Submitted to the Annals of Applied Statistics

VARIANCE COMPONENT REGULARIZATION FOR THE INFERENCE OF NEUTRAL EVOLUTION

By Kevin H. Eng and Sündüz Keleş

University of Wisconsin-Madison

We represent the Brownian model for continuous traits with phylogenetic dependence as a variance components model, allowing for the novel estimation of edge weights for every branch of a tree graph. We show that maximum likelihood estimation carries an inherent bias often shrinking internal edges to zero, which is troublesome because this estimate is intended to compare covariance structures associated with neutral drift and selection. To correct this problem, we develop a new estimator that parameterizes a path between the maximum likelihood and neutral graphs leading to more stable estimation and the retention of graphical structure. Example analyses highlight the descriptive power of these estimates. We also link insights in this evolutionary context to general concerns about predicting random effects and estimating hierarchical variance components in linear mixed models when viewed in a penalized likelihood framework.

1. Introduction to neutral evolution inference.

It is the goal of neutral evolution or neutral drift inference, a form of comparative method (Felsenstein, 1985; Martins and Hansen, 1997), to infer whether a Brownian model (Felsenstein) provides an adequate description of phenotypic variation through genotypic divergence. In addition to the topology of the genotype tree, we have a set of corresponding neutral branch lengths which reflect the expected amount of neutral drift variation on each branch of the tree.

A strict hypothesis test for neutral evolution (Fay and Wittkopp, 2008) may take the form of a test of variance components. Nuzhdin *et al.* (2004); Rifkin, Kim and White (2003) and Whitehead and Crawford (2006) use ratios of scalar variances to say whether variation is bigger or smaller than expected. Gu (2004); Oakley *et al.* (2005) and Guo *et al.* (2007) consider likelihood based model selection methods to choose among a set of prespecified models.

An alternative approach is to estimate the hierarchically structured covariance based on a sample which may be represented by a mixed effects model. Then, the problem reduces to inferring which evolutionary scenario is the best description of the apparent dependence structure. Lynch (1991) and

Keywords and phrases: Comparative method, Neutral evolution, Penalized estimation, Random effects, Variance components

Housworth, Martins and Lynch (2004) study mixed effects models where a vector of normal variates, say b with covariance $\Sigma(\theta)$ representing the phylogenetic effect for strains or species, is observed indirectly by measuring individuals within these taxa. They assume that $\Sigma(\theta)$ is fully known up to scale parameter $\tau^2 > 0$ representing between taxa variation. Butler and King (2004) vary the structure of $\Sigma(\theta)$ according to a non-linear transformation predicted by an Ornstein-Uhlenbeck process. Freckleton, Harvey and Pagel (2002); Guo *et al.* (2007) and Eng, Corrada Bravo and Keleş (2009) relax these assumptions to allow for environmental error. In all cases, the topology of the tree is assumed to be known.

While all of the methods above allow only the coarse determination of selection across the whole tree under a specific alternative hypothesis, we propose to estimate all branch lengths directly, allowing the selective force to be described in detail. We fully relax the assumptions on the branch length parameters in Section 3. Having derived the maximum likelihood estimate of the branch lengths, we observe that it prefers to amass covariation in terminal branches leading to tree estimates that are counterintuitive for evolutionary inference. In Section 4, we develop a regularized estimator which parameterizes a path of tree estimates between the neutral and maximum likelihood trees. Our regularized estimator corresponds to an estimate on this path and can be represented as a weighted average of the estimates under these two tree models. We present simulation studies illustrating how insensitive model selection methods are for comparing whole mixed model covariance structures and supporting our regularized estimator in Section 5. In addition, we show that our regularized estimator stabilizes the maximum likelihood estimator, shrinking back to the neutral tree when there is insufficient evidence for selection and retaining data driven features when they are important. This path of estimates also subsumes both the uncertainty characterized by the hypothesis test and its end goal: characterizing the evident alternative, selective pressure when it is inferred.

2. Introduction to tree-structured covariance.

Let b be a $(m \times 1)$ multivariate normal vector whose entries, called taxa, represent levels of a factor with dependence induced by common evolutionary history. The Brownian motion model for the evolution of continuous traits postulates that the variance of an observed taxon is proportional to its total evolutionary time and the covariation between two of the m observed taxa is proportional to their shared evolutionary time.

Phylogenetic trees are graphical representations of this structured covariation which may be written as $V\Theta V^T$ for a known topology-encoding matrix

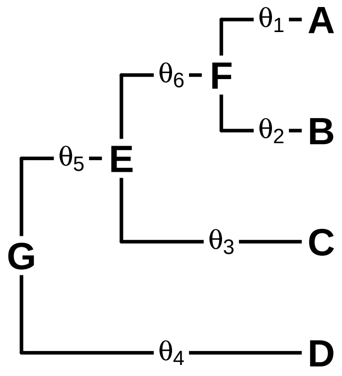

FIG 1. *An example phylogenetic tree for comparison with example covariance matrix. Letters are nodes reflecting random variables and θ edge weights are branch length parameters.*

$V = (v_1, v_2, \ldots v_{2m-2})$ and branch lengths $\Theta = diag(\theta_1, \theta_2, \ldots, \theta_{2m-2})$, $\theta_k \geq 0$. The topology vectors v_k have binary entries with $v_{jk} = 1$ if the kth branch appears in the path between the root and the jth taxon. For example, the following basis matrix V with labeled columns and the accompanying covariance matrix correspond to the phylogenetic graph in Fig. 1.

$$V = \begin{array}{c} \begin{matrix} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \mathrm{E} & \mathrm{F} \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}, \; V\Theta V^T = \begin{pmatrix} \theta_1 + \theta_5 + \theta_6 & \theta_5 + \theta_6 & \theta_5 & 0 \\ \theta_5 + \theta_6 & \theta_2 + \theta_5 + \theta_6 & \theta_5 & 0 \\ \theta_5 & \theta_5 & \theta_3 + \theta_5 & 0 \\ 0 & 0 & 0 & \theta_4 \end{pmatrix}.$$

This directed graph is read from left to right and represents dependence between the nodes in the sense that nodes A and B are conditionally independent given node F. Each node represents a normal random variable of which we observe only the taxa A, B, C, and D with covariance $V\Theta V^T$. The internal nodes, E, F, and G, represent unobserved common ancestors and the root node, G, is assumed to be a constant.

We call edge weights, θ_k, branch lengths noting that vertical distances on phylograms are ignored; length is measured as the projection onto the horizontal axis. We will call the edges between internal nodes internal branches and the edges between internal and terminal nodes terminal branches.

3. Maximum likelihood estimation.

3.1. *Estimation.* Let $Y = W\beta + Zb + e$ be a $(n\times 1)$ vector of observed trait values where binary-valued $Z = (z_{ij})$, $z_{ij} = 1$ identifies the ith subject as a member of the jth taxon. We assume that random effect $b \sim \mathcal{N}(0, \sigma^2 V\Theta V^T)$ has a phylogenetic tree covariance independent of environmental error $e \sim \mathcal{N}(0, \sigma^2 I_n)$ and include fixed effects β with $(n\times p)$ design matrix W to account for confounding environmental effects. We emphasize that branch lengths $\theta_k \geq 0$ are the scalar variances of random effects associated with column vectors $X_k = Zv_k$:

$$\begin{aligned}\mathrm{var}(Y) &= \sigma^2 I_n + \sigma^2 ZV\Theta V^T Z^T,\\ &= \sigma^2 I_n + \sum_{k=1}^{2m-2} \sigma^2 \theta_k X_k X_k^T.\end{aligned}$$

As σ^2 is a scale for phenotypic variation, this representation is for convenience: inferences about θ_k are typically scale-free. When required, we set $\sigma^2\theta_k = \tau^2\widetilde{\theta}_k$ where τ^2 is identified from θ_k using the tree norm proposed in Ives, Midford and Garland (2007): setting the maximum graph distance between taxa to be one.

The likelihood of Y is

$$\begin{aligned}L(\theta, \sigma^2, \beta; Y) = \left(2\pi\sigma^2\right)^{-n/2} \left|I_n + ZV\Theta V^T Z^T\right|^{-n/2} \times \\ \exp\left\{-\frac{1}{2\sigma^2}(Y - W\beta)^T \left(I + ZV\Theta V^T Z^T\right)^{-1}(Y - W\beta)\right\}.\end{aligned}$$

Maximum likelihood estimation can be done by Expectation Maximization (Dempster, Laird and Rubin, 1977), described in the appendix, or similar numerical methods (Bates and Maechler, 2010; McCulloch and Searle, 2001). We focus on the estimation of the structure of the random effect variance in the following remarks.

Remark 1 (Maximum likelihood as penalized estimation). For random vector $Y \sim N(0, I_n + X\Theta X^T)$, the maximum likelihood estimate is the solution to

$$\min_\theta Y^T \left(I_n + X\Theta X^T\right)^{-1} Y + \log\left|I_n + X\Theta X^T\right|.$$

Following Bates and Maechler (2010), the objective may be written as a twice penalized least squares problem where $\hat{u}_\theta$ satisfies the normal equations $(\Theta^{1/2}X^T X\Theta^{1/2} + I_{2m-2})u = \Theta^{1/2}X^T Y$:

$$\min_\theta \left\|Y - X\Theta^{1/2}\hat{u}_\theta\right\|^2 + \hat{u}_\theta^T \hat{u}_\theta + \log\left|I_n + X\Theta X^T\right|.$$

In the above representation, the first term measures data fidelity, the fit of the random effects, while the second term is a ridge-like penalty restraining the predicted random effect and the last term prefers diagonally dominant covariances. We study these penalties in detail as follows.

3.2. *Consideration of the penalty.* We make two observations about the undesirable properties of the maximum likelihood estimate and illustrate them with simple analytical examples.

Under the variance component or tree constraint (assumption 1), we interpret θ_k as a branch length. But, the likelihood remains valid so long as $\Omega_\theta = I_n + X\Theta X^T$ remains positive definite (assumption 2). Formally, we write:

ASSUMPTION 1 (Tree constraint). $\Theta = (\theta_1, \ldots, \theta_{2m-2})$ such that $\theta_k \geq 0$ for each k.

ASSUMPTION 2 (Positive definiteness). $\Theta = (\theta_1, \ldots, \theta_{2m-2})$ constrained such that Ω_θ is positive definite ($\Omega_\theta > 0$).

Clearly, all Θ satisfying tree constraints lead to positive definite Ω_θ. The effect of this constraint is not negligible; the constrained analysis of variance estimating equations are equivalent to the restricted maximum likelihood (that is, REML) equations and often yield solutions on the boundary (McCulloch and Searle, 2001), which have an extreme interpretation in an evolutionary context (Houswort h, Martins and Lynch, 2004). Hill and Thompson (1978) demonstrate that the maximum likelihood estimates of variance parameters prefer to default to zero when there is too little likelihood based information otherwise. We examine these points as follows.

EXAMPLE 1. Suppose that $Y \sim \mathcal{N}(0, \Omega_1)$. For $\Omega_1 = \theta_1 X_1 X_1^T + I_n$, the characteristic polynomial is $(1-\lambda)^{(n-1)}(1-\lambda+\theta_1 X_1^T X_1)$, implying that Ω_1 is positive definite when $\theta_1 > -1/X_1^T X_1$. Therefore, it is possible that the maximizer of the likelihood might be found outside the range of θ restricted to trees (assumption 1).

REMARK 2. The expected good behavior of the maximum likelihood estimate is not guaranteed because two regularity conditions (Lehmann and Casella, 1998) do not hold. First, under (assumption 1), the parameter space is a closed set, $\theta_k \geq 0$, and we illustrate how the likelihood is often forced to choose local maxima in the next example. Second, data do not accumulate independently; that is, $f(Y|\sigma^2, \beta, \theta) \neq \prod_{ij} f(Y_{ij}|\sigma^2, \beta, \theta)$.

TABLE 1
Example probabilities of negative maximum likelihood estimates.

n	1	5	10	20	100	1000
$\Pr\left(\hat{\theta}_M < 0\right)$	0.5205	0.3169	0.2370	0.1727	0.0793	0.0252

EXAMPLE 2. Continuing example 1, suppose that $Y \sim \mathcal{N}(0, \theta_1 X_1 X_1^T + I_n)$, with density:

$$f(Y|\theta_1) = (2\pi)^{-\frac{n}{2}}(1+\theta_1 X_1^T X_1)^{-\frac{1}{2}} \exp\left\{-\frac{Y^T Y}{2} + \frac{\theta_1 Y^T X_1 X_1^T Y}{2(1+\theta_1 X_1^T X_1)}\right\}.$$

We may derive the maximum likelihood estimator directly:

$$\hat{\theta}_M = \frac{Y^T X_1 X_1^T Y - X_1^T X_1}{(X_1^T X_1)^2}.$$

We observe that $Y^T X_1 X_1^T Y$ is the square of a (scalar) normal random variable with variance $(X_1^T X_1)(1+\theta_1 X_1^T X_1)$, so that $Y^T X_1 X_1^T Y/[X_1^T X_1(1+\theta_1 X_1^T X_1)] \sim \chi_1^2$. It follows that

$$\Pr\left(\hat{\theta}_M < 0\right) = \Pr\left(\chi_1^2 < \frac{1}{1+\theta_1 X_1^T X_1}\right).$$

For finite sample cases, the likelihood's global maximum is a negative value of θ_1 with positive probability. Under assumption 1, we would use the local maximum satisfying the non-negativity constraint and set the estimate exactly to zero.

In balanced studies, $X_1^T X_1 = n$ is the number of individuals in each taxon. As a function of n, at $\theta_1 = 1$, the probabilities in Table 1 show that for common study sizes ($n = 5$ to $n = 20$) there is a significant chance of a negative estimate.

REMARK 3 ($\log\left|I_n + X\Theta X^T\right|$ as a penalty). Let $\Omega_\theta = I_n + X\Theta X^T$, $X = ZV$ where $Z^T Z = nI_m$ represents balanced observations from the m taxa. Applying Sylvester's determinant theorem (Seber and Lee, 2003), we observe that

$$\log\left|I_n + X\Theta X^T\right| = \log\left|I_m + nV\Theta V^T\right|.$$

If $V\Theta V^T$ is positive semi-definite, it follows that $\log\left|I_m + V\Theta V^T\right| \geq \log|I_m| = 0$ for all Θ. The function is therefore non-negative, taking value zero only when $\Theta = 0$. Further, we observe in the following example that, as a penalty, the log determinant favors terminal branches.

EXAMPLE 3. Consider the determinant of $\Omega_2 = I_n + \theta_1 X_1 X_1^T + \theta_2 X_2 X_2^T$. Following the representation in Corrada Bravo *et al.* (2009), suppose we have 3 taxa $\{A, B, C\}$ with n balanced observations each ($Z^T Z = nI_3$) and that $v_1 = (1,0,0)^T$ and $v_2 = (1,1,0)^T$, where v_1 identifies the partition $\{A\}$ and is contained in v_2, the partition $\{A, B\}$. Graphically, v_2 is an internal branch and v_1 is a terminal branch. Setting $X_k = Zv_k$, it follows that $X_1^T X_1 = n$, $X_2^T X_2 = 2n$ and $X_1^T X_2 = n$. By the matrix determinant lemma and Sherman-Morrison formula (Seber and Lee, 2003):

$$\begin{aligned}\left|I_n + \theta_1 X_1 X_1^T + \theta_2 X_2 X_2^T\right| &= (1 + \theta_1 X_1^T X_1)(1 + \theta_2 X_2^T X_2) - \theta_1\theta_2 (X_1^T X_2)^2 \\ &= 1 + \theta_1 n + 2\theta_2 n - \theta_1\theta_2 n^2\end{aligned}$$

For $\theta = (\theta_1, \theta_2)^T$, we note that

$$\frac{\partial}{\partial\theta^T}|\Omega_2| = \left(n - \theta_2 n^2, 2n - \theta_1 n^2\right).$$

We are interested in the sensitivity of the log determinant for small values of $\theta \geq 0$. For values of θ close to 0, $\Omega_2 \approx I_n$ so that the gradient of the penalty is

$$\frac{\partial}{\partial\theta^T}\log|\Omega_2| = \frac{1}{|\Omega_2|}\frac{\partial}{\partial\theta^T}|\Omega_2| \approx (n, 2n)$$

This result suggests that a small perturbation in θ_2 (the nested, internal branch) is twice as costly as the same perturbation in θ_1 (the terminal branch). The net effect of this cost is that maximum likelihood estimate of Ω_2 prefers smaller off-diagonal entries when possible.

4. Regularized estimation. To incorporate the aims of neutral drift inference and to correct for the maximum likelihood estimate's propensity to shrink covariation to zero, we propose to penalize the deviations of estimates $\hat{\theta}_k$ from their known neutral tree values t_k. When the deviation is small, the estimated phenotypic branch length is near the genotypic branch length and the penalty should be small. That is, we propose the penalized estimator of (θ, σ^2, β) minimizing

$$-\log f(Y; \theta, \beta, \sigma^2) + \lambda J(\theta),$$

where $J(\theta) \geq 0$ is a penalty function which takes value zero when the estimated branches are set at the neutral branches, i.e., $J(t) = 0$.

From an average risk perspective, the choices of penalties are effectively priors. Because the maximum likelihood penalty is based on $\vartheta_k = 1 + \theta_k X_k^T X_k$, we consider a scaled inverse χ^2 density for ϑ_k with prior mode $1 + t_k X_k^T X_k$. We choose degrees of freedom ν_k and scale σ_k^2 to satisfy

$$\nu_k \sigma_k^2 = \lambda, \quad \frac{\nu_k \sigma_k^2}{\nu_k + 2} = 1 + t_k X_k^T X_k.$$

After transformation, the prior density for θ_k is proportional to

$$p(\theta_k) \propto \exp\left\{ -\frac{\lambda}{2} \frac{1 + t_k X_k^T X_k}{1 + \theta_k X_k^T X_k} - \frac{\lambda}{2} \log(1 + \theta_k X_k^T X_k) \right\}.$$

We note that the posterior mode is the minimizer of $-\log f(Y|\theta) - \sum_k \log p(\theta_k)$ and equivalently the minimizer of

$$-\log f(Y|\theta) + \frac{\lambda}{2} \sum_k \left\{ (\frac{1 + t_k X_k^T X_k}{1 + \theta_k X_k^T X_k} + \log(1 + \theta_k X_k^T X_k) \right\},$$

which leads to our choice of penalty function (satisfying $J(t) = 0$):

$$J(\theta) = \sum_{k=1}^{2m-2} \left\{ \left(\frac{1 + t_k X_k^T X_k}{1 + \theta_k X_k^T X_k} - 1 \right) + \log \left(\frac{1 + \theta_k X_k^T X_k}{1 + t_k X_k^T X_k} \right) \right\}.$$

The tuning parameter $\lambda \geq 0$ can be chosen by likelihood-based cross-validation (van der Laan, Dudoit and Keleş, 2004), which chooses the parameter λ maximizing the log validation likelihood.

Example 4. Continuing example 2, we study a mean zero normal variate with $\mathrm{var}(Y) = \theta_1 X_1 X_1^T + I_n$. If we have $t_1 > 0$, the known neutral value of θ_1, the penalized log likelihood estimate minimizes

$$-\frac{\theta_1 Y^T X_1 X_1^T Y}{1 + \theta_1 X_1^T X_1} + \log(1 + \theta_1 X_1^T X_1) + \lambda J(\theta_1).$$

As a function of the tuning parameter λ, it can be shown that the estimator, $\hat{\theta}_A(\lambda)$, is

$$\begin{aligned} \hat{\theta}_A &= \frac{1}{1+\lambda} \left(\frac{Y^T X_1 X_1^T Y - X_1^T X_1}{(X_1^T X_1)^2} \right) + \frac{\lambda}{1+\lambda} t_1 \\ &= \frac{1}{1+\lambda} \hat{\theta}_M + \frac{\lambda}{1+\lambda} t_1. \end{aligned}$$

Quite reasonably the average risk estimator under squared error loss is a weighted average of the data-driven maximum likelihood estimate and the known, prior mode t_1. As expected, $\lambda = 0$ corresponds to the maximum likelihood estimate while $\hat{\theta}_A \to t_1$ as $\lambda \to \infty$. This is shrinkage towards the neutral tree.

4.1. *Effective degrees of freedom.* We conclude in the previous section that $\lambda \geq 0$ can be considered to be a parameter indexing a path of tree models, Θ_λ, between the neutral tree and the maximum likelihood tree. Recall from Remark 1 that the phylogenetic effect $b \sim \mathcal{N}(0, V\Theta V^T)$ can be written as $b = V\Theta^{1/2}u$, where $u \sim \mathcal{N}(0, I_{2m-2})$. Therefore, a predicted random effect $\hat{u}_\lambda$ should satisfy the normal equation $(\Theta_\lambda^{1/2} X^T X \Theta_\lambda^{1/2} + I_{2m-2})u = \Theta_\lambda^{1/2} X^T Y$, reflecting the change of basis corresponding to an orthogonal set of random effects.

As in shrinkage methods like ridge regression, we suggest that the model complexity corresponding to estimate Θ_λ be measured by the effective degrees of freedom of the predicted $\hat{u}_\lambda$ (Hastie, Tibshirani and Friedman, 2001; Wahba *et al.*, 1995):

$$df(\lambda) = tr\left[X\Theta_\lambda^{1/2}\left(\Theta_\lambda^{1/2} X^T X \Theta_\lambda^{1/2} + I_{2m-2}\right)^{-1}\Theta_\lambda^{1/2} X^T\right].$$

We utilize this degrees of freedom when comparing the regularized estimator with estimators from different models.

5. Simulation studies of the estimators.

5.1. *Maximum likelihood loss of graph structure.* We observed that analytic estimates of the covariance attributable to a single branch have an unexpected propensity to be estimated at zero because of the constrained parameter space. Because $I_n + X\Theta X^T$ is still positive definite when any or all $\hat{\theta}_k = 0$, these cases represent valid covariance matrices corresponding to trees with deleted branches. The following simulation reinforces the prevalence of the problem when considering multiple branches: for an extremely simple tree with a large number of replicate individuals, we observe that the full structure is never fully estimated.

We fit the ML estimate to 10,000 simulated random vectors $Y \sim \mathcal{N}(0, I_{60} + ZV\Theta V^T Z^T)$ where $Z = I_3 \otimes 1_{20}$,

$$V = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \text{and} \quad \Theta = \frac{1}{3}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In Fig. 2, we plot the histogram of topological structures ordered by the number of non-zero branches. The true generating topology is the far right graph. For every simulated vector, at least one branch was lost. We anticipated this result from the consideration of the maximum likelihood penalty

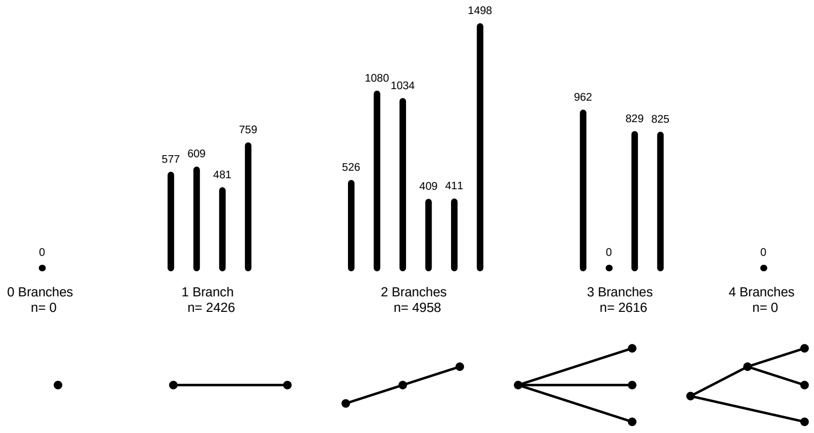

Fig 2. *Counts of topologies selected by maximum likelihood for 10,000 simulated data sets. Example topologies are shown below the axis.*

in the previous sections but did not anticipate the degree of the loss of structure. For neutral inference, the tendency to lose branches implies extreme evolutionary interpretations and will lead to anti-conservative inference.

5.2. *Model selection versus regularized estimation.* In neutral evolution inference, our goal is to infer whether the neutral, genotypic tree is an adequate description of the covariance. We have established that the maximum likelihood estimate is too extreme: it favors low cost zero branch lengths which have a less satisfactory interpretation (Housworth, Martins and Lynch, 2004). We contrast this behavior with regularized estimation in the context of neutral inference in the following study.

In Fig. 3, we show two trees of the same topology (up to zero-length branches) with lengths defined by θ_G, derived from genotypic data; and by θ_P, estimated from phenotypic data using maximum likelihood. We stress that these estimates are based on the mouse strain data used in the next section and reflect conditions found in an applied problem.

We generated 10,000 random vectors $Y_G \sim \mathcal{N}(0, I_{35n} + 2ZV\Theta_G V^T Z^T)$ and 10,000 vectors $Y_P \sim \mathcal{N}(0, I_{35n} + 2ZV\Theta_P V^T Z^T)$ where $Z = I_{35} \otimes I_n$ and allowing $n = 5$ and $n = 10$ individuals in each of the 35 strains. For this simulation, we ignore mean parameters.

Model selection is a common way to choose between θ_G and θ_P (Oakley *et al.*, 2005). Using the effective degrees of freedom described in the previous section with the AIC criterion, we tabulate the number of times the maximum likelihood estimate correctly identifies the true generating scenario in

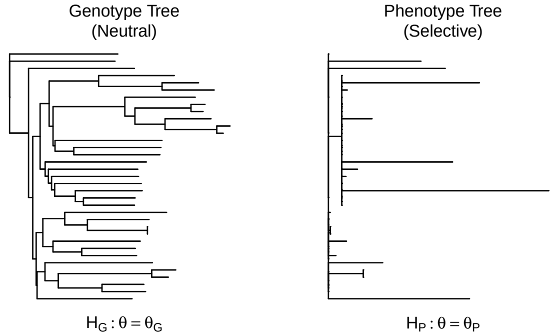

FIG 3. *Graphs for the genotypic tree (left) and the phenotypic tree (right) defining θ_G and θ_P respectively.*

Table 2(a). We observe that the maximum likelihood estimate tends to over fit the data (by estimating a simpler structure) so that we frequently select θ_P even as n increases.

While the regularized estimator does not choose one particular model, we recall that for an effectively large λ the estimate is θ_G and for $\lambda = 0$ it agrees with the maximum likelihood estimate. We tabulate the number of times the regularized estimate differs from the maximum likelihood estimate in Table 2(b). As expected, the regularized estimate is more conservative, allowing the selection of θ_G by default. In exchange, the regularized estimate appears to be less "powerful," more frequently selecting the default neutral model when the selective model is true.

We quantify the quality of the estimates in Table 2(c) with their mean squared errors. We see that regardless of model selected, the regularized estimator does no worse than the maximum likelihood estimate. Further, in cases where they disagree, the regularized estimates are closer to the true model. Combined with the implications of zero-length branches, the unreliability of model selection results strongly suggests that we rely on visual inspection of the tree estimates over the summary of a model selection criterion.

6. Case study: neutral inference of tail length phenotype in inbred mouse strains.

We illustrate neutral inference as a part of the analysis of phenotypes of inbred mouse strains. Inbred mouse strains are a laboratory science workhorse because they display wide, reproducible phenotypic diversity. However, the attribution of this divergence to their underlying genetic relationships is not fully understood. Using the mouse haplotype project genotype data (www.mousehapmap.org), we constructed a SNP-

TABLE 2

Model selection diagnostics and estimation error for maximum likelihood and regularized models. Correct decisions are highlighted.

(a) Model selection counts

		$n = 5 \times 35$		$n = 10 \times 35$	
		θ_G True	θ_P True	θ_G True	θ_P True
MLE	AIC Chooses θ_G	**14**	1034	**1**	749
	AIC Chooses θ_P	9986	**8966**	9999	**9251**

(b) Regularized estimate

		$n = 5 \times 35$		$n = 10 \times 35$	
		θ_G True	θ_P True	θ_G True	θ_P True
REG	Disagrees with MLE	**5587**	4660	**3881**	1572
	Agrees with MLE	4413	**5340**	6119	**8428**

(c) Mean Squared Error

		$n = 5 \times 35$		$n = 10 \times 35$	
		θ_G True	θ_P True	θ_G True	θ_P True
REG	Disagrees with MLE	0.001	0.014	0.002	0.013
	Agrees with MLE	0.021	0.019	0.020	0.017
MLE	REG disagrees	0.021	0.021	0.019	0.018
	REG agrees	0.021	0.019	0.020	0.017

MLE: maximum likelihood estimate, REG: regularized estimate, AIC: Akaike's Information Criterion.
Simulation standard errors are order 0.001

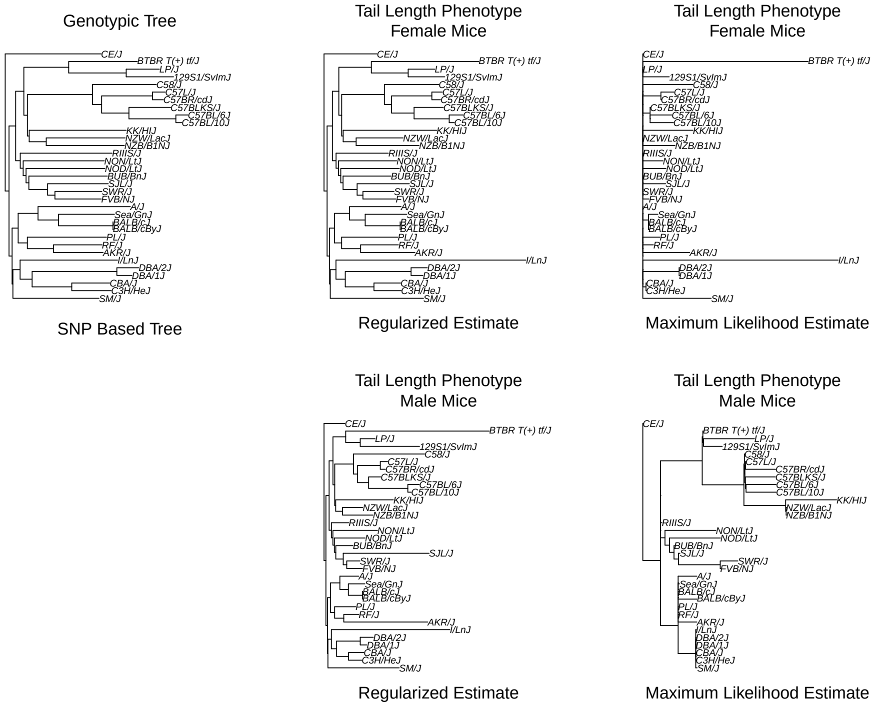

FIG 4. *Estimated trees for the tail length phenotype are expected to be predominately neutral (left). The maximum likelihood estimates (right column) and regularized estimates (center column) are given for female (top row) and male (bottom row) mice.*

based distance tree in the style of Frazer *et al.* (2007) and Kang *et al.* (2008) for $m = 33$ inbred strains. We matched these strains to morphological phenotypes measured on between $n = 15$ and $n = 20$ mice per strain in a study by Bret Payseur (University of Wisconsin-Madison) and Christopher Vinyard (Northeastern Ohio Universities Colleges of Medicine and Pharmacy). We consider the analysis of the tail length phenotype which is expected to confer no functional advantage in the laboratory and is therefore expected to be under neutral drift.

After separating mice into male and female groups, we plot the maximum likelihood and regularized tree estimates for tail length divergence in Fig. 4. Compared to the expected neutral tree on the left, we see some of the severity of the maximum likelihood estimates: many branches are exactly zero and most of the variation (edge length) is pooled into a few branches. The regularized estimates are a middle ground combining some of the extreme

TABLE 3
Variance models for tail length phenotypes.

Model	Heritable part	+	Error part	AIC, Male mice	AIC, Female Mice
Non-heritable	–		$\sigma^2 I_n$	547.3	552.5
Independent	$\sigma^2 I_m$	+	$\sigma^2 I_n$	237.4	171.9
Neutral	$\sigma^2 V \Theta_G V^T$	+	$\sigma^2 I_n$	224.1	161.0
Regularized	$\sigma^2 V \Theta_R V^T$	+	$\sigma^2 I_n$	219.2	157.2

AIC: Akaike's Information Criterion.

features of the maximum likelihood estimates and the structure of the neutral tree.

We compare the regularized estimate with estimates from alternative models for the covariance structure. The functional forms of the covariances and the corresponding AIC values are listed in table 3. In these AIC calculations, we use the effective degrees of freedom described in Section 4.1 to compute model complexity. In both male and female mice, the best two models (smallest AIC values) are the neutral and regularized estimates. That is, we are able to exclude non-heritable models where there is no random effect across strains, which matches the expectation that morphological traits have significant heritable variation. We are also able to exclude the model where there is no dependence between strains; in graph terms, this is a "star tree" topology consistent with exceptionally strong stabilizing selection. The regularized estimate has a slightly better criterion than the neutral model suggesting that the extra long branches for BTBR and I/J strains are statistically supported.

7. Discussion.

Our observations about maximum likelihood estimates are consistent with more general variance estimation as, besides their phylogenetic interpretation, trees are hierarchical representations of variance. While our problem and approach have prior tree structure available, the consideration given to the maximum likelihood penalty ought to generalize and we imagine that a differently penalized maximum likelihood estimate may lead to an adequate solution in more general cases.

The use of a prior genetic tree estimate requires both branch lengths and topology. We might relax the former by considering branches $t = I_{2m-2}$, a scenario where variation is proportional to the number of evolutionary splitting events; or t such that the tips of the previous tree are contemporary. Removing the topological assumptions remains an elusive problem.

We wish to emphasize that the plurality of possible evolutionary scenarios characterized by Brownian models makes visualizing a point estimate of the apparent variation particularly important. Model selection methods require a subset of hypotheses to compare and hypothesis testing only tells investigators about statistical significance. These models, however, are usually based on well defined selective regimes and are therefore more immediately interpretable.

In our current framework, we rely on our collaborators to interpret the difference between tree graphs. It would be ideal to formalize a metric comparing the evolutionary scenarios implied by two trees with which to quantify the selective force between them. Such a measure would naturally form the basis of a hypothesis test. To this end, likelihood based tests are unsatisfactory because our results emphasize that the fully parameterized likelihood is not sensitive to evolutionary differences under Brownian models. It is our feeling that tools like our regularized estimate will lead us closer to developing this measure.

APPENDIX

A.1. EM algorithm for maximum likelihood. The complete data log likelihood $f(Y,d)$ is based on $d = \Theta^{1/2}u \sim \mathcal{N}(0,\sigma^2\Theta)$ and $Y \mid d \sim \mathcal{N}(W\beta + ZVd, \sigma^2 I_n)$:

$$\begin{aligned}\log f(Y,d) = &-\frac{n+2m-2}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{k=1}^{2m-2}\log\theta_k \\ &- \frac{1}{2\sigma^2}\|Y - W\beta - ZVd\|^2 - \frac{1}{2\sigma^2}\sum_{k=1}^{2m-2}\frac{d_k^2}{\theta_k}.\end{aligned}$$

The E-step imputes the conditional mean of d:

$$\hat{d} = E(d|Y,\sigma^2,\theta,\beta) = \Theta^{(m)}V^TZ^T(ZV\Theta^{(m)}V^TZ^T + I_n)^{-1}(Y - W\beta^{(m)}).$$

The M-step maximizes expectation of the complete data log likelihood, $Q(\beta,\sigma^2,\theta;\hat{d}) = \log f(Y,\hat{d})$, and results in the following estimating equations:

$$\begin{aligned}\frac{\partial Q}{\partial\beta} &= -\frac{2}{\sigma^2}\left(W^T(Y - ZV\hat{d}) - W^TW\beta\right), \\ \frac{\partial Q}{\partial\sigma^2} &= -\frac{n+2m-2}{2}\frac{1}{\sigma^2} + \frac{1}{2}\left(\frac{1}{\sigma^2}\right)^2\left(\|Y - W\beta - ZV\hat{d}\|^2 + \sum_{k=1}^{2m-2}\frac{\hat{d}_k^2}{\theta_k}\right) \\ \frac{\partial Q}{\partial\theta_k} &= -\frac{1}{2}\frac{1}{\theta_k} + \frac{1}{2}\frac{1}{\theta_k}\frac{\hat{d}_k^2}{\sigma^2\theta_k}.\end{aligned}$$

The solutions for these estimating equations are straightforward:

$$\begin{aligned}\beta^{(m+1)} &= (W^T W)^{-1} W^T (Y - ZV\hat{d})\\ \sigma^{2(m+1)} &= n^{-1}\|Y - W\beta^{(m+1)} - ZV\hat{d}\|^2\\ \theta_k^{(m+1)} &= \frac{\hat{d}_k^2}{\sigma^{2(m+1)}}.\end{aligned}$$

A.2. EM algorithm for regularized estimation.

The regularized estimate M-step maximizes a penalized Q-function:

$$Q^*(\beta, \sigma^2, \theta; \hat{d}) = Q(\beta, \sigma^2, \theta; \hat{d}) + \lambda J(\theta).$$

For the penalty on $J(\theta)$, recall that $X_k = Zv_k$:

$$J(\theta) = \sum_{k=1}^{2m-2} \left\{ \left(\frac{1 + t_k X_k^T X_k}{1 + \theta_k X_k^T X_k} - 1 \right) + \log \left(\frac{1 + \theta_k X_k^T X_k}{1 + t_k X_k^T X_k} \right) \right\}.$$

The estimating equations only change for θ_k. The M-step updates for the penalized Q function are not closed form, but are based on the following adjustment.

$$\frac{\partial Q^*}{\partial \theta_k} = -\frac{1}{2}\frac{1}{\theta_k} + \frac{1}{2}\frac{1}{\theta_k}\frac{\hat{d}_k^2}{\sigma^2 \theta_k} - \lambda \frac{(1 + t_k X_k^T X_k)(X_k^T X_k)}{(1 + \theta_k X_k^T X_k)^2} + \frac{\lambda X_k^T X_k}{1 + \theta_k X_k^T X_k}.$$

ACKNOWLEDGEMENTS

The authors thank Bret Payseur and Chris Vinyard for their data and discussions. They are grateful to Doug Bates and Cécile Ané for a number of insightful discussions. The first author was supported by NIH grants T32 HL083806 and T15 LM007359.

REFERENCES

Bates, D. and Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes R package version 0.999375-35.

Butler, M. A. and King, A. A. (2004). Phylogenetic Comparative Analysis: A modeling approach for adaptive evolution. *The American Naturalist* **164** 683-695.

Corrada Bravo, H., Wright, S., Eng, K. H., Keleş, S. and Wahba, G. (2009). Estimating Tree Structured Covariance Matrices via Mixed-Integer Programming. In *Proceedings of the Twelfth International Conference on Artificial Intelligence* **5** 41-48.

Dempster, A., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39** 1-22.

Eng, K. H., Corrada Bravo, H. and Keleş, S. (2009). A Phylogenetic Mixture Model for Gene Expression Data. *Molecular Biology and Evolution* **26** 2363-2372.

Fay, J. C. and Wittkopp, P. J. (2008). Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* **100** 191-199.

Felsenstein, J. *Inferring phylogenies.*

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist* **125** 1-15.

Frazer, K., Eskin, E., Kang, H., Bogue, M., Hinds, D., Beilharz, E., Gupta, R., Montgomery, J., Morenzoni, M. and Nilsen, G. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448** 1050–1053.

Freckleton, R. P., Harvey, P. H. and Pagel, M. (2002). Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist* **160** 712-726.

Gu, X. (2004). Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics* **167** 531-542.

Guo, H., Weiss, R. E., Gu, X. and Suchard, M. (2007). Time squared: repeated measures on phylogenies. *Molecular Biology and Evolution* **24** 352-362.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer.

Hill, W. G. and Thompson, R. (1978). Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics* **34** 429-439.

Housworth, E. A., Martins, E. P. and Lynch, M. (2004). The Phylogenetic Mixed Model. *The American Naturalist* **163** 84-96.

Ives, A. R., Midford, P. E. and Garland, T. Jr. (2007). Within-Species Variation and Measurment Error in Phylogenetic Comparative Methods. *Systematic Biology* **56** 252-270.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178** 1709-1723.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2 ed. Springer.

Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45** 1065-1080.

Martins, E. P. and Hansen, T. F. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* **149** 646-667.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models.* Wiley.

Nuzhdin, S. V., Wayne, M. L., Harmon, K. and McIntyre, L. M. (2004). Common pattern of evolution of gene expression level and protein sequence in *Drosophila. Molecular Biology and Evolution* **21** 1308-1317.

Oakley, T. H., Gu, Z., Abouheif, E., Patel, N. H. and Li, W. H. (2005). Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Molecular Biology and Evolution* **22** 40-50.

Rifkin, S. A., Kim, J. and White, K. P. (2003). Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nature Genetics* **33** 138-144.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, 2 ed. Wiley.

van der Laan, M. J., Dudoit, S. and Keleş, S. (2004). Astymptotic Optimality of Likelihood-based Cross-vaidation. *Statistical Applications in Genetics and Molecular Biology* **3**.

Wahba, G., Johnson, D. R., Gao, F. and Gong, J. (1995). Adaptive Tuning of Numerical Weather Prediction Models: Randomized GCV in Three-and Four-Dimensional Data Assimilation. *Monthly Weather Review* **123** 3358-3369.

Whitehead, A. and Crawford, D. L. (2006). Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences* **103** 5425-5430.

Department of Biostatistics and Medical Informatics
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue
Madison, Wisconsin 53706
USA
E-mail: eng@stat.wisc.edu
E-mail: keles@biostat.wisc.edu