Structured covariance matrices in multivariate regression models

Peter McCullagh*

September 14, 2006

Abstract

A similarity matrix is a covariance matrix generated by additive nested common factors having independent components. The set of such matrices is a structured subset of covariance matrices, closed under permutation and restriction, which makes it potentially useful as a sub-model for the joint dependence of several responses. It is also equal to the set of rooted trees. Some issues connected with parameter estimation and Bayesian model formulation for such structured sets and subsets are discussed. Although the set of similarity matrices has a rich algebraic structure, the fact that it is not a manifold leads to difficulties in computational work.

Keywords: Commutative semi-group; Dissimilarity matrix; Exchangeable partition; Exchangeable sequence; Exchangeable tree; Infinite divisibility; Latent factor; Lévy fragmentation; Markov fragmentation; Multidimensional scaling; Multivariate dependence; Natural transformation; Rooted tree; Similarity matrix; Unrooted tree

1 Introduction

The chief goal of this paper is to introduce a class of structured covariance matrices potentially suitable for use in applications where the response for each subject is a moderately long list of values on closely related variables. Suppose that we are studying the effect of covariates x on q response variables $Y_1, \ldots, Y_q$ in an application where the responses on distinct units may be taken as independent q-vectors. The response values for n individuals constitute a random matrix Y of order $n \times q$ with independent rows, and the model matrix X of covariate values is of order $n \times p$. In the standard multivariate linear regression model the response distribution is

$$Y \sim N(X\beta,\, I_n \otimes \Sigma) \tag{1}$$

where the parameter space consists of the coefficient matrix β of order $p \times q$ plus the covariance matrix $\Sigma \in \mathcal{PD}_q$, the cone of symmetric positive semi-definite q-matrices. We call this the standard regression model because both components

[1]Support for this research was provided in part by NSF Grant No. DMS0305009

of the parameter space are unconstrained. The maximum-likelihood estimate of β and the residual maximum likelihood estimate of Σ are

$$X^T X\hat{\beta} = X^T Y, \quad (n - \text{rank}(X))\hat{\Sigma} = (Y - X\hat{\beta})^T (Y - X\hat{\beta}).$$

So far as parameter estimation is concerned, nothing is gained by considering the problem in multivariate form rather than q separate univariate linear regressions. However, the residual covariances do have an effect on the prediction of an individual response component given some or all of the remaining components.

The situation of interest in this paper is one in which the responses are correlated in a particular way. The presence of latent common factors implies that all correlations are non-negative, and also induces a characteristic pattern in the covariance matrix. A similarity matrix is a covariance matrix induced by latent common factors with independent components. Thus, instead of the full model $\Sigma \in \mathcal{PD}_q$, we consider the sub-model in which $\Sigma \in \Theta_q$ belongs to the subset of similarity matrices. The chief goal of the sub-model is to gain an understanding of the joint structure of the response variables rather than to increase the precision of parameter estimates.

A structured covariance class Θ implies a subset $\Theta_q \subset \mathcal{PD}_q$ of $q \times q$ matrices, one such subset for each positive integer q. The primary structure of these subsets is that Θ_q is the set of matrices obtained by deletion of the last row and column from matrices in Θ_{q+1}. In addition, for all classes considered here, Θ_q is closed under simultaneous permutation of rows and columns. In other words, the response variables are not exchangeable, but the order is immaterial, and the model for any subset of the responses belongs to the same class. Structured covariance classes may also be used to generate sub-models in contexts such as multivariate time series where the values on distinct units are not independent. For example (1) may be replaced by

$$Y \sim N(0,\, V_n \otimes \Sigma) \tag{2}$$

where V_n belongs to some suitable class of temporal covariance matrices such as AR1, i.e. $V_n(t, t') = \exp(-\lambda|t - t'|)$.

Assuming that Σ in (1) is adequately modelled as a similarity matrix, is it reasonable, or perhaps necessary, to consider parallel restrictions on the regression parameters? The answer naturally depends on the specifics of the application, but the rationale is based on the implications of the additive decomposition (4) for the similarity matrix, which determines a nested sequence of subspaces of $\mathcal{R}^q$. The following example of a Gaussian regression model with scalar parameter $\gamma \geq 0$, vectors $\theta \in \mathcal{R}^q$, $\beta \in \mathcal{R}^p$, and $\Sigma \in \Theta_q$, illustrates the principle in an extreme form.

$$E(Y_{ir}) = \theta_r + \sum_t x_{it}\beta_t, \quad \text{cov}(Y_{ir}, Y_{js}) = \big(\delta_{ij} + \gamma(XWX^T)_{ij}\big)\Sigma_{rs}, \tag{3}$$

where W is a given positive semi-definite matrix of order p. Although the rows of Y are no longer independent, the restriction of Y by rows or by columns

yields a model of the same additive form, with the same matrix W and the same parameters. In addition, the model (3) is such that two ordered sets of distinct units having the same list of covariate values also have the same joint response distribution. Thus, the characteristic property of a regression model is satisfied (McCullagh, 2005).

The class of similarity matrices induced by latent common factors is described in section 2. Section 3 shows that the set of similarity matrices is a commutative semi-group, equal to the set of rooted trees. Various properties of trees are described, including positive definiteness, closure under monotone transformation, and the minimax projection. Finally, some problems connected with parameter estimation and Bayesian model formulation are described.

2 Independent common factors

Social survey questionnaires are often designed with a battery of related questions, all addressing the same or similar issue. For example, a questionnaire on juvenile delinquency might well have six or more questions addressing specific aspects of each of the following topics: (i) incidents of theft, (ii) damage to property, (iii) violence to individuals and (iv) gang activity. Incidents of theft might be further sub-divided by type, for example car theft versus breaking and entering of buildings. Gang activities could be classified as drug-related or other. Most likely the survey contains numerous additional questions, some related to family background, some related to the history of encounters with the police and judicial system, and others to psychological issues such as self-esteem, attitude to authority and attitude to others.

The response from one individual is a list of answers to questions, some closely related, others less so. Answers to questions about family background may best be regarded as explanatory, others as the primary response. For example, if the focus is on the effect of family background on teenage violence, family background variables are explanatory, but it may be helpful to consider the psychological variables as responses in addition to incidents of actual violence. Ordinarily we should not expect these responses to be mutually independent, so any reasonable statistical model must allow for correlations. If no restrictions are imposed on the class of correlation matrices, the model has $q(q+1)/2$ covariance parameters for q response variables. Such a large number of free parameters is, in most cases, unnecessarily wasteful of resources because the range of plausible correlations in situations of this sort is a small subset of the entire model space.

Suppose that the set of q variables can be partitioned into disjoint subsets, corresponding to the primary class types or major themes in the questionnaire. The types in turn can be partitioned into sub-types, and the sub-types into sub-sub-types, and so on. This recursive partitioning by types and sub-types may be overt and observed, or it may be latent and unobserved, or there may be ambiguities in the classification. Our interest here is in the class of covariance matrices induced by exchangeable random effects in a latent recursive partition.

Associate with each variable v a primary type $t(v)$, and with each primary type t a random effect X_t, the components for different types being independent. A further random effect is associated with sub-types, these also being independent and independent of the primary-type effects. Assuming that the effects on the response are additive, the recursive construction leads to conditional covariance matrices of the form

$$\Sigma = \sigma_0^2 E_0 + \sigma_1^2 E_1 + \sigma_2^2 E_2 + \cdots \tag{4}$$

in which $E_0(r,s) = 1$, $E_1(r,s) = 1$ if variables r, s have the same primary type, $E_2(r,s) = 1$ if they have the same sub-type, and so on. Each of the matrices $E_0, E_1, E_2, \ldots$ is an equivalence relation, reflexive, symmetric, transitive and positive semi-definite. The sequence $E_0 \geq E_1 \geq E_2 \cdots$ is decreasing the sense that E_{r+1} is a sub-partition of E_r. A matrix Σ having such an additive non-increasing partition representation (4) is called a spherical similarity matrix. The components are non-negative and the diagonal elements are equal.

If the random effects X_t for the primary types are independent but not identically distributed, the decomposition takes a similar form in which the matrix $\sigma_1^2 E_1$ is replaced by the corresponding block-diagonal matrix, constant on blocks, but having possibly different variances for different blocks. To understand what this decomposition entails, consider a similarity matrix Σ of order 3 in which the initial partition by primary types is $E_1 = 12|3$. Then the response variables have the decomposition

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} X_0 \\ X_0 \\ X_0 \end{pmatrix} + \begin{pmatrix} X_1 \\ X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

in which all variables are conditionally independent given the partition. As a consequence, the conditional covariances satisfy

$$\begin{aligned}
\Sigma_{11} &= \mathrm{var}(X_0 + X_1 + \epsilon_1) \geq \mathrm{var}(X_0 + X_1) = \Sigma_{12} \\
\Sigma_{22} &= \mathrm{var}(X_0 + X_1 + \epsilon_2) \geq \mathrm{var}(X_0 + X_1) = \Sigma_{12} \\
\Sigma_{33} &= \mathrm{var}(X_0 + X_2 + \epsilon_3) \geq \mathrm{var}(X_0) = \Sigma_{13} \\
\Sigma_{12} &= \mathrm{var}(X_0 + X_1) \geq \mathrm{var}(X_0) = \Sigma_{13} = \Sigma_{23}.
\end{aligned}$$

Regardless of the primary decomposition E_1, the defining property of a similarity matrix is that the two smaller similarities from each triple $\Sigma_{ij}, \Sigma_{jk}, \Sigma_{ik}$ are equal and non-negative. Consequently the maximal element in each row occurs on the diagonal.

The general version of the additive decomposition (4) for factors having independent but non-identically distributed effects is as follows. Each Boolean matrix E_r is a partition of a subset of $[n]$ and $E_0 \geq E_1 \geq \cdots$ component-wise. The coefficients are non-negative.

3 Similarity matrices and rooted trees

3.1 Definition

A symmetric matrix T of order n whose components satisfy the two conditions

$$0 \leq T_{ij} \leq \infty, \quad T_{ij} \geq \min\{T_{ik}, T_{jk}\} \tag{5}$$

for all i, j, k is called a rooted n-tree. The term 'rooted tree' is used in the standard graph-theoretic sense of a connected undirected leaf-labelled acyclic graph. The n leaves or terminal nodes are labelled by the elements of $[n] = \{1, \ldots, n\}$, and the label on each branch is the set of terminal nodes on that branch so the first node following the root is $[n]$. Each edge has a length, and T_{ij} is the distance measured along the tree edges from the root node to the first junction at which leaves i, j occur on separate branches. The tree inequality (5) is the condition for a non-negative symmetric matrix to admit a tree-diagram representation. The relationship between rooted trees and similarity matrices is straightforward; every similarity matrix is a rooted tree, and vice-versa. In fact, it is convenient to adopt (5) as the definition of of a similarity matrix, thereby erasing any possible distinction.

The inequality $T_{ii} \geq T_{ik}$ means that the distance to terminal node i is greater than the distance to the junction at which leaves i, k occur on different branches. In terms of similarity matrices, object i is more similar to itself than to any other object, and the degree of similarity can be infinite. Furthermore, for any set of three elements i, j, k, the inequality $T_{ij} \geq \min\{T_{ik}, T_{jk}\}$ is satisfied by all permutations. The only possibilities, not mutually exclusive, are

$$T_{ij} = T_{jk} \leq T_{ik}, \quad T_{ij} = T_{ik} \leq T_{jk} \quad \text{and} \quad T_{ik} = T_{jk} \leq T_{ij}.$$

In other words, each set of three numbers T_{ij}, T_{jk}, T_{ik} contains a duplicate pair equal to the minimum.

3.2 Closure under minima

Let S, S' be two rooted n-trees. Then the matrix $T = S \wedge S'$ with components $T_{ij} = \min\{S_{ij}, S'_{ij}\}$ is also a rooted n-tree. The proof follows directly from the rooted tree inequality.

$$\begin{aligned} T_{ij} &= \min\{S_{ij}, S'_{ij}\} \\ &\geq \min\{\min\{S_{ik}, S_{jk}\},\ \min\{S'_{ik}, S'_{jk}\}\} \\ &= \min\{S_{ik}, S'_{ik}, S_{jk}, S'_{jk}\} \\ &= \min\{\min\{S_{ik}, S'_{ik}\},\ \min\{S_{jk}, S'_{jk}\}\} \\ &= \min\{T_{ik}, T_{jk}\}. \end{aligned}$$

Formally, this binary operation on matrices makes the set of rooted n-trees into a commutative semi-group containing an identity element, the matrix with all components infinite, and a zero element, the zero matrix. In addition, we

observe that if $x = (x_1, \ldots, x_n)$ is a sequence of non-negative numbers, the matrix $(x \wedge x)_{ij} = \min(x_i, x_j)$ is a tree. Further, the set of such trees is a sub-semi-group, closed under minima, $(x \wedge x) \wedge (x' \wedge x') = \min(x, x') \wedge \min(x, x')$. It is also closed under permutation and restriction to subsets. Further subsets having these properties are described in the next section.

The transformation from the sequence x to the matrix $x \wedge x$ enables us to construct a limited class of exchangeable random trees from exchangeable random sequences. The importance of the semi-group property is that it permits the construction of infinitely divisible distributions (section 7), and continuous-time tree-valued Markov processes with independent increments, similar in many respects to Lévy processes (section 7.4–7.7).

3.3 Structured subsets

The set $\mathcal{RT}_n$ of rooted n-trees, i.e. the set of $n \times n$ matrices satisfying (5), is a simply-connected set of positive semi-definite matrices. It is not convex for $n \geq 3$, nor is it a simple manifold of fixed dimension. It is in fact a finite union of $2n-1$-dimensional manifolds, each manifold corresponding to a distinct Boolean tree, and all containing the central $n+1$-dimensional set of radial trees defined as follows.

A tree such that $T_{ij} = T_{kl}$ for $i \neq j$ and $k \neq l$ is called radial. Equivalently,

$$T_{ij} = \begin{cases} x_0 + x_i & \text{if } i = j \\ x_0 & \text{if } i \neq j \end{cases}$$

with $x_0 \geq 0$ and $x_i \geq 0$. For $n \geq 1$, the set of radial trees is a manifold of dimension $n+1$. It is an additive semi-group containing zero and closed under addition of matrices.

If $T_{ii} = \rho > 0$ for all i, the tree is said to be spherical with radius ρ: for finite ρ, all leaves are equi-distant from the root. The set $\mathrm{Sph}(\rho)$ of spherical rooted trees or similarity matrices of radius ρ is a subset of dimension $n-1$, a finite union of $n-1$-dimensional manifolds. The set of all spherical trees is equal to the set of exchangeable similarity matrices (4).

An exchangeable covariance matrix is a tree such that $T_{ii} = x$ and $T_{ij} = x'$ for $i \neq j$. The set of such trees is a manifold of dimension 2, equal to the intersection of radial trees and spherical trees.

The set of rooted trees contains numerous subsets of the above type, each closed under permutation and restriction, and each also a semi-group under the component-wise minimum operation. For example, every partition of $[n]$, regarded as an equivalence relation $E\colon [n] \times [n] \to \{0, 1\}$, is a tree. The set $\mathcal{E}_n$ of partitions of $[n]$ is a closed finite subset of trees, and an exchangeable random partition is an exchangeable random tree with a discrete distribution.

The set $\mathcal{RT}_n$ of rooted n-trees or similarity matrices is not convex, but it is closed under addition of radial trees. In other words, if $T \in \mathcal{RT}_n$ is a rooted tree and T' is a radial tree, the matrix sum $T + T'$ is a rooted tree.

3.4 Monotone transformation

Let T be a rooted tree and let T' be defined by component-wise transformation

$$T'_{ij} = g(T_{ij})$$

where g is non-decreasing and non-negative. Since the defining inequality $T_{ij} \leq \min\{T_{ik}, T_{jk}\}$ is preserved by monotone transformation, the transformed matrix is also a rooted tree. Furthermore, the set of spherical trees and the set of radial trees are also closed under monotone transformation.

Two examples suffice to illustrate such transformations. If $T_{ii} = T_{jj}$ for each i, j (possibly infinite), the component-wise transformation $T' = T/(1+T)$ yields a finite spherical tree. Second, if h is the heaviside function in left-continuous form, $h(x) = 1$ for $x > 0$ and zero for $x \leq 0$, then $h(T)_{ij}$ is either zero or one. The transformed tree is a pseudo-partition, a partition of a subset of $[n]$, symmetric and transitive but not necessarily reflexive.

3.5 Positive definiteness

Let T be a rooted n-tree. For each $t \geq 0$, define the Boolean matrix E_t by

$$E_t(i,j) = \begin{cases} 1 & \text{if } T_{ij} \geq t \\ 0 & \text{otherwise.} \end{cases}$$

Evidently $E_t(i,j) = E_t(j,i)$ is symmetric, non-increasing in t, and continuous on the left. If $E_t(i,j) = E_t(j,k) = 1$ then $T_{ij} \geq t$ and $T_{jk} \geq t$. By inequality (5), $T_{ik} \geq \min\{T_{ij}, T_{jk}\} \geq t$, so $E_t(i,k) = 1$. Thus E_t is transitive. However, E_t is not necessarily reflexive unless $T_{ii} \geq t$ for each i, which is not guaranteed.

For present purposes, we observe that E_t is a partition of a subset of the units, namely those such that $T_{ii} \geq t$. For those leaves occurring at height t or above, it tells us which ones belong to the same branch at this height. The matrix is thus symmetric and positive semi-definite. Furthermore, the matrix T is expressible as an integral

$$T(i,j) = \int_0^\infty E_t(i,j)\, dt \tag{6}$$

i.e. a non-negative linear combination of positive semi-definite matrices similar to (4). It follows that T is positive semi-definite if it is finite, i.e. if $T_{ii} < \infty$ for each i.

Another interpretation of T uses Brownian motion with T as the index set as follows. Let W be a Brownian motion on T starting at zero at the root node. The process is continuous on T with independent zero-mean Gaussian increments on non-overlapping tree intervals. Given the value at a junction, the process proceeds independently on each branch. The values $W_i = W(T_{ii})$ at the terminal nodes are jointly Gaussian with covariance matrix T.

The set $\mathcal{PD}_n$ of covariance matrices of order n is a convex manifold of dimension $n(n+1)/2$, where the set $\mathcal{T}_n$ of similarity matrices is a finite union

of manifolds of dimension $2n-1$. Thus, similarity matrices constitute a small subset of covariance matrices. Their main attraction for statistical purposes is that each similarity matrix is a covariance matrix that has a simple statistical interpretation and can be depicted as a tree. This graphical representation illustrates in quantitative fashion, the relationships between units or specimens, or between the responses to related items in a survey. Although most similarity matrices are strictly positive definite, the range of relationships that can be represented by a similarity matrix is fairly limited. For example, all covariances are non-negative. Furthermore, the set of similarity matrices is not closed under re-scaling of variables, $T_{ij} \mapsto x_i x_j T_{ij}$, so there is an implicit assumption that all variables are measured on the same scale or on closely related scales. If the covariance matrix V belongs to $\mathcal{RT}_n$ the associated correlation matrix may not belong to the same set. This remark does not apply to spherical trees, which are constant on the diagonal.

3.6 Graph representation

To construct a graph-theoretic tree diagram from a matrix $T \in \mathcal{RT}_n$ we proceed as follows. The root node is first established on the baseline at height zero, and the edge attached to the root node has length $t = \min(T)$. Each leaf i such that $T_{ii} = t$ is attached directly to the vertex at the end of the root edge by an edge of zero length. The indicator matrix E_{t^+} for $T > t$ is a partition of the surviving subset of $[n]$ into blocks, each block defining a branch attached to the same vertex. The residual height matrix $T - t$ has one block for each branch. The construction then proceeds recursively on each non-zero branch with T replaced by the appropriate block of $T - t$. Two such trees are shown in section 6.2.

3.7 Minimax projection

The minimax projection is a transformation on symmetric n-matrices whose image is the set $\mathcal{RT}_n$ of rooted n-trees. Let A be a symmetric matrix, and let $A^+ = \max(A, 0)$ be the component-wise maximum. The projection is defined by $(TA)_{ii} = \max\{A^+_{i1}, \ldots, A^+_{in}\}$, and for $i \neq j$ by

$$(TA)_{ij} = \min_{I|J} \max_{r \in I, s \in J} A^+_{rs},$$

where the minimum is taken over partitions $I|J$ of $[n]$ with $i \in I$ and $j \in J$.

The minimax projection is equivariant in two senses, with respect to permutation and with respect to monotone transformation. Equi-variance is a commutativity condition $TgA = gTA$ in which g denotes either permutation or component-wise non-decreasing transformation. Clearly, $(TA)_{ij} \geq A_{ij}$, so the projection is component-wise non-decreasing.

The minimax projection has a direct max-flow min-cut interpretation. Let A be a network represented as a graph in which $A_{rs} \geq 0$ is the capacity (bandwidth) of the direct physical link or edge from node r to s. The capacity of a

path from i to j is limited by the weakest link, the minimum of the capacities of the edges on that path. In most physical systems such as electrical grids or fluid transmission pipes, the total transmission capacity of two non-overlapping paths from i to j is the sum of the two capacities. However, the total capacity in our mathematical system is not the sum but the maximum over all paths $i \to j$, which is given by transformation TA shown above. The idea is that any flow along a path from i to j is also a flow from the set I into the set J through some edge (r, s) with $r \in I$ and $s \in J$.

Since $TA \in \mathcal{RT}_n$,

$$TA = \sigma_0^2 E_0 + \sigma_1^2 E_1 + \cdots$$

in which each matrix E_r is a partition of a subset of $[n]$, and $E_0 \geq E_1 \geq \cdots$ component-wise. Thus σ_0^2 is the minimum element of TA, $\sigma_0^2 + \sigma_1^2$ is the second-smallest element, and so on. If A is a covariance matrix, this decomposition is helpful as an initial step in computing the maximum-likelihood estimate. Given the matrices $E_0, E_1, \ldots$ as determined by the projection, the maximum-likelihood coefficients are easily computed by standard techniques. However, the matrices themselves are a part of the parameter space, and there is no guarantee that the maximum-likelihood projection is a non-negative combination of the minimax projection matrices.

4 Dissimilarity matrices

4.1 Distance function

If $x_1, \ldots, x_n$ are points in Euclidean space, and $T_{ij} = \langle x_i, x_j \rangle$ is the inner product, then the inter-point squared distances are given by

$$d_{ij} = T_{ii} + T_{jj} - 2T_{ij}. \tag{7}$$

For present purposes we use the same relationship to associate with each finite similarity matrix T a corresponding dissimilarity or distance matrix d. In the graph-theoretic representation of T, d_{ij} is the distance (not squared distance) measured along the tree edges from leaf node i to leaf node j. The larger the value of d_{ij}, the more dissimilar are the objects.

Gower (1966) uses the spectral decomposition of the matrix of dissimilarities to create a Euclidean point configuration whose inter-point distances are in approximate agreement with the given dissimilarities. The ideas behind dimension reduction and multi-dimensional scaling can be traced back to Torgersen (1958), and Shepard (1962a, b). Dissimilarity matrices are also used in non-metric scaling (Kruskal, 1964), to give a low-dimensional Euclidean representation of the points in such a way that the distance between points is approximately monotone in the dissimilarities. Techniques of this sort have been used for seriation analysis in archaeology (Hodson, Sneath and Doran 1966). The terms similarity matrix and dissimilarity matrix are used in a narrower, more tightly defined, sense than in the multi-dimensional scaling literature.

A symmetric matrix d of order n is called an unrooted n-tree if $0 \leq d_{ij} < \infty$, $d_{ii} = 0$ and

$$d_{ij} + d_{kl} \leq \max\{d_{ik} + d_{jl},\ d_{il} + d_{jk}\}$$

for all i, j, k, l not necessarily distinct. The rationale for this condition can be seen from the graph in Fig. 1, where five edge lengths determine the six pairwise distances. The triangle inequality is obtained by setting $k = l$, so d is a distance function. It is not necessarily a metric because $d_{ij} = 0$ need not imply $i = j$. The set of unrooted n-trees is denoted by $\mathcal{UT}_n$.

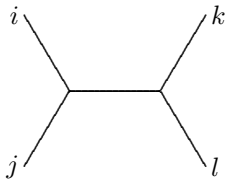

Figure 1: An unrooted tree with four terminal nodes.

An unrooted tree is called radial if $d_{ij} = x_i + x_j$ for $i \neq j$ with $x_i \geq 0$. For $n \geq 2$, the set of unrooted radial trees is a manifold of dimension n, containing zero and closed under addition of matrices. The set $\mathcal{UT}_n$ is not closed under addition, but it is closed under addition of radial trees.

An unrooted tree is called linear or one-dimensional if $d_{ij} = |x_i - x_j|$, with $x_i \geq 0$. The set of linear trees is an $n-1$-dimensional manifold.

A symmetric matrix d is called a coalescent tree if $0 \leq d_{ij} < \infty$, $d_{ii} = 0$ and

$$d_{ij} \leq \max\{d_{ik},\ d_{jk}\}$$

for all i, j, k not necessarily distinct. The triangle inequality follows directly, so d is a distance function but not necessarily a metric. Kingman's n-coalescent (Kingman, 1982) is a probability distribution on the subset of coalescent trees, exchangeable and self-consistent for different n.

Since the coalescent tree inequality is satisfied by all permutations, each set of three numbers d_{ij}, d_{ik}, d_{jk} contains a duplicate pair equal to the maximum of the three. Equivalently, every set of three points $\{i, j, k\}$ forms an isoceles triangle that is also acute in the sense that the unequal angle does not exceed $60°$ in a Euclidean representation. Thus Coal_n is the set of n-point Euclidean configurations in which each of the $\binom{n}{3}$ triangles is acute and isoceles.

As a set of matrices, Coal_n is closed under component-wise non-decreasing transformations such that $g(0) = 0$. Furthermore, $d \in \text{Coal}_n$ implies that the matrix with components $T_{ij} = \rho/(1 + \rho d_{ij})$ is a rooted n-tree with $T_{ii} = \rho$, a spherical tree of radius ρ. Conversely, every strictly positive rooted n-tree with $T_{ii} = \rho$ determines a coalescent tree by the inverse transformation $d = (\rho - T)/(\rho T)$ applied component-wise.

It is fairly obvious geometrically that every coalescent tree is also an unrooted tree, but the proof is less straightforward than it might appear. Suppose first that the configurations of the two triangles $\{i, j, k\}$ and $\{j, k, l\}$ are such that

$$d_{ij} = d_{ik} \geq d_{jk}, \quad d_{kl} = d_{jk} \geq d_{jl},$$

implying $d_{ij} \geq d_{jl}$. Since the triangle $\{i, j, l\}$ is acute and isoceles, we must have $d_{ij} = d_{il} \geq d_{jl}$. Consequently $d_{ij} + d_{kl} = d_{il} + d_{jk}$. All other configurations are such that

$$d_{ij} \leq d_{ik}; \quad d_{kl} \leq d_{jl},$$

from which we obtain directly $d_{ij} + d_{kl} \leq d_{ik} + d_{jl}$, or $d_{ij} + d_{kl} \leq d_{il} + d_{jk}$ if the indices need to be permuted.

Although radial and coalescent trees are in a sense complementary, there exist unrooted trees that cannot be expressed as the sum of a radial and a coalescent tree. However, given an unrooted tree d, we can find a coalescent d' and a radial tree d'' such that $d = d' - d''$. Likewise, an arbitrary finite rooted tree cannot be expressed as the sum of a spherical and a radial tree, but it can be expressed as a difference of two such trees.

4.2 Operations on unrooted trees

The set of unrooted trees is closed under permutation and restriction, so this is a set with the structure needed for a parameter space in a statistical model. Unrooted trees are widely used in genetics to model divergence times between species, but the tree topology is usually assumed known, so only branch lengths need to be estimated. The matrices $d \in \mathcal{UT}_n$ are not positive semi-definite, but the points can be embedded isometrically in a Euclidean space, so $\exp(-d_{ij})$ is positive definite, and $-d_{ij}$ is positive definite on contrasts. The set of unrooted trees is not a semi-group, nor is it closed under monotone transformation. In these respects, dissimilarity matrices do not behave like similarity matrices.

Despite the limitations of the algebraic structure, at least two operations on unrooted trees occur naturally in statistical work. An unrooted tree $\Delta \in \mathcal{UT}_n$ associates with each ordered pair of elements (i, j), a path in the tree, and each pair of paths (i, j) and (k, l) has an intersection whose signed length is

$$2\Delta^{\wedge 2}_{ij,kl} = -\Delta_{ik} - \Delta_{jl} + \Delta_{il} + \Delta_{jk}.$$

The sign is positive if the intersection is traversed in the same direction on each path, and negative otherwise. For the particular configuration of labels shown in Fig. 1, $\Delta^{\wedge 2}_{il,jk}$ is positive and $\Delta^{\wedge 2}_{il,kj}$ is negative. The array $\Delta^{\wedge 2}$ has the following symmetries

$$\Delta^{\wedge 2}_{ij,kl} = \Delta^{\wedge 2}_{kl,ij} = -\Delta^{\wedge 2}_{ij,lk}.$$

In particular, self-intersections satisfy $\Delta^{\wedge 2}_{ij,ij} = \Delta_{ij}$. Clearly $\Delta^{\wedge 2}$ is not a tree, nor is $|\Delta^{\wedge 2}|$.

In genetical work, each path (i, j) in the tree represents a period of history during which a record of genetic events such as mutations and substitutions accumulates on the genome. Thus, Δ_{ij} is the genetic divergence time between species. The intersection of two paths (i, j) and (k, l) is the period of genetic history that is shared by the two divergence paths. Observed divergences Y_{ij} may be obtained from homologous genomic sequences by counting the number of segregating loci for each pair of species. Because of their shared history, observed

divergences are not independent. In the simplest models, the covariance of Y_{ij} and Y_{kl} is proportional to the path intersection length $|\Delta^{\wedge 2}_{ij,kl}|$.

The second operation is akin to the Kronecker product. It measures the tree distance between paths, so it is a distance function on paths.

$$2\Delta^{\otimes 2}_{ij,kl} = \max\{\Delta_{ik} + \Delta_{jl},\ \Delta_{il} + \Delta_{jk}\} - \Delta_{ij} - \Delta_{kl}.$$

It also satisfies the conditions for an unrooted tree. Clearly, $\Delta^{\otimes 2}_{ii,jj} = \Delta_{ij}$ and $\Delta^{\otimes 2}_{ij,ij} = 0$. By definition, each product $\Delta^{\wedge 2}_{ij,kl}\Delta^{\otimes 2}_{ij,kl}$ is zero. For non-intersecting paths, $\Delta^{\otimes 2}_{ij,kl} = 0$ implies an additional symmetry $\Delta^{\wedge 2}_{il,jk} = \Delta^{\wedge 2}_{ik,jl}$, which is evident from Fig. 1.

5 Natural transformation

A transformation on trees that commutes with permutation and restriction is called natural. Let $g_n\colon \mathcal{RT}_n \to \mathcal{RT}_n$ be a transformation on rooted n-trees defined for each integer n, and let $T \in \mathcal{RT}_n$ be a rooted tree whose leading sub-matrix of order $n-1$ is $T^\dagger$. Then g is natural if the restriction $(g_nT)^\dagger$ of the transformed tree g_nT is equal to the transformation $g_{n-1}(T^\dagger)$ of the restricted tree or sub-matrix. Natural transformations predominate in applied work, where it is conventional to write g without the subscript. Natural transformations $\mathcal{RT}_n \to \mathcal{UT}_n$ from rooted to unrooted trees, and $\mathcal{UT}_n \to \mathcal{UT}_n$ are defined in the same way. The composition of composable natural transformations is also natural. The term 'restriction' is understood to include permutations, so a natural transformation also commutes with permutations.

On the partition lattice, the two transformations $\mathcal{E}_n \times \mathcal{E}_n \to \mathcal{E}_n$, least upper bound and greatest lower bound, illustrate the issues involved. One is natural and the other is not, as can be seen from the following table. The middle column shows two partitions $E, E' \in \mathcal{E}_4$, their greatest lower bound and least upper bound, and the last column shows the restriction of each partition to [3].

	$[n] = [4]$	$[n] = [3]$
E_n	1|2|34	1|2|3
E'_n	1|24|3	1|2|3
$(E \wedge E')_n$	1|2|3|4	1|2|3
$(E \vee E')_n$	1|234	1|23

We see that the restriction of $E \wedge E'$ to [3] is equal to the minimum of the two restrictions. But the restriction of $E \vee E'$ to [3] is 1|23, whereas the maximum of the restrictions is 1|2|3. This example suffices to show that the lattice maximum does not commute with restriction to subsets.

Natural transformations are of interest for the following reason. If T is an infinitely exchangeable random rooted tree and $g\colon \mathcal{RT}_n \to \mathcal{UT}_n$ is natural, then gT is a random unrooted tree, also infinitely exchangeable. The same holds for transformations $\mathcal{RT}_n \to \mathcal{RT}_n$, $\mathcal{UT}_n \to \mathcal{UT}_n$, $\mathcal{E}_n \times \mathcal{E}_n \to \mathcal{E}_n$, and so on.

The insertion maps $\mathrm{Coal}_n \to \mathcal{UT}_n$, and radial into $\mathcal{UT}_n$ are both natural, as are the corresponding insertions for rooted trees. Component-wise monotone transformation $\mathcal{RT}_n \to \mathcal{RT}_n$ is natural, as is monotone transformation on the subsets of spherical and radial trees. Monotone transformation does not preserve unrooted trees, but it is natural on the subset of coalescent trees provided that $g(0) = 0$. The distance map

$$(gT)_{ij} = T_{ii} + T_{jj} - 2T_{ij}$$

from rooted to unrooted n-trees is natural. The most important natural transformation in this paper is the semi-group product $\wedge\colon \mathcal{RT}_n \times \mathcal{RT}_n \to \mathcal{RT}_n$, which is also commutative.

Less obvious natural transformations on rooted trees include $T \mapsto \mathrm{diag}(T)$ with off-diagonal values set to zero, the matrix with components $\min(T_{ii}, T_{jj})$, and the block factor transformation (10) from sequences into partitions. From rooted to unrooted trees, we have

$$d_{ij} = \begin{cases} g(T_{ij}) & \text{if } i \neq j \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$$d_{ij} = g(T_{ij}) - \max\{g(T_{ii}),\, g(T_{jj})\} \tag{9}$$

where g is non-negative, non-increasing and finite. The first transformation gives a coalescent tree, and the second a non-coalescent in general. In the reverse direction, the component-wise reciprocal transformation $\mathrm{Coal}_n \to \mathcal{RT}_n$ is natural, and the image is a spherical rooted trees with infinite radius. If $g(\cdot) = \infty$ is permitted in (8), the image is a distance matrix containing infinite values, which is interpretable as a graph in which each connected component is an unrooted tree, i.e. d is forest of unrooted trees.

Given an unrooted tree $d \in \mathcal{UT}_n$, it is possible to construct a rooted tree $T \in \mathcal{RT}_n$ in various ways such as

$$2T_{ij} = \begin{cases} 0 & \text{if } i = n \text{ or } j = n \\ d_{ni} + d_{nj} - d_{ij} & \text{otherwise.} \end{cases}$$

This transformation is not natural, but T can be used for graphical purposes to display the unrooted tree d. Each multiple of the identity map $\mathcal{UT}_n \to \mathcal{UT}_n$ on unrooted trees is natural. Apart from these there does not appear to be a non-trivial natural transformation from unrooted trees into any other type or sub-type.

The term 'natural' is taken from category theory, and within the realms of that theory, only natural transformations are relevant. Although natural transformations have a special place in probabilistic work, the term 'natural' has a slightly unfortunate pejorative connotation. As it happens, non-natural transformations are important in statistical work. For example, the minimax projection is not natural even though it commutes with permutation. In fact, covariance estimation, considered as a transformation on matrices, is almost never natural.

6 Elementary statistical models

6.1 Estimation of similarity matrices

We consider in this section the simplest type of statistical model in which the observations $Y_1, \ldots, Y_n$ are independent and identically distributed $N_q(0, \Sigma)$ with parameter space $\Sigma \in \mathcal{RT}_q$. The unrestricted estimate $S = n^{-1}YY^T$ is sufficient, so it is natural to begin with the minimax projection $\tilde{\Sigma} = TS$. The projected value determines a tree consisting of a decreasing sequence of pseudo-partitions $E_0 > E_1 > \cdots$, together with non-negative coefficients such that

$$TS = \sigma_0^2 E_0 + \sigma_1^2 E_1 + \cdots.$$

Given the pseudo-partitions, we estimate the coefficients by maximum likelihood in the standard way using Newton-Raphson iteration. The global maximum could occur on a tree of a different Boolean type, so there is no guarantee that this computational strategy will produce the global maximum.

The data for the following example were generated using a covariance matrix consisting of two uncorrelated blocks of three variables each. In the first block, the variances are 2.0 and covariances 1.0; in the second block, the variances are 4.5 and covariances 2.25. The sample covariance matrix on 20 degrees of freedom, the minimax projection TS, and the maximum-likelihood estimate $\hat{\Sigma} \in \mathcal{RT}_6$ are

$$S = \begin{pmatrix} 2.23 & 1.11 & 1.14 & -0.47 & -0.60 & -0.52 \\ 1.11 & 2.11 & 0.72 & 0.44 & 0.11 & 0.55 \\ 1.14 & 0.72 & 1.87 & 0.32 & 0.16 & 0.57 \\ -0.47 & 0.44 & 0.32 & 4.74 & 3.55 & 5.00 \\ -0.60 & 0.11 & 0.16 & 3.55 & 3.99 & 4.43 \\ -0.52 & 0.55 & 0.57 & 5.00 & 4.43 & 6.71 \end{pmatrix},$$

$$TS = \begin{pmatrix} 2.23 & 1.11 & 1.14 & 0.57 & 0.57 & 0.57 \\ 1.11 & 2.11 & 1.11 & 0.57 & 0.57 & 0.57 \\ 1.14 & 1.11 & 1.87 & 0.57 & 0.57 & 0.57 \\ 0.57 & 0.57 & 0.57 & 5.00 & 4.43 & 5.00 \\ 0.57 & 0.57 & 0.57 & 4.43 & 4.43 & 4.43 \\ 0.57 & 0.57 & 0.57 & 5.00 & 4.43 & 6.71 \end{pmatrix},$$

$$\hat{\Sigma} = \begin{pmatrix} 2.06 & 0.91 & 1.14 & 0.00 & 0.00 & 0.00 \\ 0.91 & 2.11 & 0.91 & 0.00 & 0.00 & 0.00 \\ 1.14 & 0.91 & 2.04 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 5.46 & 3.92 & 4.86 \\ 0.00 & 0.00 & 0.00 & 3.92 & 3.99 & 3.92 \\ 0.00 & 0.00 & 0.00 & 4.86 & 3.92 & 5.70 \end{pmatrix}.$$

For this example, E_0 is the single-block partition with all elements one, and E_1 is the partition into two blocks of size three each. The coefficients in TS are 0.57 and 0.54, while the coefficients in $\hat{\Sigma}$ are 0.0 and 0.91. The maximum-likelihood estimate satisfies $\mathrm{tr}(\hat{\Sigma}^{-1}S) = q$.

The unconstrained log likelihood is maximized at S: the values at S, $\hat{\Sigma}$ and TS are -91.96, -95.92 and -107.16. The block structure is evident in $\hat{\Sigma}$, and to a lesser extent in TS.

6.2 Graphical displays for similarity matrices

Apart from the fact that similarity matrices arise in a natural way, one added advantage is that the tree representation provides a simple and easily interpretable graphical representation. The left tree in Fig. 2 shows the block-diagonal matrix Σ described in the preceding section, with variables as terminal nodes labelled a–f. The sample covariance matrix S is not a tree, but the minimax projection TS is the tree shown on the right. The covariances of the variables a–c with d–f are equal to the length of the root edge, which is zero for Σ and 0.57 for TS. The covariance of d with e is the length of the common path from the root node, which is 2.25 in the left tree and 3.86 in the right tree.

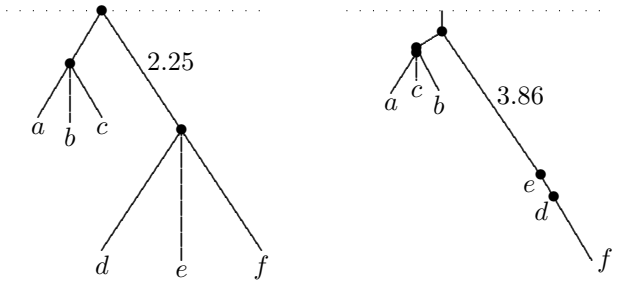

Fig. 2: Graphical depiction of two similarity matrices, Σ on the left and TS on the right, with variables labelled a–f. The root edge of TS has length 0.57, and the root edge of Σ has length zero, Two additional edge lengths are given to indicate the scale.

6.3 Structured correlations

The methods and models of this paper are designed for covariance matrices, not correlation matrices. Nevertheless, the following example taken from Ehrenberg (1981) shows that the computational 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. In Table 1a, the first five are sports programmes, and the last five are news and current affairs.

A function for fitting similarity matrices was written in R. It takes the $2q-1$ binary basis matrices generated by the minimax projection and uses Newton-Raphson to compute the coeffients. 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 given in Table 3 of Ehrenberg (1981), with ITV programmes followed bt BBC programmes in alphabetical order. The output of the composite function `round(TreeFit(corr)[perm,`

Table 1a. Viewer preference correlations 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 viewer preference correlations

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

`perm], 2)` 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 a contrast between sports programmes and current affairs. The same contrast and the resulting simplification were also noted by Ehrenberg, who used this permutation to illustrate the superiority of tables over graphs for conveying quantitative information. Within sports programmes, the main contrast is between Rugby Special and the others, which are soccer and professional boxing. The main point here is that the tree model automatically looks for and detects this structure by fitting a similarity matrix. In fact, the permutation determined by the minimax projection differs slightly from Ehrenberg's permutation by reversing the order of MoD and GrS.

The correlations in Table 1a are obtained from binary responses with probabilities likely to be in the range $(0.2, 0.8)$. Consequently, the variances are nearly equal for all ten response variables, so the transformation to correlations is approximately a scalar multiple in the range 4–6. The minimax projection of a correlation matrix is a correlation matrix, but the maximum-likelihood calculations are based on the assumption that the input matrix is a covariance matrix, not a correlation matrix. Consequently, if the model space includes all rooted trees, the maximum-likelihood matrix is not a correlation matrix. The diagonal elements of the fitted matrix in Table 1b range from 0.96 to 1.03. Finally, the fitted correlation between sports programmes and current affairs is

approximately 8% smaller than the average of the observed correlations.

It may appear that the matrix of fitted covariances is not a good approximation to the observed covariances. However, the deviance is only $0.051n$, where n is the sample size or degrees of freedom. Thus, if $n = 1000$, the deviance is 51.0. To see whether this value is large, a small-scale simulation was run with Gaussian data generated from the distribution with covariance matrix in Table 1b. For each simulation, the fitted similarity matrix was obtained, and the deviance computed. For $n = 100$, the null distribution of deviances is roughly $1.06\chi^2_{36}$, only slightly larger than the nominal χ^2 on $(q-1)(q-2)/2$ degrees of freedom. The nominal approximation is even better for $n = 1000$. Thus, if the sample size is 1500 or less, the discrepancy between the observed covariance matrix and the fitted similarity matrix is not appreciable.

It is straightforward to do exactly the same sort of calculations for the reduced model in which the parameter space is the set of spherical trees, constant on the diagonal. The fitted matrix obtained from `TreeFit(corr, sph=1)` 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 $0.053n$.

6.4 Models using unrooted trees

Most models involving unrooted trees come from population genetics, where the fundamental observation for each species is an aligned sequence from the genetic alphabet. The following account is greatly simplified and ignores many of the complications that arise in the analysis of actual sequences.

All of the information about the ancestral tree $\Delta \in \mathcal{UT}_n$ resides in differences between sequences, and much of that information comes from the matrix Y of pairwise distances, where Y_{rs} is the number of loci at which the sequences for species r and s differ. If there are no insertions or deletions, and all substitutions occur at distinct loci, the observed matrix Y is an unrooted tree. In practice, the matrix of observed distances is usually not a tree. Its expectation depends on Δ, and the same tree also governs covariances. The covariance of Y_{rs} and Y_{tu} is proportional to the length $|\Delta^{\wedge 2}_{rs,tu}|$ of the common ancestral lineage.

Although Poisson models predominate in the genetics literature, we consider here a simplified Gaussian version in which the joint distribution of Y is determined by the moments

$$E(Y_{rs}) = \Delta_{rs}, \quad \operatorname{cov}(Y_{rs}, Y_{tu}) = \sigma_0^2 \delta_{rs,tu} + \sigma_1^2 |\Delta^{\wedge 2}_{rs,tu}|$$

with mean Δ and two variance components. The white-noise variance with uncorrelated components $\delta_{rs,tu}$ is included because the model is otherwise singular, i.e. $\sigma_0^2 = 0$ implies that $Y \in \mathcal{UT}_n$.

If the observed matrix happens to belong to $\mathcal{UT}_n$, then clearly $\hat{\Delta} = Y$ and the estimated variance components are both zero. Otherwise, $\hat{\Delta}$ is obtained by maximum likelihood using special-purpose software, which is unlikely to be developed in the absence of a compelling need.

The motivation for the preceding model comes from genetics. However, models that are at least superficially similar arise in totally unrelated areas. Suppose that Y_i is the concentration of pollutant at site i in a river network. If we are dealing with airborne pollutants, the covariance between sites is likely to decrease with physical distance. If we are dealing with water-borne pollutants or pests transported by fish, the covariance may have an additional component best described by river distance, which is the tree metric. In other words, an additive variance-components model with three or more sources of variance may be needed. In examples of this sort (Cressie et al. 2006), the tree is observed and is not a part of the parameter space.

6.5 Unresolved matters

Apart from matters of computation, there is no difficulty in extending the regression model to include covariate effects. The technique described in section 6.1, of first computing the unrestricted covariance matrix and using the minimax projection as a first step, is applicable quite generally, although not to the sub-model (3). However, it is not guaranteed to produce the maximum-likelihood estimate because it may not identify the partitions $E_0, E_1, \ldots$ correctly. A more reliable computational technique is needed, but this does not appear to be a major obstacle.

Assuming that a reliable algorithm is available for maximum-likelihood computation, it is necessary to test the adequacy of the sub-model $\Sigma \in \mathcal{RT}_q$ by comparison with the full model $\Sigma \in \mathcal{PD}_q$. For this purpose, the approximate null distribution is needed. Since $\mathcal{RT}_q \subset \mathcal{PD}_q$ is not a sub-manifold, the problem is not entirely regular. It is not clear whether standard asymptotic theory treating $\mathcal{RT}_q$ locally as a manifold of dimension $2q-1$ is likely to prove satisfactory for $q \geq 3$.

It may be the case that the most effective use of similarity matrices in applied work lies in an associated decomposition of the space of regression coefficients, i.e. multivariate regression models intermediate between the extremes represented by (1) and (3). The use of random similarity matrices in Bayesian models for regression coefficients is also a distinct possibility.

From a more pragmatic point of view connected with specific applications, we also need to ask the following:

- whether, in this specific area of application, covariances are adequately modelled by similarity matrices;
- whether our understanding of the dependence structure is appreciably improved from the use of the sub-model and the graphical representation of covariances;
- whether the similarity matrices estimated from similar surveys in the past have any implications for the design of future questionnaires. In other words, the within-block covariances may be so large that further questionnaire items on the topic may provide little additional information.

7 Random similarity matrices

7.1 Exchangeable matrices

Certain types of applications involving similarity matrices call for probability distributions on the set $\mathcal{RT}_n$ or on $\mathcal{UT}_n$ or on certain structured subsets. For example, genetic models for studying ancestral lineages frequently use Kingman's coalescent model. Additional examples include the Gauss-Ewens model (McCullagh and Yang, 2006) and more complicated Dirichlet allocation schemes (Blei, Ng and Jordan, 2003) for classification and clustering. Since the set of similarity matrices is closed under permutation and under restriction to subsets, it is natural to begin by studying exchangeable similarity matrices of indefinite size, i.e. exchangeable random matrices. In order to build probability distributions, additional structure is needed in the form of a σ-algebra of subsets in $\mathcal{RT}_n$ or $\mathcal{UT}_n$, such that restriction and permutation are measurable, and Borel sets serve this purpose.

A probability distribution P_n on $\mathcal{RT}_n$ is finitely exchangeable if, for each permutation $\sigma\colon [n] \to [n]$, the random matrix $T \sim P_n$ has the same distribution as the σ-permuted matrix whose (i, j) element is $T_{\sigma(i),\sigma(j)}$. Exchangeability is founded on the principle of egalitarianism, which is commendable in the absence of more appealing alternatives, and occasionally survives superficial scrutiny in applied work. For almost all statistical applications where a reasonable justification can be produced for finite exchangeability, infinite exchangeability is equally compelling, possibly more so. Infinite exchangeability has one great advantage for statistical applications: it implies the existence of units (subjects, patients, plots,..) beyond those that happen to occur in the sample. To insist on finite exchangeability without an extension to larger sets is to defeat the purpose of inference by denying the existence of further units. This point of view may seem excessively dogmatic, but it is tempered by the fact that, in practice, pragmatism always trumps principle.

An infinitely exchangeable similarity matrix is, in effect, an infinite random matrix whose restriction to $[n]$ is a random matrix in $\mathcal{RT}_n$ with distribution P_n. For each n, the distribution is unaffected by permutation, i.e. finitely exchangeable. In addition, P_n is the marginal distribution of P_{n+1} under deletion of the last row and column, or in fact any row and the same column.

This section describes a range of infinitely exchangeable similarity matrices and provides tools for the construction of others.

7.2 Exchangeable partitions

A partition E of $[n]$ is a set of disjoint non-empty subsets whose union is $[n]$. The number of blocks or subsets is denoted by $\#E$, and for each block $b \in E$, i.e. $b \subset [n]$, the number of elements is $\#b$. Equivalently, a partition is a symmetric binary matrix or Boolean function $E\colon [n] \times [n] \to \{0, 1\}$ that is also reflexive and transitive. The set $\mathcal{E}_n$ of equivalence relations on $[n]$ is called the partition lattice; it is a finite subset of $\mathcal{RT}_n$, closed under permutation

and restriction. Thus, any infinitely exchangeable random partition is also an infinitely exchangeable random tree, albeit of a rather special type.

Finite exchangeability implies that the probability assigned by P_n to $E \in \mathcal{E}_n$ depends only on the group orbit. Since the block sizes are maximal invariant, two partitions having the same block sizes also have the same probability. Infinite exchangeability implies in addition that

$$\begin{aligned} P_2(12) &= P_3(123) + P_3(12|3) \\ P_2(1|2) &= P_3(13|2) + P_3(1|23) + P_3(1|2|3) \\ P_3(13|2) &= P_4(134|2) + P_4(13|24) + P_4(13|2|4) \\ P_4(12|3|4) &= P_5(125|3|4) + P_5(12|35|4) + P_5(12|3|45) + P_5(12|3|4|5) \end{aligned}$$

and so on, which is a substantial restriction. Blocks are not ordered, so $12|3 = 3|21$ is an abbreviation for $\{\{12\},\{3\}\}$. This is the Kolmogorov consistency property for an infinite random partition (Pitman 2005, section 2.2).

Kingman (1978) gives a characterization of infinitely exchangeable random partitions. Pitman (2005, section 4.3) characterizes the class of exchangeable Gibbs distributions, the Ewens-Pitman class. Hartigan's (1990) product partition models are finitely exchangeable if the cohesion function depends only on the block size, and the set of finitely exchangeable product partition models is a proper subset of the finitely exchangeable Gibbs models. The Ewens distribution (11) with parameter $\lambda > 0$ is a product partition model with cohesion function $\lambda(\#b-1)!$ for blocks of size $\#b$. Among finitely exchangeable product partition distributions, only the Ewens distributions are infinitely exchangeable. However, there are exchangeable Gibbs-type partitions other than Ewens.

Exchangeable random partitions are easy to construct from exchangeable sequences, but only in exceptional cases does this construction yield explicit non-trivial distributions. Let $Y_1, Y_2, \ldots$ be an infinitely exchangeable sequence, and let E be the block factor defined by

$$E_{ij} = \begin{cases} 1 & \text{if } Y_i = Y_j \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Then E is an infinitely exchangeable random partition. In fact, every infinitely exchangeable partition can be generated from an exchangeable sequence in this way (Kingman 1978). Note that if $g\colon \mathcal{R} \to \mathcal{R}$ is Borel measurable and invertible, the sequence $Y' = gY$ transformed component-wise, is exchangeable and the partition E is unaffected.

The Ewens partition (Ewens, 1972) is the archetype of an exchangeable random partition. For $\lambda > 0$, the distribution on $\mathcal{E}_n$ is given by

$$P_n(E) = \frac{\lambda^{\#E}\Gamma(\lambda)}{\Gamma(n+\lambda)} \prod_{b \in E} \Gamma(\#b). \tag{11}$$

The process is generated from the sequence in which $Y_1, Y_2, \ldots$ are conditionally independent with distribution generated from a Dirichlet process. The set

of distributions constitutes an exponential family with canonical statistic $\#E$, canonical parameter $\theta = \log \lambda$, and cumulant function $\log \Gamma(n + e^\theta) - \log \Gamma(e^\theta)$.

The Chinese restaurant process (Pitman, 2005, chapter 3) is an elegant sequential method for simulating the Ewens process directly. After k customers are seated in a configuration E with $\#b$ diners at table $b \in E$, the next customer sits at table b' with probability $\#b'/(k + \lambda)$, or at an unoccupied table with probability $\lambda/(k+\lambda)$. This is the conditional distribution $P_{k+1}(\cdot \mid E)$ given the current configuration, so the partition distribution after n customers are seated is (11). The first two customers are seated at the same table with probability $1/(1 + \lambda)$, and by exchangeability, the same is true for each pair. The process is not ergodic, but the blocks have limiting relative frequencies, and the number of blocks or occupied tables is asymptotically Poisson with mean $\lambda \log(n)$.

If the block sizes are written in integer-partition form $1^{m_1} 2^{m_2} \cdots n^{m_n}$, with m_r tables having exactly r diners, the multiplicity vector m has distribution

$$P_n(m) = \frac{n!\,\Gamma(\lambda)}{\Gamma(n + \lambda)} \prod_{j=1}^{n} \frac{\lambda^{m_j}}{j^{m_j}\, m_j!}$$

as if the components were independent Poisson counts $m_j \sim \mathrm{Po}(\lambda/j)$ conditional on $\sum j m_j = n$. This is the log series limit of the negative binomial distribution derived by Fisher, Corbet and Williams (1943) as a model for insect counts. The aim is to estimate the number of unseen species or to predict the number of additional species that will occur in an additional sample of size n'. The same negative binomial model was subsequently used by Efron and Thisted (1976) to estimate to estimate Shakespeare's vocabulary. Such models often give infinite point estimates.

7.3 Derived trees

Let $E_0, E_1, \ldots, E_k$ be independent random partitions, all infinitely exchangeable but not necessarily identically distributed. For example, we might well choose $E_0 = 1$ to be the one-block partition and E_k to be the complete partition into singletons. Then the random linear combination

$$V = \sigma_0^2 E_0 + \sigma_1^2 (E_0 E_1) + \cdots + \sigma_k^2 (E_0 \cdots E_k),$$

with non-negative coefficients independent of the partitions, is an exchangeable random tree. In this context, the product $E_0 E_1 = E_0 \wedge E_1$ is the greatest lower bound or component-wise minimum of the two partitions, not the matrix product.

In computational work, it is more economical to represent a partition in $\mathcal{E}_n$ as a list Y of levels with arbitrarily labelled blocks than as the block-factor matrix E. In the computational version of Chinese restaurant process, tables are most conveniently labelled in order of occupation, so the ith customer is seated at table number $Y_i \leq i$. Although the sequence Y is not exchangeable, it

can be made so by an independent uniform random permutation of block labels. However, this step is unnecessary because the block-factor transformation (10) discards the labels and generates the same partition however the levels are labelled. In practice, the matrix E is seldom explicitly computed. Let Y' be a second list generated in a similar manner with partition E'. If Y and Y' are regarded as two factors with interaction $Y * Y'$ in the standard sense of model formulae in statistical computing, the component-wise minimum partition $E \wedge E'$ is just the block factor generated by the ordered pairs $Y * Y'$. Thus, all of the preceding operations are elementary and need not involve matrices.

7.4 Infinite divisibility

Let P be a probability distribution on the set of rooted trees, and let $T_1, T_2, \ldots$ be independent trees with distribution P. For each integer $k \geq 0$ let $P^{\star k}$ be the distribution of the product $T_1 \cdots T_k$. The word product is used here in the semi-group sense of component-wise minimum rather than matrix product, so the operation is commutative and the product is a random tree. In particular, $P^{\star 1} = P$, and $P^{\star 0}$ puts unit mass on the semi-group identity element, the matrix with all components infinite. A distribution Q is said to be infinitely divisible if, for each integer $k \geq 1$, there exists a distribution P such that $Q = P^{\star k}$. In other words, the random tree $V \sim Q$ is equal in distribution to the k-fold product $T_1 \cdots T_k$ of independent trees each distributed as P.

For an upper tree interval $[x, \infty]$ consisting of trees $T \geq x$ component-wise, the minimum belongs to $[x, \infty]$ if and only if each component belongs to the interval, so $P^{\star k}([x, \infty]) = P^k([x, \infty])$. A distribution Q is infinitely divisible if and only if the function H_λ defined for $\lambda > 0$ by $H_\lambda(x) = Q^\lambda([x, \infty])$ determines a probability distribution on trees. If Q is infinitely divisible, it is convenient to define the family of distributions Q_λ for $\lambda > 0$ by $Q_\lambda([x, \infty]) = Q^\lambda([x, \infty])$, so that $Q_\lambda \star Q_{\lambda'} = Q_{\lambda+\lambda'}$. Every probability distribution on the real line is infinitely divisible in this sense, but the same is not true for trees. For example, the Ewens distribution with $\lambda = 1$ is not infinitely divisible for $n = 3$, and hence not infinitely divisible for any $n \geq 3$. The same appears to be true for all $\lambda > 0$.

Poisson mixtures are used to generate an infinitely divisible distribution Q from an arbitrary distribution P as follows:

$$Q_\lambda = \sum_{k=0}^{\infty} e^{-\lambda} \lambda^k P^{\star k} / k! \tag{12}$$

for $\lambda > 0$. In other words, first generate k from the Poisson distribution with mean λ, and then $V \sim Q$ is equal to the k-fold product $T_1 \cdots T_k$ of independent copies of $T \sim P$. A simple calculation shows that $Q_\lambda = Q_{\lambda/k}^{\star k}$, so Q_λ is infinitely divisible. It is also evident that only Poisson mixtures are infinitely divisible.

The preceding remarks apply to all sub-semi-groups such as the set $\mathcal{E} \subset \mathcal{RT}$ of partitions. However, the identity element in $\mathcal{E}$ is the maximal partition with one block. Since the semi-group product commutes with permutation and restriction, the distributions $P_n^{\star k}$ for fixed k determine an exchangeable

proccess when P_n is an exchangeable process. Likewise, the Poisson mixture is also infinitely exchangeable.

7.5 Convolution semi-group

Let $T \sim P_n$ and $T' \sim P'_n$ be independent rooted n-trees, and let $P_n \star P'_n$ be the distribution of the product $TT' = T \wedge T'$. In this way, the semi-group product $\wedge$ on trees induces a homomorphic semi-group operation $\star$ on distributions. Both semi-groups are commutative. Furthermore, the convolution operation $P_n \star P'_n$ preserves exchangeability, infinite divisibility and Poisson mixtures. The set of exchangeable distributions, the set of infinitely divisible distributions, and the set of Poisson mixtures are all sub-semi-groups. The set of exchangeable distributions is convex, i.e. closed under mixtures, but the set of infinitely divisible distributions is not.

7.6 Lévy fragmentations

A non-increasing partition sequence $(E_t)_{t\geq 0}$ with $E_t \in \mathcal{E}_n$ determines an n-tree by the relation $T_{ij} = \inf\{t : E_t(i,j) = 0\}$, which is the first separation time, or fragmentation time, for the two elements. Such trees are infinite on the diagonal $T_{ii} = \infty$, meaning that particles survive indefinitely. A Lévy fragmentation tree is a non-increasing partition-valued process in continuous time, exchangeable with independent multiplicative increments in non-overlapping temporal intervals. Such processes are constructed as follows.

Let $Q_{n,\lambda}$ be the distribution on $\mathcal{E}_n$ of an infinitely divisible and infinitely exchangeable random partition such that $Q_{n,\lambda} \star Q_{n,\lambda'} = Q_{n,\lambda+\lambda'}$ for all $\lambda, \lambda' > 0$. In other words, $E_t \sim Q_{n,t}$ is distributed as the k-fold product of independent partitions with distribution $Q_{n,t/k}$, so the distributions are consistent with a partition process evolving by independent multiplicative increments in non-overlapping temporal intervals. Consistency for different values of n ensures the existence of an extension to an infinite set, an infinite random partition evolving as a Markov process by successive splitting of blocks.

The process restricted to $[n]$ is evidently Markov with stationary multiplicative transition distributions determined by $Q_{n,t'-t}$ as follows. Given that the process is in state E at time t, the probabiity of being in state E' at time $t' \geq t$ is $Q_{n,t'-t}(E^{-1}E')$, where

$$E^{-1}E' = \{X \in \mathcal{E}_n \mid XE = E'\}$$

is the set of partitions X whose greatest lower bound $X \wedge E$ is E'.

Consider now the evolution of a Lévy fragmentation process beginning from the standard configuration $E_0 = 1$ with n particles as a single block at time zero. The reproductive property of the distributions $Q_{n,t}$ implies that the probability of the lattice interval $[E, 1]$ satisfies $Q_{n,t}([E,1]) = Q^t_{n,1}([E,1]) = \exp(-t\varphi_n(E))$. Thus, the first fragmentation occurs at a random time exponentially distributed with intensity parameter $\varphi_n(1)$. The larger the value of n, the shorter the

waiting time, so $\varphi_1(1) = 0$ and the fragmentation intensity increases with n. The state distribution immediately following the initial fragmentation, is the limit $\lim_{\lambda\to 0} Q_{n,\lambda}(\cdot \mid E \neq 1)$, i.e. $P(\cdot \mid E \neq 1)$ in (12). This limit also determines the block sizes, and exchangeability implies that the distribution depends only on the block sizes. Given the state E_1 following the initial fragmentation, the waiting time for a change to a lower state $E_2 = E_1 E_2'$ is exponential with parameter $\varphi_n(E_1)$, and E_2' is distributed as the conditional limit $\lim_{\lambda\to 0} Q_{n,\lambda}(\cdot \mid E \notin [E_1, 1])$. The process continues in this way with E_1 replaced by E_2 until an absorbing state E is reached such that $\varphi_n(E) = 0$.

As an example, suppose that Q is given by the Poisson mixture (12) in which P is the Ewens process with parameter $\theta > 0$. Then $P_n(1) = \Gamma(n)\Gamma(1+\theta)/\Gamma(n+\theta)$ and $\varphi_n = 1 - P_n(1)$ is the fragmentation intensity. The initial fragmentation distribution is the Ewens distribution conditional on $E < 1$, so the distribution of the number of branches and their sizes is governed by θ. If we now let $\theta \to 0$ and re-scale time accordingly, the limit intensity is $\varphi_n = \sum_{r=1}^{n-1} 1/r$ and the limiting fragmentation distribution has two blocks of sizes $(r, n-r)$ with probability $(1/r+1/(n-r))/(2\varphi_n)$. The limiting distribution $Q_{n,\lambda}$ is determined by $Q_{n,\lambda}(1) = \exp(-\lambda\varphi_n)$, and, for $E < 1$ by

$$\log Q_{n,\lambda}([E,1]) = -\lambda\varphi_n + \lambda \sum_{\substack{E' \geq E \\ \#E'=2}} \frac{(b_1-1)!(b_2-1)!}{(n-1)!},$$

where b_1, b_2 are the block sizes of E'. The probabilities are shown below by partition type for $n = 3, 4$.

		4	$e^{-11\lambda/6}$
3	$e^{-3\lambda/2}$	2^2	$e^{-10\lambda/6} - e^{-11\lambda/6}$
12	$e^{-\lambda} - e^{-3\lambda/2}$	13	$e^{-9\lambda/6} - e^{-11\lambda/6}$
1^3	$1 - 3e^{-\lambda} + 2e^{-3\lambda/2}$	$1^2 2$	$e^{-\lambda} - 2e^{-9\lambda/6} - e^{-10\lambda/6} + 2e^{-11\lambda/6}$
		1^4	$1 - 6e^{-\lambda} + 8e^{-9\lambda/6} + 3e^{-10\lambda/6} - 6e^{-11\lambda/6}$

Although it is by no means obvious from these formulae, the probabilities are non-negative for all $\lambda > 0$, and $Q_{3,\lambda}$ is the marginal distribution of $Q_{4,\lambda}$.

7.7 Homogeneous Markov fragmentations

In the Lévy-type construction, the tree increments are independent and identically distributed regardless of the current state. A different sort of fragmentation tree, arguably more natural, is obtained by allowing the increment distributions to depend on the current state, but not on time. In a homogeneous fragmentation tree, all branches following a fragmentation event evolve independently in a way that is statistically similar to the entire tree. Unlike a Lévy fragmentation tree, splits do not occur simultaneously on disjoint branches.

The Markov fragmentation beginning with a single block evolves in the same way as the Lévy fragmentation up to the first fragmentation event. Suppose that the first fragmentation results in a partition E having $k \geq 2$ blocks or branches.

Then the subsequent evolution on branch b is governed by $Q_{\#b,\lambda}$, independently on each branch until the next fragmentation event. Each fragmentation initiates a renewal, and the process continues independently on all branches. The waiting time for fragmentation on a branch of size n is exponential with intensity or rate $-\log Q_{n,\lambda}(1) = \lambda(1 - P_n(1))$, and the fragmentation distribution is $P_n(E)/(1 - P_n(1))$ for partitions $E < 1$.

The class of homogeneous Markov fragmentation trees has been characterized by Bertoin (2001) in terms of an exchangeable fragmentation intensity measure $\kappa = \lambda P$, which governs both the splitting intensity and the partition distribution. The connection between the splitting rule and the intensity measure, also called the dislocation measure, is well described in proposition 3 of Haas, Miermont, Pitman and Winkel (2006). These authors also show that the set of sampling-consistent Markov binary splitting rules coincides with Aldous's (1996) beta-splitting models. The distribution described at the end of the preceding section corresponds to Aldous's beta-splitting model with $\beta = -1$. The fragmentation intensity on a branch of size n is proportional to $\varphi_n = \sum_{j=1}^{n-1} 1/j$, and the block sizes after fragmentation are $(r, n-r)$ with probability $\big(1/r + 1/(n-r)\big)/(2\varphi(n))$.

If necessary, particle deaths can be incorporated to make $T(i,i)$ finite. One option is monotone transformation putting all leaves at the same height. Another option is to define X to be an iid exponential sequence independent of T, so that $(X \wedge X)_{ij} = \min(X_i, X_j)$ is a finite random tree, exchangeable and infinitely divisible. Then the product $T \cdot (X \wedge X)$ is a finite tree, also exchangeable.

7.8 Markovian tree processes

The construction of section 7.6 can also be used to generate a Markovian tree-valued process $(T_t)_{t \geq 0}$ in continuous time with stationary independent multiplicative increments. Simply take $Q_{n,t}$ to be infinitely exchangeable and infinitely divisible on trees rather than partitions. Such trees are necessarily decreasing in time, though they need not decrease to zero. It is not at all clear what sort of physical or biological process might be represented by such a process.

8 Acknowledgements

I am grateful to Jim Pitman and Mathias Winkel for references and helpful comments, especially on section 7.

References

[1] Aldous, D. (1966) Probability distributions on cladograms. In Random Discrete Structures (eds. D. Aldous and R. Pemantle). Springer, New York, 1-18.

[2] Bertoin, J. (2001) Homogeneous fragmentation processes. *Probab. Theory and Related Fields* **121**, 301-318.

[3] Blei, D., Ng, A., Jordan, M. (2003) Latent Dirichlet allocation. *J. Machine learning Research* **3** 993-1022.

[4] Cressie, N., Frey, J., Harch, B. and Smith, M. (2006) Spatial prediction on a river network. *J. Agricultural, Biological and Environmental Statistics* (to appear).

[5] Ehrenberg, A.S.C. (1981) The problem of numeracy. *The American Statistician* **35** 67-71.

[6] Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3** 87-112.

[7] Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika*, 63, 435-447.

[8] Hartigan, J. A. (1990) Partition models. *Communications in Statistics* **19** 2745-2756.

[9] 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.

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

[11] Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12, 42-58.

[12] Haas, B., Miermont, G., Pitman, J. and Winkel, M. (2006) Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. arXiv:math.PR/0604350v1

[13] Kingman, J. F. C. (1978). Random partitions in population genetics. *Proceedings of the Royal Society of London: Series A (Mathematical and Physical Sciences)*, 361, 1-20.

[14] Kingman, J.F.C. (1980) Mathematics of Genetic Diversity. SIAM, Philadelphia.

[15] Kingman, J.F.C. (1982) On the genealogy of large populations. *Journal of Applied Probability* 19, 27-43.

[16] Kingman, J.F.C. (1982b) The coalescent. *Stochastic processes and their applications* 13, 235-248.

[17] Kruskal, J.B. (1964) Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika* 29, 1–27.

[18] McCullagh, P. (2005) Exchangeability and regression models. In Celebrating Statistics: Papers in honour of Sir David Cox on his 80th birthday A.C. Davison, Y. Dodge and N. Wermuth, editors. Oxford Statistical Science Series No. 33.

[19] Pitman, J. (2005) Combinatorial Stochastic Processes. Springer.

[20] Shepard, R.N. (1962a) The analysis of proximities: multidimensional scaling with an unknown distance function I. *Psychometrika* 27, 125–139.

[21] Shepard, R.N. (1962b) The analysis of proximities: multidimensional scaling with an unknown distance function II. *Psychometrika* 27, 219–246.

[22] Torgersen, W.S. (1958) Theory and Methods of Scaling. New York, Wiley.