normal ML estimators in terms of the MSE under two scenarios: (1) when the traits of interest are from species of the same phylogenetic tree and (2) when the traits are from species of two different phylogenetic trees. The insights from (2) can also be extended to more than two phylogenetic trees, although we omit it here for simplicity.

3.1. Traits From One Phylogenetic Tree

In this scenario, we consider multiple traits of a set of species coming from one phylogenetic tree. Since all traits evolve on the same phylogenetic tree, they

have the same phylogenetic correlation matrix $\mathbf{V}$. Under this setting, we can obtain an analytical solution for (3.1) as follows.

First, since the likelihood functions are Gaussian, we can rewrite (3.1) as:

$$(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m) = \underset{\mu_1,\mu_2,\ldots,\mu_m}{\operatorname{argmax}} \left\{ -\sum_{i=1}^{m} (\mathbf{Y}_i - \mu_i \mathbf{1})^\top \mathbf{V}^{-1} (\mathbf{Y}_i - \mu_i \mathbf{1}) - \lambda \Big[\sum_{1 \le k,l \le m} (\mu_k - \mu_l)^2 \Big] \right\}. \tag{3.3}$$

Take the partial derivatives of the objective function above w.r.t. each μ_i and set them to 0. We then obtain $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m)$ as a solution of the following system of equations:

$$(\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1})\mu_i - \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{Y}_i + 2\lambda m \mu_i - 2\lambda \sum_{k=1}^{m} \mu_k = 0, \quad \text{for } i = 1, 2, \ldots, m. \tag{3.4}$$

Taking the summation of all equations in (3.4), we have:

$$(\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1}) \sum_{i=1}^{m} \mu_i - \sum_{i=1}^{m} \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{Y}_i = 0. \tag{3.5}$$

Thus,

$$\sum_{i=1}^{m} \mu_i = \sum_{i=1}^{m} \frac{\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{Y}_i}{\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1}}. \tag{3.6}$$

From (3.4) and (3.6), we obtain the solution for this scenario:

$$\hat{\mu}_i = \frac{\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{Y}_i}{\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1} + 2\lambda m} + \frac{2\lambda (\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1})^{-1} \sum_{k=1}^{m} \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{Y}_k}{\mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1} + 2\lambda m}, \quad \text{for } i = 1, 2, \ldots, m. \tag{3.7}$$

We note that the traits may not be independent since they come from species at the leaves of the same phylogenetic tree. Let c_{kl} be the correlation between trait k and trait l. The covariance between the observations $\mathbf{Y}_k$ and $\mathbf{Y}_l$ is $c_{kl}\mathbf{V}$. We normally assume that different traits are not perfectly positive correlated; that is, $c_{kl} < 1$ if $k \neq l$. It is worth noticing that our results hold even when traits are negative correlated ($c_{kl} < 0$). Moreover, the improvement of our method compared to the ML estimators actually increases in such scenarios (see equation A.1).

Denote

$$\lambda_s = \frac{(m-1)(1-\max_{k\neq l} c_{kl})}{(m-1)^2(\max_i\{\mu_i\}-\min_i\{\mu_i\})^2 + \left[m^2 - \left(\sum_{k,l} c_{kl}\right)\right](\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1})^{-1}}. \tag{3.8}$$

We have the following theorem which proves the estimates in (3.7) using our method are better than normal ML estimates in terms of the MSE for appropriate values of the regularizer parameter λ (see section Appendix A for proof of this theorem).

Theorem 3.1. *Simultaneously reconstructing ancestral states of m traits from species of the same phylogenetic tree using* (3.7) *is better than reconstructing them separately using ML estimators, that is* $\mathrm{MSE}(\hat{\mu}_i^{ML}) > \mathrm{MSE}(\hat{\mu}_i)$ *for all* $i = 1, 2, \ldots, m$, *when*

$$\lambda \in \begin{cases} (0, +\infty) & \text{if } (\max_i\{\mu_i\} - \min_i\{\mu_i\})^2 \leq \frac{m^2 - \left(\sum_{k,l} c_{kl}\right)}{(m-1)^2 \mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}} \\ (0, \lambda_s) & \text{otherwise} \end{cases}. \tag{3.9}$$

We remark that the condition

$$(\max_i\{\mu_i\} - \min_i\{\mu_i\})^2 \leq \frac{m^2 - \left(\sum_{k,l} c_{kl}\right)}{(m-1)^2 \mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}} \tag{3.10}$$

means the ancestral states $\{\mu_i\}_{i=1}^m$ of these m traits are concentrated. Although under this condition, the multi-task estimator improves the ancestral state reconstruction with any $\lambda > 0$, we often do not know if this condition is satisfied in practice. So, in this case, we suggest to use $\lambda = \hat{\lambda}_s/2$, where $\hat{\lambda}_s$ is the following estimator of λ_s:

$$\hat{\lambda}_s = \frac{(m-1)(1-\max_{k\neq l} \hat{c}_{kl})}{(m-1)^2(\max_i\{\hat{\mu}_i^{\mathrm{ML}}\}-\min_i\{\hat{\mu}_i^{\mathrm{ML}}\})^2 + \left[m^2 - \left(\sum_{k,l} \hat{c}_{kl}\right)\right](\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1})^{-1}}$$

with the empirical estimation $\hat{c}_{kl}$ of the correlation c_{kl} evaluated by:

$$\hat{c}_{kl} = \frac{(\mathbf{Y}_k - \mathbf{1}^\top \mathbf{Y}_k/n)^\top (\mathbf{Y}_l - \mathbf{1}^\top \mathbf{Y}_l/n)}{\sqrt{(\mathbf{Y}_k - \mathbf{1}^\top \mathbf{Y}_k/n)^\top (\mathbf{Y}_k - \mathbf{1}^\top \mathbf{Y}_k/n)}\sqrt{(\mathbf{Y}_l - \mathbf{1}^\top \mathbf{Y}_l/n)^\top (\mathbf{Y}_l - \mathbf{1}^\top \mathbf{Y}_l/n)}},$$

for $k, l \in \{1, 2, \ldots, m\}$ and n is the number of species (the length of $\mathbf{Y}_k$).

In our simulations in section 4, we will show that if λ is large and the condition (3.10) does not hold, the multi-task estimator can be worse than the ML method.

In the above formulas, $\hat{c}_{kl}$ is the sample correlation coefficient estimated from the trait values. This is a well-known estimate that has been implemented in many statistical softwares such as `R`. Using these coefficients and the ML estimators, we can compute an estimate $\hat{\lambda}_s$ of λ_s and set $\lambda = \hat{\lambda}_s/2$ so that it is small enough. We emphasize here that our method and the ML method both require $O(nm)$ time to compute using the tree traversal algorithm proposed by Ho & Ané (2014). So, computing the ML solution to estimate λ does not increase the complexity of our method asymptotically.

3.2. Traits From Two Different Phylogenetic Trees

The second scenario we consider is when we have traits from two different sets of species that come from two different phylogenetic trees. For simplicity, we consider only two traits in this section. However, we note that this consideration is still useful, especially when we want to use an old data set to improve the reconstruction of ancestral states from a new data set. The idea in this section can also be used for more than two traits.

Since we have two different phylogenetic trees, there are two different phylogenetic correlation matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. In this case, (3.1) becomes:

$$(\hat{\mu}_1, \hat{\mu}_2) = \underset{\mu_1,\mu_2}{\operatorname{argmin}} \Big\{ (\mathbf{Y}_1 - \mu_1\mathbf{1})^\top \mathbf{V}_1^{-1} (\mathbf{Y}_1 - \mu_1\mathbf{1}) + (\mathbf{Y}_2 - \mu_2\mathbf{1})^\top \mathbf{V}_2^{-1} (\mathbf{Y}_2 - \mu_2\mathbf{1}) + 2\lambda(\mu_1 - \mu_2)^2 \Big\}. \quad (3.11)$$

Setting the partial derivatives of this objective w.r.t. μ_1 and μ_2 to zero, we obtain $(\hat{\mu}_1, \hat{\mu}_2)$ as a solution of the following system of equations:

$$\begin{cases} \mu_1 - \frac{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{Y}_1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}}(\mu_1 - \mu_2) = 0 \\ \mu_2 - \frac{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{Y}_2}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}(\mu_2 - \mu_1) = 0 \end{cases}. \quad (3.12)$$

By subtracting these two equations, we have:

$$\mu_1 - \mu_2 = \left(\frac{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{Y}_1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} - \frac{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{Y}_2}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right) \Bigg/ \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right).$$

From this equation and (3.12), we obtain the solution for this scenario:

$$\begin{aligned}
\hat{\mu}_1 &= \frac{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{Y}_1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} - \\
&\quad \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} \left(\frac{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{Y}_1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} - \frac{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{Y}_2}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right) \Bigg/ \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right), \\
\hat{\mu}_2 &= \frac{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{Y}_2}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}} - \\
&\quad \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}} \left(\frac{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{Y}_2}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}} - \frac{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{Y}_1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}}\right) \Bigg/ \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right).
\end{aligned} \tag{3.13}$$

Denote $\lambda_d = \dfrac{1}{(\mu_1 - \mu_2)^2 + (\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^{-1} + (\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1})^{-1}}$.

We have the following theorem which proves the estimates in (3.13) are better than normal ML estimates in terms of the MSE for appropriate values of the regularizer parameter λ (see section Appendix B for proof of this theorem).

Theorem 3.2. *Simultaneously reconstructing ancestral states of two traits from species of two different phylogenetic trees using* (3.13) *is better than reconstructing them separately using ML estimators, that is* $\mathrm{MSE}(\hat{\mu}_i^{ML}) > \mathrm{MSE}(\hat{\mu}_i)$ *for* $i = 1, 2$, *when*

$$\lambda \in \begin{cases} (0, +\infty) & \text{if } \ (\mu_1 - \mu_2)^2 \leq (\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^{-1} + (\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1})^{-1} \\ (0, \lambda_d) & \text{otherwise} \end{cases}. \tag{3.14}$$

As with the previous scenario, we remark that the condition

$$(\mu_1 - \mu_2)^2 \leq (\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^{-1} + (\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1})^{-1} \tag{3.15}$$

means the two ancestral states are similar and our method improves the ancestral state reconstruction with any $\lambda > 0$. In practice, since we often do not know whether condition (3.15) is satisfied, we also suggest to use:

$$\lambda = \frac{\hat{\lambda}_d}{2} = \frac{1}{2[(\hat{\mu}_1^{\mathrm{ML}} - \hat{\mu}_2^{\mathrm{ML}})^2 + (\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^{-1} + (\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1})^{-1}]}. \tag{3.16}$$

Since $m = 2$, our method and the ML method both require $O(n)$ time to compute the solutions using the tree traversal algorithm (Ho & Ané, 2014).

As another remark, our theoretical results in theorems 3.1 and 3.2 are useful and perhaps surprising. First, they point out that if the considered traits are related (i.e., (3.10) and (3.15) hold), the multi-task estimator is always better than ML for any value of the regularizer parameter λ. On the other hand, if they are not related, there still exists a small range of λ values that the multi-task estimator is better than ML, but this range would get smaller if the traits become more unrelated (for example, when $\max_i\{\mu_i\} - \min_i\{\mu_i\}$ gets larger). Nevertheless, there always exists some value of λ such that our method is better than ML, regardless of the relatedness between the traits. Thus, this implies the following corollary.

Corollary 3.1. *For the problem of ancestral state reconstruction under the Brownian motion model, the maximum likelihood method can be improved.*

The argument for this corollary is as follows: given a problem of reconstructing ancestral states of any trait on a fixed phylogenetic tree under the Brownian motion model, we can improve the accuracy of the maximum likelihood estimator by simultaneously reconstructing the ancestral states of interest and the ancestral states of a fixed template trait using Equation (3.13). The template trait can be chosen arbitrarily, as long as the evolution of the trait follows the Brownian motion model and can be created by simulating a BM trait along a fixed tree.

This surprising result reinforces a popular statistical observation, referred to as *Stein's paradox*, that leveraging data from multiple tasks can yield better performance over learning from each task independently, even if the underlying random variables come from seemingly unrelated distributions (Stein et al., 1956; Feldman et al., 2012). Most notably, Stein et al. (1956) showed that it is better (using MSE as the measure of accuracy) to estimate each of the means of multiple Gaussian random variables using data sampled from all of them. Our paper shows that such results still hold true for trait evolution on trees.

4. Simulations

We use simulations to illustrate the performance of our proposed multi-task learning method. We implement our method in `R` and apply the tree traversal algorithm proposed by Ho & Ané (2014) (implemented in the `R` package `phylolm`) to avoid inverting the phylogenetic correlation matrices. This package also provides a function for simulating traits along a phylogenetic tree under the BM model and a function for estimating the ancestral states using the ML estimators.

4.1. Comparing Multi-task and Maximum Likelihood Estimators

In this simulation, we compare the performance of the multi-task estimator with the standard ML method. We use the `rTrait` function in the `R` package `phylolm` to generate data according to the scenarios considered in this paper:

- Traits from the species of one phylogenetic tree: we simulate three independent continuous traits along the 4507-species mammal tree in Figure 1 under the BM model with $(\mu_i, \sigma_i^2) = (0,1), (1,1), (2,2)$ for $i = 1, 2, 3$ respectively.

- Traits from the species of two different phylogenetic trees: we simulate one trait along the mammal tree in Figure 1 under the BM model with $(\mu_1, \sigma_1^2) = (0, 1)$ and another trait along the 140-species phylogeny of ants in Figure 3 under the BM model with $(\mu_2, \sigma_2^2) = (2, 2)$.

The traits are standardized using the ML estimators $\mathbf{Y}'_i = \mathbf{Y}_i / \hat{\sigma}_i^{\mathrm{ML}}$ for $i = 1, 2, \ldots, m$, as suggested in section 3. Then $\{\hat{\mu}'_i\}_{i=1}^m$ are computed for $\{\mathbf{Y}'_i\}_{i=1}^m$ via (3.7) with $\lambda = \hat{\lambda}_s/2$ in the first scenario and via (3.13) with $\lambda = \hat{\lambda}_d/2$ in the second one. After that, we scale back $\{\hat{\mu}'_i\}_{i=1}^m$ to recover the estimated ancestral states $\{\hat{\mu}_i\}_{i=1}^m$ by $\hat{\mu}_i = \hat{\sigma}_i^{\mathrm{ML}} \hat{\mu}'_i$ for all $i = 1, 2, \ldots, m$. We also use the function `phylolm` to compute the ML estimators $\{\hat{\mu}_i^{\mathrm{ML}}\}_{i=1}^m$ for comparison. This procedure is repeated $1,000$ times and the MSE is estimated by the empirical MSE. Table 1 summarizes the results of this simulation.

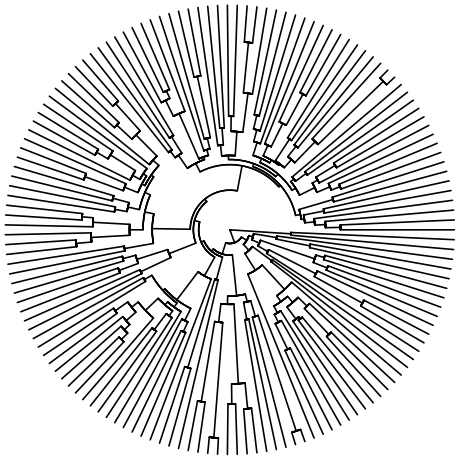


Figure 3: The 140-species phylogeny of ants from Moreau et al. (2006).

Compared to the ML method, our method returns a slightly biased estimators but reduces the MSE by 20% and the standard deviation by 10% for mammals. For ants, our method reduces the MSE by 12% and the standard deviation by 7%. This shows the proposed multi-task learning procedure improves the accuracy of ancestral state reconstruction.

4.2. Effect of the Regularizer Parameter λ

In this second simulation, we aim to investigate the behavior of our method as λ varies regarding to the conditions (3.10) and (3.15). We simulate two traits evolving independently along the mammal tree under the BM model with two settings:

- Condition (3.10) holds: $(\mu_1, \sigma_1^2) = (0, 1)$ and $(\mu_2, \sigma_2^2) = (2, 2)$.
- Condition (3.10) does not hold: $(\mu_1, \sigma_1^2) = (0, 1)$ and $(\mu_2, \sigma_2^2) = (16, 2)$.

In both settings, we reconstruct the ancestral states using (3.7) with $\lambda = 0,\ 1.25 \times 10^{-3},\ 2.5 \times 10^{-3},\ 5 \times 10^{-3},\ 7.5 \times 10^{-3},\ 10^{-2},\ 2 \times 10^{-2}$, and 3×10^{-2}. Note that $\lambda = 0$ corresponds to the ML method. To estimate the MSEs, we also repeat this procedure $1,000$ times. Figure 4(a) shows that when the condition (3.10) holds, our method outperforms the ML method for all λ. On the other

Scenario		Same set of species			Different sets of species	
Species		Mammals			Mammals	Ants
Trait		1	2	3	1	2
(μ_i, σ_i^2)		$(0,1)$	$(1,1)$	$(2,2)$	$(0,1)$	$(2,2)$
	MSE	33.59	33.22	65.68	34.54	31.58
ML	Mean	-0.1	0.99	2	-0.34	1.97
	Sd	5.63	5.73	8.02	5.63	5.44
	MSE	27.31	27.15	52.84	25	27.5
Multi-task	Mean	0.04	0.96	1.85	-0.08	1.8
	Sd	5.07	5.17	7.19	4.77	5.08

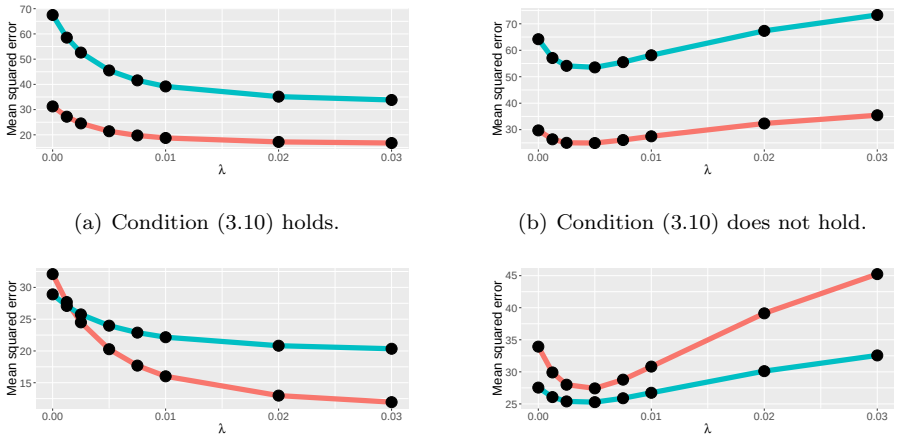
Table 1: Estimated MSEs, means and standard deviations (Sd) of multi-task learning and ML method for reconstructing ancestral states.

hand, when the condition (3.10) does not hold, our method only outperforms the ML method for small λ (Figure 4(b)).

We also repeat the simulation for condition (3.15) using the mammal tree for the first trait (μ_1, σ_1^2) and the ant tree for the second trait (μ_2, σ_2^2) in the following two settings:

- Condition (3.15) holds: $(\mu_1, \sigma_1^2) = (0,1)$ and $(\mu_2, \sigma_2^2) = (2,2)$.
- Condition (3.15) does not hold: $(\mu_1, \sigma_1^2) = (0,1)$ and $(\mu_2, \sigma_2^2) = (16,2)$.

In these cases, our method behaves similarly to the simulation for the condition (3.10) (see Figures 4(c) and 4(d)). The results show that it is necessary to be conservative when choosing λ. The simulations also suggest that the gain from using the multi-task estimator is larger when the condition (3.10) or (3.15) is satisfied.



(a) Condition (3.10) holds.

(b) Condition (3.10) does not hold.

(c) Condition (3.15) holds.

(d) Condition (3.15) does not hold.

Figure 4: MSEs of the multi-task learning method with respect to λ. Note that $\lambda = 0$ is the ML estimators. For the first trait (red), $(\mu_1, \sigma_1^2) = (0, 1)$. For the second trait (blue), $(\mu_2, \sigma_2^2) = (2, 2)$ in 4(a), and 4(c), and $(\mu_2, \sigma_2^2) = (16, 2)$ in 4(b) and 4(d).

5. Discussion and Conclusion

Our paper proposed and analyzed a new multi-task estimator for ancestral state reconstruction. This estimator uses the regularized maximum likelihood method to reconstruct the ancestral states of multiple traits simultaneously. Our theoretical results show the advantage of the proposed method compared to the usual independent maximum likelihood approach for the problem. We confirm our theories using several simulated data sets from the phylogenies of mammals and ants with known ancestral states. Our multi-task learning method provides slightly biased estimators but can reduce their standard deviations, leading to better MSEs compared to the ML estimators. The simulations also verify that our method always outperforms ML method when the regularizing parameter λ is small enough.

The idea in this paper can also be applied to other trait evolutionary models such as the phylogenetic two-state model (see e.g. Li et al., 2008), the phylogenetic threshold model (see e.g. Felsenstein, 2011), and the Ornstein-Uhlenbeck model (see e.g. Ho & Ané, 2013). However, the theoretical approach in this

paper relies heavily on the Gaussian models. Therefore, extending our results to non-Gaussian models is not straightforward. On the other hand, while we only consider the ℓ_2-penalty in our framework, the shrinkage effect has been observed on a wide class of penalty functions. For that reason, the same theoretical results might hold for other penalties such as ℓ_1 and SCAD (Fan & Li, 2001).

Our method can also be applied to reconstruct the state at any internal node by re-rooting the tree to that node. Note that this re-rooting technique, which has been applied for the ML estimators (see Goolsby, 2017, and the references therein), is appropriate because the Brownian motion is time-reversible. Therefore, one must be cautious when using the technique for other models.

We note that in the context of ancestral state reconstruction, the accuracy of an ML estimator depends on the structure of the tree rather than the sample size (number of tips). For example, Ané (2008) introduces the notion of effective sample size, which depends on the tree, to measure how much information is contained in a given data set. Similarly, the accuracy of our method depends on the structure of the tree and the correlation between traits.

Acknowledgments

LSTH was supported by startup funds from Dalhousie University, the Canada Research Chairs program, and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2018-05447. CVN was supported by EPSRC grant EP/M0269571.

Appendix A. Proof of Theorem 3.1

For $i = 1, 2, \ldots, m$, we can compute the MSE of $\hat{\mu}_i$ as follows:

$$\begin{aligned}\mathrm{MSE}(\hat{\mu}_i) &= (\mathbb{E}\hat{\mu}_i - \mu_i)^2 + \mathrm{Var}(\hat{\mu}_i) \\ &= \frac{4\lambda^2(\sum_{k=1}^m \mu_k - m\mu_i)^2}{(\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1} + 2\lambda m)^2} \\ &\quad + \frac{(\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1})^2 + 4\lambda^2\left(\sum_{k,l} c_{kl}\right) + 4\lambda\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}\left(\sum_{k=1}^m c_{ik}\right)}{\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1} + 2\lambda m)^2}.\end{aligned}$$

Recall that $\mathrm{MSE}(\hat{\mu}_i^{\mathrm{ML}}) = (\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1})^{-1}$. Thus, we have:

$$\begin{aligned}\mathrm{MSE}(\hat{\mu}_i^{\mathrm{ML}}) - \mathrm{MSE}(\hat{\mu}_i) &= \frac{4\lambda\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}[m - (\sum_{k=1}^m c_{ik})]}{\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1} + 2\lambda m)^2} \\ &\quad + \frac{4\lambda^2\left[m^2 - \mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(\sum_{k=1}^m \mu_k - m\mu_i)^2 - \left(\sum_{k,l} c_{kl}\right)\right]}{\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1} + 2\lambda m)^2}. \quad \text{(A.1)}\end{aligned}$$

Note that: $m - \left(\sum_{k=1}^m c_{ik}\right) \geq (m-1)(1 - \max_{k\neq l} c_{kl}) > 0$.

Therefore, if $(\max_i\{\mu_i\} - \min_i\{\mu_i\})^2 \leq \dfrac{m^2 - \left(\sum_{k,l} c_{kl}\right)}{(m-1)^2\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}}$, then

$$m^2 - \mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(\sum_{k=1}^m \mu_k - m\mu_i)^2 - \left(\sum_{k,l} c_{kl}\right) \geq 0.$$

Thus, $\mathrm{MSE}(\hat{\mu}_i^{\mathrm{ML}}) > \mathrm{MSE}(\hat{\mu}_i)$ for every $\lambda > 0$.

Otherwise, $\mathrm{MSE}(\hat{\mu}_i^{\mathrm{ML}}) > \mathrm{MSE}(\hat{\mu}_i)$ when

$$\lambda < \frac{\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(m-1)(1 - \max_{k\neq l} c_{kl})}{\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(\sum_{k=1}^m \mu_k - m\mu_i)^2 + \left(\sum_{k,l} c_{kl}\right) - m^2}.$$

We also notice that $\lambda_s \leq \dfrac{\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(m-1)(1 - \max_{k\neq l} c_{kl})}{\mathbf{1}^\top\mathbf{V}^{-1}\mathbf{1}(\sum_{k=1}^m \mu_k - m\mu_i)^2 + \left(\sum_{k,l} c_{kl}\right) - m^2}$ for all $i = 1, 2, \ldots, m$.

Thus, the theorem holds.

Appendix B. Proof of Theorem 3.2

From (3.13), we have:

$$(\mathbb{E}\hat{\mu}_1 - \mu_1)^2 = \frac{4\lambda^2(\mu_1 - \mu_2)^2}{(\mathbf{1}^\top\mathbf{V}_1^{-1}\mathbf{1})^2} \Bigg/ \left(1 + \frac{2\lambda}{\mathbf{1}^\top\mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top\mathbf{V}_2^{-1}\mathbf{1}}\right)^2.$$

Note that

$$\hat{\mu}_1 = \left[\left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)\frac{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{Y}_1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}}\frac{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{Y}_2}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right] \Bigg/ \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)$$

and $\mathbf{Y}_1$ is independent of $\mathbf{Y}_2$. Hence,

$$\mathrm{Var}(\hat{\mu}_1) = \left[\left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)^2 \frac{1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{4\lambda^2}{(\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^2}\frac{1}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right] \Bigg/ \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)^2.$$

Therefore, we have:

$$\begin{aligned}\mathrm{MSE}(\hat{\mu}_1) &= (\mathbb{E}\hat{\mu}_1 - \mu_1)^2 + \mathrm{Var}(\hat{\mu}_1) \\ &= \left[\frac{4\lambda^2(\mu_1-\mu_2)^2}{(\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^2} + \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)^2 \frac{1}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{4\lambda^2}{(\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^2}\frac{1}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right] \Bigg/ \\ &\quad \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)^2.\end{aligned}$$

Recall that $\mathrm{MSE}(\hat{\mu}_1^{\mathrm{ML}}) = (\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^{-1}$. So, $\mathrm{MSE}(\hat{\mu}_1^{\mathrm{ML}}) > \mathrm{MSE}(\hat{\mu}_1)$ is equivalent to:

$$\frac{4\lambda^2(\mu_1-\mu_2)^2}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)^2 + \frac{4\lambda^2}{(\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})(\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1})} < \left(1 + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{2\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}}\right)^2,$$

which means

$$\lambda(\mu_1-\mu_2)^2 < \frac{\lambda}{\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1}} + \frac{\lambda}{\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1}} + 1.$$

Therefore, we conclude that if $(\mu_1-\mu_2)^2 \le (\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^{-1} + (\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1})^{-1}$, then $\mathrm{MSE}(\hat{\mu}_1^{\mathrm{ML}}) > \mathrm{MSE}(\hat{\mu}_1)$ for any $\lambda > 0$.

Otherwise, if $\lambda < \dfrac{1}{(\mu_1-\mu_2)^2 - (\mathbf{1}^\top \mathbf{V}_1^{-1}\mathbf{1})^{-1} - (\mathbf{1}^\top \mathbf{V}_2^{-1}\mathbf{1})^{-1}}$, then we also have $\mathrm{MSE}(\hat{\mu}_1^{\mathrm{ML}}) > \mathrm{MSE}(\hat{\mu}_1)$.

The above argument can also be applied for $\hat{\mu}_2$, which completes the proof.

References

Ané, C. (2008). Analysis of comparative data with hierarchical autocorrelation. *The Annals of Applied Statistics*, (pp. 1078–1102).

Argyriou, A., Evgeniou, T., & Pontil, M. (2006). Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 41–48).

Baum, D. A., Ané, C., Larget, B., Solís-Lemus, C., Ho, L. S. T., Boone, P., Drummond, C. P., Bontrager, M., Hunter, S. J., & Saucier, W. (2016). Statistical evidence for common ancestry: Application to primates. *Evolution*, *70*, 1354–1363.

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A. et al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, *523*, 217–220.

Bininda-Emonds, O., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L., & Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, *446*, 507–512.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., & Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. *Science*, *337*, 957–960.

Cooper, N., & Purvis, A. (2010). Body size evolution in mammals: complexity in tempo and mode. *The American Naturalist*, *175*, 727–738.

Davis, C. C., Latvis, M., Nickrent, D. L., Wurdack, K. J., & Baum, D. A. (2007). Floral gigantism in rafflesiaceae. *Science*, *315*, 1812–1812.

Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-task learning for multiple language translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1723–1732).

Evgeniou, T., & Pontil, M. (2004). Regularized multi-task learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 109–117).

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J. et al. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, *346*, 56–61.

Feldman, S., Gupta, M., & Frigyik, B. (2012). Multi-task averaging. In *Advances in Neural Information Processing Systems* (pp. 1169–1177).

Feldman, S., Gupta, M. R., & Frigyik, B. A. (2014). Revisiting Stein's paradox: multi-task averaging. *Journal of Machine Learning Research*, *15*, 3441–3482.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.

Felsenstein, J. (2011). A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist*, *179*, 145–156.

Gill, M. S., Ho, L. S. T., Baele, G., Lemey, P., & Suchard, M. A. (2017). A relaxed directional random walk model for phylogenetic trait evolution. *Systematic Biology*, *66*, 299.

Goolsby, E. W. (2017). Rapid maximum likelihood ancestral state reconstruction of continuous characters: A rerooting-free algorithm. *Ecology and Evolution*, *7*, 2791–2797.

Ho, L. S. T., & Ané, C. (2013). Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models. *The Annals of Statistics*, *41*, 957–981.

Ho, L. S. T., & Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, *63*, 397–408.

Li, G., Steel, M., & Zhang, L. (2008). More taxa are not necessarily better for the reconstruction of ancestral character states. *Systematic Biology*, *57*, 647–653.

Li, J., Tian, Y., Huang, T., & Gao, W. (2010). Probabilistic multi-task learning for visual saliency estimation in video. *International Journal of Computer Vision*, *90*, 150–165.

Lu, W., Chieu, H. L., & Löfgren, J. (2016). A general regularization framework for domain adaptation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Moreau, C. S., Bell, C. D., Vila, R., Archibald, S. B., & Pierce, N. E. (2006). Phylogeny of the ants: diversification in the age of angiosperms. *Science*, *312*, 101–104.

Revell, L. J., & Harmon, L. J. (2008). Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research*, *10*, 311–331.

Stein, C. et al. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Zhang, J., Ghahramani, Z., & Yang, Y. (2006). Learning multiple related tasks using latent independent component analysis. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1585–1592).

Zhang, T., Ghanem, B., Liu, S., & Ahuja, N. (2012). Robust visual tracking via multi-task sparse learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2042–2049).