Statistica Sinica **19** (2009), 631-649

MARGINAL LIKELIHOOD FOR DISTANCE MATRICES

Peter McCullagh

University of Chicago

Abstract: A Wishart model is proposed for random distance matrices in which the components are correlated gamma random variables, all having the same degrees of freedom. The marginal likelihood is obtained in closed form. Its use is illustrated by multidimensional scaling, by rooted tree models for response covariances in social survey work, and unrooted trees for ancestral relationships in genetic applications.

Key words and phrases: Generalized Gaussian distribution, maximum-likelihood tree, multi-dimensional scaling, REML, residual maximum likelihood, tree-structured covariance matrix, unrooted tree, Wishart distribution.

1. Introduction

Distance matrices are widely used in genetic work to study the ancestral relationships among extant species or taxa. The emphasis in early work was on distance measures based on quantitative traits supposedly evolving by Brownian diffusion with occasional speciation splits (Cavalli-Sforza and Edwards (1967) and Felsenstein (1973)). In later work, distances were measured from aligned sequence data by counting the fraction of homologous sites at which each pair of species differs. In order to use this fraction to estimate the ancestral tree, a transformation is needed to correct for multiple and back-substitutions. In addition, it is necessary, or at least helpful for efficient estimation, to know the variances and covariances of these transformed proportions. These transformation and variance formulae are based on models that are specific to genetic evolution. For details, see Bulmer (1991).

Distance matrices, also called dissimilarity matrices, are widely used in archaeological work to discern relationships among artifacts. The aim is to understand trading patterns, the migration of populations, and the transfer of technology among early civilizations. The methods used to measure the distance between two artifacts based on expert assessment of stylistic elements or technological properties are necessarily subjective and ad hoc. By contrast with genetic applications, there is little theory to use as a guide for measuring distance or constructing statistical models. Such applications call for generic models and methods of estimation. For a range of examples and areas of application, see the 1970 volume edited by Hodson, Kendall and Tăutu (1970).

In both areas of application, relationships are expressed mathematically as an unrooted tree with leaves labelled by artifacts or taxa. The aim is to estimate this tree, including edge lengths. For example, Bulmer's genetic substitution model gives rise naturally to a criterion for estimation, which is weighted least squares since the model is specified only by means and covariances. Felsenstein's Brownian diffusion model is more detailed because it determines the joint distribution of the distances, not just their means and covariances. The diffusion model gives rise to two criteria, weighted least squares and maximum likelihood, which are not equivalent when applied to distance matrices. Thus all three criteria derived from two models are different.

The variances and covariances implied by the diffusion model are exactly quadratic in tree distances, whereas those derived from sequence substitution data are approximately linear (Gascuel (1997)). Exact linearity of covariances is seldom a reasonable assumption because it implies that the the *observed* distance matrix D is a tree (Sections 4 and 7). Thus the slightest departure of D from tree form contradicts exact linearity. For non-genetic applications, the generic Brownian diffusion model seems best because it is not degenerate in this sense. It is also scale equivariant.

In addition to specific models and areas of application, there is a parallel literature that emphasizes computational algorithms, neighbor joining (NJ) being the best known algorithm for the estimation of an unrooted tree from a distance matrix (Saitou and Nei (1987)). Generally speaking, a criterion such as maximum likelihood or weighted least squares has several local optima, each tree achieving roughly the same value of the criterion. By contrast, an algorithm such as NJ identifies a single tree, which may not be a stationary point of any statistically natural criterion. The relation between algorithms, criteria and statistical models is further complicated by the fact that many authors seem unaware of the distinctions. For a discussion of the relation between neighbor-joining and least squares, see Gascuel and Steel (2006).

The first goal of this paper is to obtain the likelihood function in closed form for Wishart distance matrices, in essence Felsenstein's Brownian diffusion model not restricted to trees. The second is to illustrate the application of rooted trees as a model for response covariances in social survey applications.

2. Wishart Distance Matrices

Let $S \sim \mathcal{W}_d(\Sigma)$ be a random symmetric matrix of order n having the Wishart distribution on d degrees of freedom with parameter $\Sigma = E(S)$ of rank n. If S is observed, the log likelihood function is

$$l(\Sigma; S) = -\tfrac{1}{2} d \operatorname{tr}(\Sigma^{-1} S) - \tfrac{1}{2} d \log \det(\Sigma)$$

for Σ in the space of positive definite symmetric matrices. This is an exponential-family model with canonical parameter $W = \Sigma^{-1}$, dispersion parameter $2d^{-1}$, cumulant function $-\log\det(\Sigma)$, and deviance function

$$-\log\det(\Sigma^{-1}S) - n + \mathrm{tr}(\Sigma^{-1}S).$$

If S has full rank, the deviance is minimized at the observation $\Sigma = S$. The parameterization used here matches the gamma parameterization for generalized linear models with quadratic covariance function

$$\mathrm{cov}(S_{ij}, S_{kl}) = \frac{(\Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk})}{d}.$$

Thus, Wishart-based generalized linear models can be constructed, and certain models used in the analysis of spatial data are linear on the mean-value scale. However, most of the models considered in this paper are not linear on any transformed scale.

In a number of applications the matrix S is not fully observed. Instead, only the distance matrix with components

$$D_{ij} = S_{ii} + S_{jj} - 2S_{ij} \tag{2.1}$$

is available. Note that if $S_{ij} = \langle y_i, y_j\rangle$ is the inner product of two vectors in $\mathcal{R}^d$, $D_{ij} = \langle y_i - y_j, y_i - y_j\rangle$ is the *squared* distance between the points. Consequently $D_{ii} = 0$, and the square roots satisfy the triangle inequality. We denote by $S \in \mathcal{PD}_n$ the set of positive definite symmetric matrices, and by $\mathcal{D}_n$ the image of $\mathcal{PD}_n$ under the linear transformation (2.1). Apart from the diagonal elements being zero, the characteristic property of a distance matrix $D \in \mathcal{D}_n$ is that D is negative definite on contrasts. A contrast is a linear combination α whose coefficients add to zero, and negative definiteness means

$$\alpha' D\alpha = -2\alpha' S\alpha \le 0$$

since S is positive definite. Some, but not all, elements of $\mathcal{D}$ satisfy the triangle inequality. In some contexts, such as multi-dimensional scaling, the terms similarity matrix and dissimilarity matrix are used (Hodson, Sneath and Doran (1966), Sattath and Tversky (1977) and Semple and Steel (2003, 2004))

The expected value of D is a matrix with components

$$\Delta_{ij} = E(D_{ij}) = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}.$$

The covariances are

$$\begin{aligned} d\ \mathrm{cov}(D_{ij}, D_{kl}) &= 2|\Sigma_{ik} + \Sigma_{jl} - \Sigma_{il} - \Sigma_{jk}|^2 \\ &= \frac{1}{2}|\Delta_{ik} + \Delta_{jl} - \Delta_{il} - \Delta_{jk}|^2. \end{aligned} \tag{2.2}$$

The latter expression has a natural geometric interpretation in the context of unrooted tree models (Section 4). Thus, even if the distinct components of S are uncorrelated, strong correlations are present among the components of D.

It turns out that the marginal Wishart model is also of the natural exponential-family type, although this is not entirely obvious. Because of the relation between quasi-likelihood and exponential family models (McCullagh and Nelder (1989)), the quasi-likelihood estimating function derived from (2.2) coincides with the log likelihood derivative as determined by the Wishart model. Thus maximum quasi-likelihood coincides with maximum likelihood provided that the maximum is unique. Unfortunately, most of the examples discussed here exhibit multiple local maxima. The likelihood function is needed to discriminate among competing local maxima, so quasi-likelihood alone is not satisfactory.

The likelihood function based on a statistic $T(Y)$ is usually obtained directly from the marginal density function $f(t;\theta)$ of T by computing density ratios $f(t;\theta)/f(t;\theta')$. In the present context, the density of S is available, and it is straightforward in principle to compute the joint moments or cumulants of D. No convenient expression is available for the density, so the likelihood function is not easily obtained in this way. The solution is to calculate the likelihood function indirectly without deriving the density function.

3. Gaussian Models

3.1. General

It is convenient here to introduce the distributional symbol $Y \sim N(\mathcal{K}, \mu, \Sigma)$ for a generalized Gaussian random vector in $\mathcal{R}^n$. The subspace $\mathcal{K} \subset \mathcal{R}^n$ is called the kernel. The meaning is that for any linear transformation such that $L\mathcal{K} = 0$, the linearly transformed vector LY is Gaussian $LY \sim N(L\mu, L\Sigma L')$ in the conventional sense. This implies that the matrix $L\Sigma L'$ is positive semi-definite. Here we assume strict positive definiteness in the sense that $\alpha'\Sigma\alpha > 0$ for non-zero $\alpha \in \mathcal{K}^0$, the space of linear functionals or contrasts that take the value zero on $\mathcal{K}$.

Two parameter values (μ_1, Σ_1) and (μ_2, Σ_2) are equivalent if $L(\mu_1 - \mu_2) = 0$ and $L(\Sigma_1 - \Sigma_2)L' = 0$. In other words, $\mu_1 - \mu_2 \in \mathcal{K}$ and $\Sigma_1 - \Sigma_2 \in \operatorname{sym}^2(\mathcal{K} \otimes \mathcal{R}^n)$, the space of matrices spanned by $xv' + vx'$ with $x \in \mathcal{K}$ and $v \in \mathcal{R}^n$. Equivalent parameter values determine the same distribution. If an identifiable parameterization is required, we can take μ to be a point (coset) in $\mathcal{R}^n/\mathcal{K}$, and similarly for Σ. Identifiable parameterizations are not especially helpful and we make little use of them.

Certain spatial covariance functions such as $-|x - x'|^\nu$ for $0 < \nu < 2$ are not positive definite in the ordinary sense, but are nonetheless positive definite on the space of simple contrasts (Stein (1999, Sec. 2.9)). The associated Gaussian process is also defined on contrasts by setting $\mathcal{K} = \mathbf{1}$, the space of constant functions. If observations are made at n points $x_1, \ldots, x_n$ in the plane, with Y_i observed at x_i, the distribution may be written in the form $Y \sim N(\mathbf{1}, 0, \Sigma)$ with $\Sigma_{ij} = -|x_i - x_j|^\nu$. The mean can be replaced by any vector $\mu \in \mathbf{1}$ without affecting the distribution. With additive treatment effects superimposed in the usual way, the distribution becomes $Y \sim N(\mathbf{1}, X\beta, \Sigma)$, where X is the model matrix. Once again, all points in the coset $X\beta + \mathbf{1}$ determine the same distribution, which means that the intercept is not identifiable. In practice, we would usually include a white-noise component with $\Sigma_{ij} = \sigma_0^2\delta_{ij} - \sigma_1^2|x - x'|^\nu$, so that there are two variance components to be estimated. For specific examples, see McCullagh and Clifford (2005).

In general, the matrix Σ in $N(\mathcal{K}, \mu, \Sigma)$ is not positive definite. However, if $x \in \mathcal{K}$ and $v \in \mathcal{R}^n$ we may add to Σ any matrix of the form $xv' + vx'$ without affecting the value of $L\Sigma L'$, and thus without affecting the distribution of contrasts. All such versions are equivalent. If the columns of the matrix K are vectors in $\mathcal{K}$, we may add to Σ a suitably large multiple of KK' so that $\Sigma + KK'$ is positive definite. There is no loss of generality in assuming Σ to be positive definite, so we write $W = \Sigma^{-1}$ for the inverse.

The log likelihood for (β, Σ) can be obtained by choosing a full-rank linear transformation LY such that $\ker(L) = \mathcal{K}$, and using the conventional expression for the density. The quadratic form in the exponent is

$$(y - X\beta)'L'(L\Sigma L')^{-1}L(y - X\beta).$$

The matrix $Q = \Sigma L'(L\Sigma L')^{-1}L$ is a projection with kernel $\mathcal{K}$, and self-adjoint with respect to the inner product $\langle u, v\rangle = u'Wv$, i.e. $\langle u, Qv\rangle = \langle Qu, v\rangle$. It follows that Q is the unique orthogonal projection with kernel $\mathcal{K}$. Specifically, if the columns of K form a basis for $\mathcal{K}$, $Q = I - K(K'WK)^{-1}K'W$, and the matrix of the quadratic form is WQ. The determinantal factor in the likelihood is the square root of

$$|L\Sigma L'|^{-1} = |(L\Sigma L')^{-1}| = \frac{\operatorname{Det}(L'(L\Sigma L')^{-1}L)}{|LL'|} = \frac{\operatorname{Det}(WQ)}{|LL'|},$$

where Det() is the product of the non-zero eigenvalues. Thus, the log likelihood function for the parameter $\theta = (\beta, \Sigma)$ in the model $Y \sim N(\mathcal{K}, X\beta, \Sigma)$ is

$$l(\theta; y, \mathcal{K}) = \tfrac{1}{2}\log\operatorname{Det}(WQ) - \tfrac{1}{2}(y - X\beta)'WQ(y - X\beta). \tag{3.1}$$

This function is constant on equivalence classes in the parameter space, which means that equivalent versions of (β, Σ) give the same likelihood.

The particular case of most interest here is the one in which the kernel is $\mathcal{K} = \mathcal{X} = \mathrm{span}(X)$. Then β is eliminated from the likelihood and (3.1) reduces to the residual likelihood

$$l(\theta; y, \mathcal{K}) = \tfrac{1}{2}\log \mathrm{Det}(WQ) - \tfrac{1}{2}y'WQy = \tfrac{1}{2}\log \mathrm{Det}(WQ) - \tfrac{1}{2}\,\mathrm{tr}(WQS)$$

(Patterson and Thompson (1971), Harville (1974, 1977) and Stein (1999, Sec.6.4)). At this point it is convenient to introduce the symbol $S \sim \mathcal{W}_1(\mathcal{K}, \Sigma)$ for the generalized Wishart distribution of the random matrix $S = YY'$ when $Y \sim N(\mathcal{K}, 0, \Sigma)$. From the preceding calculations we observe that S is sufficient, and the log likelihood function is $l(\theta; S, \mathcal{K}) = \log \mathrm{Det}(WQ)/2 - \mathrm{tr}(WQS)/2$.

The relevance of this calculation to distance matrices is as follows. The mapping (2.1) is a linear transformation on symmetric matrices. It has a kernel equal to the space sym^+ of additive symmetric matrices, i.e. matrices of the form $A_{ij} = v_i + v_j$ with $v \in \mathcal{R}^n$, which is the same as $\mathrm{sym}^2(\mathbf{1} \otimes \mathcal{R}^n)$. Thus, if S has the ordinary Wishart distribution $\mathcal{W}_1(\Sigma)$ with rank one, the distance matrix D is distributed as

$$-D \sim \mathcal{W}_1(\mathbf{1}, 2\Sigma) = \mathcal{W}_1(\mathbf{1}, -\Delta),$$

which we denote by $D \sim \mathcal{W}_1^-(\mathbf{1}, \Delta)$ with kernel $\mathbf{1} \subset \mathcal{R}^n$. Consequently the log likelihood function based on D is

$$l(\Delta; D) = \tfrac{1}{2}\log \mathrm{Det}(WQ) + \tfrac{1}{4}\,\mathrm{tr}(WQD), \tag{3.2}$$

with $K = \mathbf{1}$ and $Q = I - K(K'WK)^{-1}K'W$. The log likelihood is expressed as a function of $W = \Sigma^{-1}$, but it is constant on equivalence classes, so it depends only on Δ.

Tunnicliffe-Wilson (1989) and Cruddas, Reid and Cox (1989) have taken the REML argument a step farther by ignoring scalar multiples of Y in addition to translation by $\mathcal{K}$. The marginal log likelihood based on the standardized residual $QY/\|QY\|$ is

$$\begin{aligned}\check{l}(\Sigma; y, \mathcal{K}) &= -\tfrac{n-p}{2}\log(y'\Sigma^{-1}Qy) - \tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}\log|K'\Sigma^{-1}K| + \tfrac{1}{2}\log|K'K'| \\ &= -\tfrac{n-p}{2}\log \mathrm{tr}(WQS) + \tfrac{1}{2}\log \mathrm{Det}(WQ), \end{aligned} \tag{3.3}$$

which is constant on scalar multiples of Σ. Thus, if only relative distances are available, the log likelihood is

$$\tfrac{n-p}{4}\log \mathrm{tr}(WQD) + \tfrac{1}{2}\log \mathrm{Det}(WQ),$$

where $p = \dim(\mathcal{K})$ and $n - p = \mathrm{rank}(Q)$. Note that S necessarily has rank one, so $\mathrm{tr}(WQS) = \mathrm{Det}(WQS)$.

3.2. Gaussian matrix model

In order to deal with distance matrices of rank $d > 1$ we proceed as follows. Let Y be a Gaussian random matrix of order $n \times d$ with moments

$$E(Y) = X\beta, \quad \operatorname{cov}(Y_{ir}, Y_{js}) = \Sigma_{ij}\Gamma_{rs}.$$

The columns of the model matrix X span a subspace $\mathcal{X} \subset \mathcal{R}^n$ of dimension p, and β is a matrix of order $p \times d$. The conventional way of writing this model is $Y \sim N(X\beta, \Sigma \otimes \Gamma)$, with unknown parameters β, Σ, Γ to be estimated. For present purposes it is replaced by the generalized Gaussian distribution $N(\mathcal{X}^{\oplus d}, 0, \Sigma \otimes \Gamma)$ with kernel $\mathcal{K} = \mathcal{X}^{\oplus d}$ of dimension pd in $\mathcal{R}^{nd}$. If $LX = 0$, then $LY \sim N(0, (L\Sigma L') \otimes \Gamma)$ in the conventional sense. The argument used in the preceding section gives the log likelihood for (Σ, Γ) in the form

$$\begin{aligned} l(\Sigma, \Gamma; y, \mathcal{X}) &= \tfrac{1}{2} \log \operatorname{Det}(WQ \otimes \Gamma^{-1}) - \tfrac{1}{2} \operatorname{tr}(y'WQy\Gamma^{-1}) \\ &= \tfrac{1}{2} \operatorname{rank}(\Gamma) \log \operatorname{Det}(WQ) - \tfrac{n-p}{2} \log \operatorname{Det}(\Gamma) - \tfrac{1}{2} \operatorname{tr}(y'WQy\Gamma^{-1}), \end{aligned}$$

where $Q = I - X(X'WX)^{-1}X'W$ is of order n and rank $n - p$.

If $\Gamma_{rs} = \delta_{rs}$ is known, then each column of Y is an independent replicate, and the log likelihood reduces to

$$\tfrac{d}{2} \log \operatorname{Det}(WQ) - \tfrac{d}{2} \operatorname{tr}(WQS) = \tfrac{d}{2} \log \operatorname{Det}(WQ) + \tfrac{d}{4} \operatorname{tr}(WQD), \tag{3.4}$$

where $S = YY'/d$. For $\mathcal{X} = \mathbf{1}$, this is the log likelihood for the Wishart model $D \sim \mathcal{W}_d^-(\mathbf{1}, \Delta)$ on which all calculations in Section 5 are based. The degrees of freedom enters only as a multiplicative factor, so it is irrelevant for most purposes whether d is known or not. If $\Gamma = \gamma\delta_{rs}$, the maximum-likelihood estimator of γ for fixed Σ is $\operatorname{tr}(WQS)/(n - p)$, and the profile log likelihood is

$$\tfrac{d}{2} \log \operatorname{Det}(WQ) - \tfrac{d(n-p)}{2} \log \operatorname{tr}(WQS).$$

This function is constant on scalar multiples of Σ, a generalization of (3.3) to matrices S of rank $d \geq 1$.

4. Trees Rooted and Unrooted

A non-negative symmetric matrix Σ of order n is called a rooted $[n]$-tree if

$$\Sigma_{ij} \geq \min(\Sigma_{ik}, \Sigma_{jk}) \tag{4.1}$$

for all i, j, k in $[n] = \{1, \ldots, n\}$. The tree inequality (4.1) is the condition that permits a non-negative symmetric matrix to be represented graphically as a rooted tree. The value Σ_{ij} is the distance from the root to the junction at which the leaves i, j occur on separate branches. For a diagram and further explanation, see Felsenstein (2004, p.395). No distinction is drawn here between the matrix representation (Table 1b) and the graphical representation (Figure 1) of the same tree.

As a graph, each edge in the tree is labelled naturally by the set of leaves that occur as terminal nodes on that branch. Thus, the root edge is labelled $[n] = \{1, \ldots, n\}$, each leaf edge is a singleton $\{i\}$, and each leaf node is an element. Every tree has a unique canonical decomposition

$$\Sigma = \sum_r \lambda_r b_r b_r' = B\Lambda B'$$

in which b_r is the indicator vector for edge r, Λ is diagonal and $\lambda_r > 0$ is the edge length. The associated Boolean tree (or topological type) BB' is obtained by replacing each edge length by one. In a binary or bifurcating tree, each non-leaf edge splits into exactly two branches, so the number of edges is $2n - 1$. A non-binary tree has fewer edges.

The canonical decomposition has a natural variance-components interpretation in which the response Y_i is measured at leaf i. Given the Boolean tree BB', we associate with each edge r an independent random variable η_r with variance λ_r. Then the sum $Y = B\eta$ has covariance matrix Σ. In other words, $\mathrm{var}(Y_i)$ is the sum of the variances of the ηs on all edges from the root to the leaf, and $\mathrm{cov}(Y_i, Y_j)$ is the sum of the variances on all branches that include both leaves.

To each rooted $[n]$-tree Σ there corresponds an unrooted $[n]$-tree defined by

$$\Delta_{ij} = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}$$

which is the distance between the two leaf nodes. More directly, a non-negative symmetric matrix Δ is an unrooted tree if $\Delta_{ii} = 0$ and

$$\Delta_{ij} + \Delta_{kl} \le \max\{\Delta_{ik} + \Delta_{jl},\ \Delta_{il} + \Delta_{jk}\} \tag{4.2}$$

for all i, j, k, l not necessarily distinct. This is called Buneman's four-point metric condition after Buneman (1971). For further details, see Semple and Steel (2003, Chap. 7). The set of rooted trees is a subset of $\mathcal{PD}_n$, and the unrooted trees are a subset of $\mathcal{D}_n$.

Each edge r of an unrooted tree splits the leaves into two non-empty blocks, those on the left and those on the right. Denote the split by b_r in matrix form, i.e. $b_r(i, j) = 1$ if leaves i, j occur in the same block, or the same side of edge r. Every unrooted tree has a canonical decomposition of the form

$$\Delta_{ij} = \sum_r \lambda_r \bar{b}_r(i, j),$$

where $\lambda_r > 0$ is the edge length, and $\bar{b}_r$ is the Boolean complement of b_r. Thus $\bar{b}_r(i,j) = 1$ if the path from i to j includes edge r. The associated Boolean tree $\sum \bar{b}_r$ has edges of unit length. The number of edges is at most $2n-3$.

To each pair of terminal nodes (i,j) there corresponds a directed path of length Δ_{ij} from i to j. To each pair of paths (i,j) and (k,l) there corresponds an intersection whose signed length is

$$\Delta_{ij,kl} = \tfrac{1}{2}(\Delta_{il} - \Delta_{lj} + \Delta_{jk} - \Delta_{ki}).$$

The sign is positive if the intersection is traversed in the same direction on each path, otherwise negative or zero. If $\{\eta_r\}$ are independent non-negative random variables with mean and variance λ_r, the sum $\tilde{D}_{ij} = \sum \eta_r \bar{b}_r(i,j)$ is a random matrix whose mean is Δ. Furthermore, $\tilde{D}$ is an unrooted tree, and if the ηs are strictly positive $\tilde{D}$ has the same topology as Δ. The covariance matrix of order $n^2 \times n^2$ is $\mathrm{cov}(\tilde{D}_{ij}, \tilde{D}_{kl}) = |\Delta_{ij,kl}|$. It follows that the $n^2 \times n^2$ path intersection matrix $|\Delta_{ij,kl}|$ has rank equal to the number of edges in Δ, i.e. at most $2n-3$. By contrast, the upper trianglular components in the Wishart matrix $D \sim \mathcal{W}_d^-(\mathbf{1}, \Delta)$ are linearly independent even for $d=1$. Thus the matrix of squared path intersection lengths $\Delta_{ij,kl}^2$ in (2.2) has rank $n(n-1)/2$.

If Σ is constant on the diagonal all leaves are equi-distant from the root, and the tree is called spherical. The associated unrooted tree contains a central point such that all leaves are equi-distant from the centre. Spherical trees arise in genetic models under the assumption that mutations occur at the same rate on all lineages, so spherical trees are called ‘clock-like’. Spherical unrooted trees satisfy the ultrametric inequality $\Delta_{ij} \le \max(\Delta_{ik}, \Delta_{jk})$. The coalescent model (Kingman (1982a,b)) is a probability distribution on unrooted spherical trees.

5. Single-Matrix Applications

5.1. Multi-dimensional scaling

Gower (1966) considered the problem of recovering the Euclidean configuration of a set of n points from the matrix D of observed squared distances. We can express this as a formal model $D \sim W^-(\mathbf{1}, \Delta)$ with

$$\Delta = \Delta_0 - M - \sigma_0^2 I_n,$$

where $\Delta_0 \in \mathrm{sym}^+$ belongs to the kernel, M is positive definite of rank 2, and $\sigma_0^2 I_n$ represents departures from the target two-dimensional configuration. Gower assumed that the eigenvalues beyond the first two were small, but he did not assume that they were equal, nor did he use a formal model for the joint distribution of the distances. His solution was to transform the squared distances to inner product form

$$S_{ij} = -\frac{D_{ij} - \bar{D}_{i\boldsymbol{\cdot}} - \bar{D}_{\boldsymbol{\cdot}j} + \bar{D}}{2},$$

preserving distances but eliminating Δ_0. In matrix form $S = -QDQ/2$ where $Q_{ij} = \delta_{ij} - 1/n$ is the exchangeable projection with kernel $\mathbf{1}$. In distributional form, $S \sim \mathcal{W}(\mathbf{1}, \Sigma)$ with $\Sigma = M' + \sigma_0^2 Q/2$ where M' is symmetric with rank 2. The matrix M' was estimated by choosing the best rank 2 least-squares approximation to S. The fitted two-dimensional configuration is the set of n points (ξ_{1i}, ξ_{2i}) where the eigenvectors are scaled so that $\|\xi_r\|^2 = \lambda_r$. The configuration is unique up to planar rotations and reflections. This least-squares projection coincides with the maximum-likelihood solution in the Wishart model.

5.2. Rooted trees for correlated responses

Although the models of this paper are designed for covariance matrices and distance matrices, the following example taken from Ehrenberg (1981) shows that the techniques may be used to good effect on correlation matrices. In the course of a questionnaire for U.K. television viewers, adults were asked whether the 'really liked to watch' each of ten programmes, four broadcast by ITV and six by the BBC. Table 1a shows the sample correlation matrix of the ten responses, reordered so that the first five are sports programmes, and the last five are news and current affairs.

A function for fitting rooted and unrooted trees was written in R. It takes the canonical decomposition generated by an initial tree and uses Newton-Raphson to compute the coefficients by maximum likelihood in the Wishart model $\mathcal{W}_d(\mathcal{K}, \Sigma)$. If at any iteration some of these coefficients are zero, the procedure moves to an equivalent binary tree of an adjacent, randomly chosen, topological type until a local maximum is found. In general, the likelihood function has several local maxima, but the number of local maxima is much less than the number of topological types. All 19 fitted coefficients are positive, so the fitted matrix is at least a local maximum of the likelihood function. The analysis was actually performed on the correlation matrix in Table 3 of Ehrenberg (1981), with ITV programmes followed by BBC programmes in alphabetical order. The output is shown in Table 1b. The programmes have been permuted for visual effect, so that the structure can more easily be seen from the fitted matrix. The main partition is a contrast of the first five programmes with the remainder, which happens to be the contrast between sports programmes and current affairs. The same contrast and the resulting simplification were also noted by Ehrenberg, who used this permuted matrix to argue that tables are superior to graphs for conveying quantitative information. Within sports programmes, the main contrast is

Table 1a. Viewer preference correlation matrix for 10 programmes

WoS	1.000	0.581	0.622	0.505	0.296	0.140	0.187	0.145	0.093	0.078
MoD	0.581	1.000	0.593	0.473	0.326	0.121	0.131	0.082	0.039	0.049
GrS	0.622	0.593	1.000	0.474	0.341	0.142	0.181	0.132	0.070	0.085
PrB	0.505	0.473	0.474	1.000	0.309	0.124	0.168	0.106	0.065	0.092
RgS	0.296	0.327	0.341	0.309	1.000	0.121	0.147	0.064	0.051	0.097
24H	0.140	0.122	0.142	0.124	0.121	1.000	0.524	0.395	0.243	0.266
Pan	0.187	0.131	0.181	0.168	0.147	0.524	1.000	0.352	0.200	0.197
ThW	0.145	0.082	0.132	0.106	0.064	0.395	0.352	1.000	0.270	0.188
ToD	0.093	0.039	0.070	0.065	0.051	0.243	0.200	0.270	1.000	0.155
LnU	0.078	0.049	0.085	0.092	0.097	0.266	0.197	0.188	0.155	1.000

Table 1b. Fitted tree for viewer preference correlation matrix

WoS	0.99	0.59	0.61	0.48	0.32	0.10	0.10	0.10	0.10	0.10
MoD	0.59	1.01	0.59	0.48	0.32	0.10	0.10	0.10	0.10	0.10
GrS	0.61	0.59	0.99	0.48	0.32	0.10	0.10	0.10	0.10	0.10
PrB	0.48	0.48	0.48	1.00	0.32	0.10	0.10	0.10	0.10	0.10
RgS	0.32	0.32	0.32	0.32	1.00	0.10	0.10	0.10	0.10	0.10
24H	0.10	0.10	0.10	0.10	0.10	0.96	0.51	0.36	0.25	0.20
Pan	0.10	0.10	0.10	0.10	0.10	0.51	1.01	0.36	0.25	0.20
ThW	0.10	0.10	0.10	0.10	0.10	0.36	0.36	0.99	0.25	0.20
ToD	0.10	0.10	0.10	0.10	0.10	0.25	0.25	0.25	1.03	0.20
LnU	0.10	0.10	0.10	0.10	0.10	0.20	0.20	0.20	0.20	1.01

between Rugby Special and the others, which are soccer and professional boxing. The fitted correlation matrix is also illustrated by a conventional tree diagram in Figure 1.

One objection to the preceding analysis is that the model is geared for covariance matrices rather than correlation matrices. Although the fitted matrix is positive definite, it is not a correlation matrix because the diagonal entries are not exactly one. This objection can be partially answered by using the reduced model consisting of spherical trees, constant on the diagonal. The fitted matrix is then a multiple of a correlation matrix. In this instance, the multiple is 0.9990 and the fitted matrix differs only slightly from Table 1b. The deviance for the reduced model is 0.053.

It may appear that the matrix of fitted covariances is not a good approximation to the observed covariances. However, the deviance is only 0.051, distributed approximately as $d^{-1}\chi^2_{36}$ where d is the sample size or degrees of freedom, which is not reported. To see whether this value is large, a small-scale simulation was run with Gaussian data generated from the distribution with covariance matrix

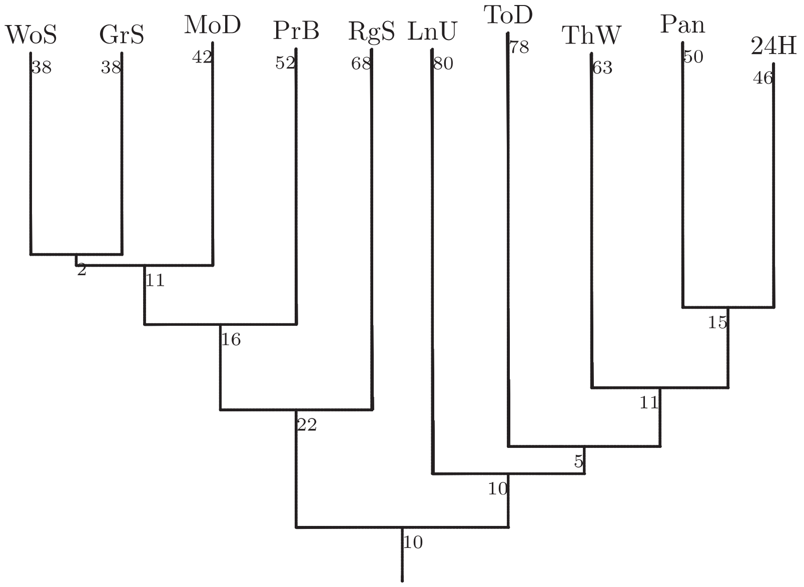

Figure 1. Rooted tree illustrating an approximate correlation matrix. For each pair of variables, the correlation in percent is the distance from the root to the junction at which the pair split.

in Table 1b. For each simulation, the fitted tree was obtained, and the deviance computed. For $d = 100$, the null distribution of deviances is roughly $1.06\chi^2_{36}$, only slightly larger than the nominal χ^2 on $(n-1)(n-2)/2$ degrees of freedom. The nominal approximation is even better for $d = 1,000$. Thus, if the sample size is 1,500 or less, the discrepancy between the observed covariance matrix and the fitted tree is compatible with the tree model.

This tree model is a standard tool in genetics where the supporting argument is based on evolutionary processes. It is remarkable that it works so well in this application where no comparable supporting argument is available.

5.3. Unrooted trees for genetic distances

Table 2a shows the genetic distances between seven species, dog, bear, raccoon, weasel, seal, sea lion, cat and monkey, as given by Sarich (1969) and reported by Felsenstein (2004) to illustrate tree clustering algorithms. For distance matrices of this sort, the normal practice in the genetics literature is to fit an unrooted tree by least squares as if the distinct components of D were independent. Weights are sometimes used, and if these are based on the current estimate of Δ, the procedure is equivalent to quasi-likelihood. In the Wishart model $D \sim W_d^-(\mathbf{1}, \Delta)$, each component has a gamma distribution with variance proportional to the square of its expected value. Furthermore, the covariance of

Table 2a. Immunological distances for eight species

	Dog	Bear	Raccn	Weasel	Seal	S-lion	Cat	Monkey
Dog	0	32	48	51	50	48	98	148
Bear	32	0	26	34	29	33	84	136
Raccoon	48	26	0	42	44	44	92	152
Weasel	51	34	42	0	44	38	86	142
Seal	50	29	44	44	0	24	89	142
Sea lion	48	33	44	38	24	0	90	142
Cat	98	84	92	86	89	90	0	148
Monkey	148	136	152	142	142	142	148	0

Table 2b. Unrooted fitted tree for eight species

	Dog	Bear	Raccn	Weasel	Seal	S-lion	Cat	Monkey
Dog	0.00	32.00	45.57	51.95	50.28	50.14	100.83	154.85
Bear	32.00	0.00	26.62	33.00	31.33	31.19	81.88	135.90
Raccoon	45.57	26.62	0.00	44.41	42.73	42.59	93.29	147.31
Weasel	51.95	33.00	44.41	0.00	40.83	40.69	86.63	140.65
Seal	50.28	31.33	42.73	40.83	0.00	24.00	89.71	143.73
Sea lion	50.14	31.19	42.59	40.69	24.00	0.00	89.57	143.59
Cat	100.83	81.88	93.29	86.63	89.71	89.57	0.00	148.00
Monkey	154.85	135.90	147.31	140.65	143.73	143.59	148.00	0.00

Table 2c. Spherical unrooted tree for eight species

	Dog	Bear	Raccn	Weasel	Seal	S-lion	Cat	Monkey
Dog	0.00	45.21	45.21	45.21	45.21	45.21	90.26	145.50
Bear	45.21	0.00	27.40	39.12	38.35	38.35	90.26	145.50
Raccoon	45.21	27.40	0.00	39.12	38.35	38.35	90.26	145.50
Weasel	45.21	39.12	39.12	0.00	39.12	39.12	90.26	145.50
Seal	45.21	38.35	38.35	39.12	0.00	23.58	90.26	145.50
Sea lion	45.21	38.35	38.35	39.12	23.58	0.00	90.26	145.50
Cat	90.26	90.26	90.26	90.26	90.26	90.26	0.00	145.50
Monkey	145.50	145.50	145.50	145.50	145.50	145.50	145.50	0.00

two components is proportional to the square of their path intersection length. This is in close accord with evolutionary notions, such as Brownian diffusion for quantitative traits, provided that squared differences are used.

Table 2b gives the unrooted tree, and Table 2c gives the unrooted spherical tree, both fitted by maximum likelihood using software described in the preceding section. The graphical representations in Figure 2a,b shows the edge lengths but emphasizes the topology. Figure 11.8 of Felsenstein (2004), obtained by the NJ algorithm, is similar to Figure 2a, but the topologies are different. The spherical tree in Figure 2b is different from the UPGMA tree in Figure 11.6 of Felsenstein (2004), but the topologies are the same.

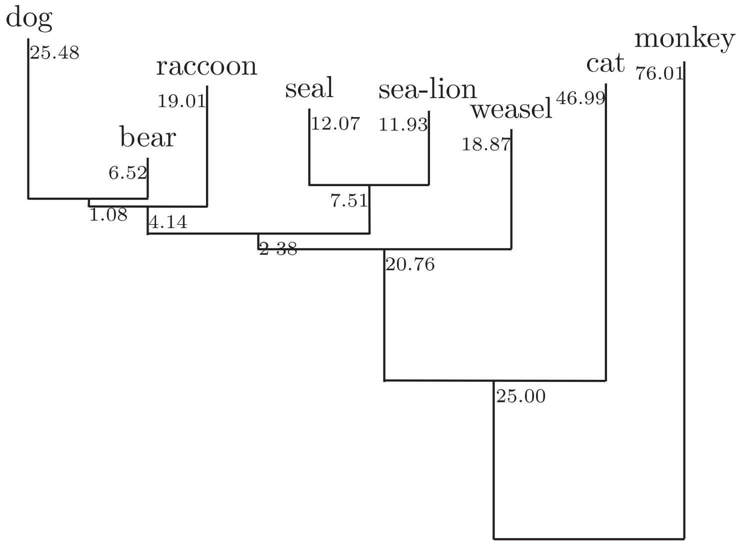

Figure 2a. The tree in Table 2b with deviance 0.0584. There is no root or central point, but the monkey branch has been arbitrarily split into two vertical parts for aesthetic reasons.

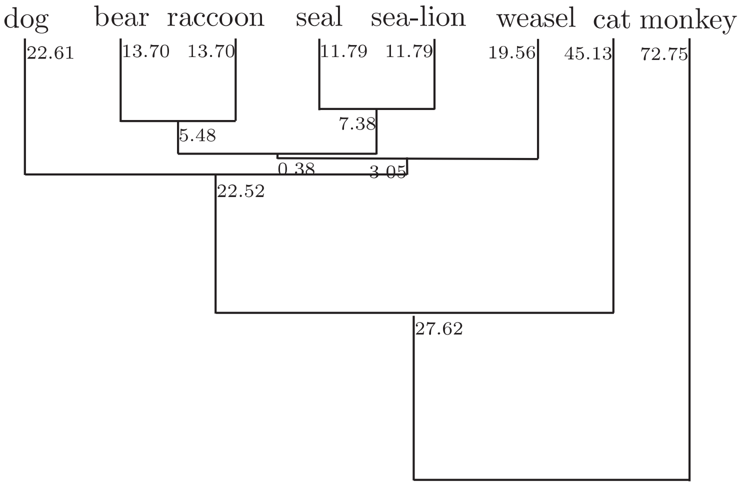

Figure 2b. Spherical tree in Table 2c with deviance 0.2090. Each sub-tree has a central point at the mid-point of the longest path.

The residual deviances for the two trees are 0.0584 and 0.2090, respectively. Neither the set of trees nor the subset of spherical trees is a manifold. However, if the tree is binary with all edge lengths positive, each sufficiently small neighbourhood is a manifold of dimension $f = 2n - 3$, or $f = n - 1$ for spherical trees.

Standard asymptotic theory for large d or small dispersion implies that residual deviances are distributed as χ^2/d on $n(n-1)/2 - f$ degrees of freedom. The hypothesis that the tree is spherical may be tested by comparing the reduction in deviance with the residual deviance by the ratio

$$F = \frac{(0.2090 - 0.0584)/6}{0.0584/15} = 6.44.$$

The standard $F_{6,15}$ approximation indicates that the spherical model is not consistent with the data. Although the limit distribution is correct for fixed n as $d \to \infty$, the approximation is suspect because some of the fitted edge lengths are small and the two fitted trees do not have the same topology. Nevertheless, the F approximation is in reasonably good agreement with simulation.

Both likelihood functions have further local maxima that are only slightly less than the values reported. For example, the log likelihood for spherical trees has a local maximum with deviance 0.2531 at a tree of the topological type shown in Figure 2a. The log likelihood for general trees has local maxima on trees of several different shapes. The NJ-tree shown in Figure 11.8 of Felsenstein (2004) has a deviance of 0.0607, and the likelihood has a local maximum with deviance 0.0587 on trees of the same shape. However, the likelihood does not appear to have a local maximum on non-spherical trees of the shape illustrated in Figure 2b.

6. Generalized Linear Models

6.1. Link functions and power transformation

Although linear models are not especially useful for distance matrices, it is possible to introduce a link function acting component-wise in such a way that $D \sim \mathcal{W}^-(\mathbf{1}, \Delta)$ is the sampling distribution, $\Delta = E(D)$ is the mean-value parameter and $g(\Delta)$ is a tree. Such transformations are potentially useful if the distance measure is non-linearly related to the genetic distance. Power transformations are the most natural in this context. When the power model is used for Sarich's data in Section 5.3, the residual deviance is reduced by about 8%, from 0.0584 at $\lambda = 1$ to 0.0539 at $\lambda = 1.4$. This is certainly not a significant decrease, so there is no evidence that the model is improved by transformation.

An alternative approach following the lines of Box and Cox (1964) is to apply the power transformation directly to the matrix D, so the model is $D^\lambda \sim \mathcal{W}^-(\mathbf{1}, \Delta)$, where $\Delta = E(D^\lambda)$ is a tree. If the density $h(D; \Delta)$ of the distribution $\mathcal{W}^-(\mathbf{1}, \Delta)$ were available, the likelihood function $h(D^\lambda; \Delta) \prod_{i<j} (\lambda D_{ij}^{\lambda-1})$ for (λ, Δ) could be used for inference. Unfortunately, this density is not available, and the calculations in Section 3 are not sufficient to determine the likelihood function for this transformation model.

6.2. Comparison of Wishart distance matrices

Suppose that $D_1 \sim \mathcal{W}_d^-(\mathbf{1}, \Delta_1)$ and $D_2 \sim \mathcal{W}_d^-(\mathbf{1}, \Delta_2)$ are two independent distance matrices indexed by the same set of objects. For example these might be genetic distance matrices for the same set of species, but determined by two different methods or distinct traits. Alternatively, for a given set of landmarks, D_1 and D_2 might be distances measured on two images of the same or similar object. It is natural to ask whether the distance matrices are homogeneous, or similar, and if so to combine them. More generally, there might be k matrices on $d_1, \ldots, d_k$ degrees of freedom, all supposedly independent and measuring the same configuration.

This structure suggests a number of simple models of the generalized linear type for k matrices as follows:

$$\Delta_i = \begin{cases} \Delta & \text{(homogeneous configurations)} \\ \Delta \exp(\beta_i) & \text{(similar configurations)} \\ \Delta_i & \text{(general),} \end{cases} \tag{6.1}$$

where $\Delta \in \mathcal{D}$, and $\beta_1, \ldots, \beta_k$ are scalars. The intermediate model implies that the matrices $\Delta_1, \ldots, \Delta_k$ are proportional to Δ. In genetic applications, it is natural to replace Δ by Δ^γ, for some positive scalar γ, and impose the condition that Δ be an unrooted tree.

The log likelihood function is

$$\tfrac{1}{2} \sum_i d_i \log \mathrm{Det}(W_i Q_i) + \tfrac{1}{4} \sum_i d_i \operatorname{tr}(W_i Q_i D_i),$$

where $W_i^{-1} = \text{const} - \Delta_i/2$, and Q_i is the associated orthogonal projection with kernel $\mathbf{1}$. The likelihood ratio statistic may be used for model comparisons in the usual way provided that the degrees of freedom are either known or equal.

7. Discussion

The likelihood function (3.4) is most easily computed using standard matrix operations including eigenvalue decompositions. The derivative vector and the Fisher information matrix, both of which are needed for the Newton-Raphson algorithm, are also easily evaluated using standard matrix operations. For the important special case in which $\mathcal{X} = \mathbf{1}$ and Σ is a tree, Felsenstein's (1973) pruning algorithm may be used to compute the likelihood. This sequential algorithm exploits the tree structure to generate implicitly the spectral decomposition of the matrix WQ.

The distinction between rooted and unrooted trees is not always obvious. The social-science example in Section 5.2 uses rooted trees as a model for structured covariance matrices. Most of the discussion in the genetics literature concerns rooted trees, and most diagrams show a root, but virtually all of the fitted models are unrooted (Felsenstein (2004, p.256)). For example, the Brownian diffusion model for the value of a phenotype at the terminal leaves is formulated initially as the standard Gaussian model $Y \sim N(\mathbf{1}\mu, \Sigma)$, with scalar μ, and Σ in the space of rooted trees. This model is identifiable but the parameters are not estimable, in part because the space of rooted trees includes the space $\mathbf{1} \otimes \mathbf{1}$. By reduction to contrasts, Felsenstein (1973) replaces this model with the generalized Gaussian model $N(\mathbf{1}, 0, \Sigma)$ with kernel $\mathbf{1}$, and notes that only the unrooted tree is identifiable from observations at the leaves. Although Felsenstein obtains an algorithm for computing the likelihood, the induced Wishart model $\mathcal{W}_d^-(\mathbf{1}, \Delta)$ is not commonly used for the estimation of phylogenetic trees from distance matrices. The reasons for this are not entirely clear, but computational efficiency may be a consideration.

In the estimation of genetic phylogenies, it is necessary to distinguish between distances computed by counting substitutions in homologous sequence data, and squared distances computed using Euclidean distances for quantitative traits. Poisson-type models are appropriate for the former, Wishart-type models for the latter. Were it not for complications associated with insertions and deletions, and multiple substitutions at the same locus, the distance matrix D computed from sequence data would itself be a tree. This degeneracy is implied not only by the Poisson model $\mathrm{cov}(D_{ij}, D_{kl}) = |\Delta_{ij,kl}|$, but by any model for covariances that is an additive function of edge lengths. An additive covariance matrix implies that the matrix $\mathrm{cov}(D_{ij}, D_{kl})$ has rank equal to the number of edges in Δ. Since the distance matrix measured from sequence data is not ordinarily a tree, exact additivity of covariances is ruled out and the unmodified Poisson model must be rejected. Thus, an adequate model for the additional effects of insertions and deletions, multiple substitutions, transitions, transversions and non-synonymous substitutions is essential in order to model departures of D from Δ (Bulmer (1991)). This argument does not rule out approximate linearity (Gascuel (1997)) provided that the approximation is used judiciously.

The Wishart model is not degenerate in the same way that the unmodified Poisson model is degenerate: no linear combination of the components has zero variance. In addition, if each leaf edge in Δ is positive and $d \geq n$, the Wishart model has positive density at all points in the space of distance matrices. It does not assign zero probability to any event that is likely to occur. Although the Wishart model is not designed with sequence data in mind, it generates a consistent estimate of the tree, though not with maximum efficiency.

Acknowledgements

I am grateful to an associate editor and two referees for helpful comments on an earlier version of this paper. Support for this research was provided in part by NSF Grant No. DMS0305009.

References

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26**, 211-252.

Bulmer, M. (1991). Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution* **8**, 868-883.

Buneman, P. (1971). The recovery of trees from measurements of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences* (Edited by F. R. Hodson, D. G. Kendall and P. Tǎutu), 387-395. Edinburgh University Press.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *Amer. J. Human Genetics* **19**, 233-257.

Cruddas, A. M., Reid, N. and Cox, D. R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika* **76**, 231-237.

Ehrenberg, A. S. C. (1981). The problem of numeracy. *Amer. Statist.* **35**, 67-71.

Felsenstein, J. (1973). Maximum likelihood estimation of evolutionary trees from continuous characters. *Amer. J. Human Genetics* **25**, 471-492.

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

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and Evolution* **14**, 685-695.

Gascuel, O. and Steel, M. (2006). Neighbor-joining revealed. *Molecular biology and Evolution* **23**, 1997-2000.

Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325-338.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems (with discussion). *J. Amer. Statist. Assoc.* **72**, 320-340.

Hodson, F. R, Sneath, P. H. A. and Doran, J. E. (1966). Some experiments in the numerical analysis of archaeological data. *Biometrika*, **53**, 311-324.

Hodson, F. R, Kendall, D. G. and Tǎutu, P. (1970). *Proceedings of the Anglo-Romanian Conference on Mathematics in the Archaeological and Historical Sciences.* Edinburgh University Press.

Kingman, J. F. C. (1982a). On the genealogy of large populations. *J. Appl. Probab.* **19**, 27-43.

Kingman, J. F. C. (1982b). The coalescent. *Stochastic Process. Appl.* **13**, 235-248.

McCullagh, P. and Clifford, D. (2005). Evidence for conformal invariance of crop yields. *Proc. Roy. Soc. A* **462**, 2119-2143.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd edition. Chapman and Hall, London.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425.

Sarich, V. M. (1969). Pinniped phylogeny. *Systematic Zoology* **18**, 416-422.

Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika* **42**, 320-344.

Semple, C. and Steel, M. (2003). *Phylogenetics.* Oxford Series in Mathematics and its Applications, **24**. Oxford University Press.

Semple, C. and Steel, M. (2004). Cyclic permutations and evolutionary trees. *Adv. Appl. Math.* **32**, 669-680.

Stein, M.L. (1999). *Interpolation of Spatial Data.* Springer Series in Statistics.

Tunnicliffe Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *J. Roy. Statist. Soc. Ser. B* **51**, 15-27.

Department of Statistics, University of Chicago, 5734 University Ave., Chicago, Il 60637 U.S.A.

E-mail: pmcc@galton.uchicago.edu

(Received July 2006; accepted October 2007)