Chapter 10

Multidimensional Scaling and Distance Geometry

10.1 The Data Matrix

Consider p variables, and a *random sample* $\underline{x}_1, \ldots, \underline{x}_n$, where $\underline{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})^t$. Each observation is a p vector, and there are n observations. A *random sample* means that $\underline{x}_1, \ldots, \underline{x}_n$ is an observation of $\underline{X}_1, \ldots, \underline{X}_n$, where the $(\underline{X}_j)_{j=1}^n$ are independent, identically distributed random p-vectors.

Notation A random p-vector, where each component corresponds to a different variable, is usually taken as a *column* vector, but when presented in a data matrix of n independent observations, the transpose is taken and each p-variate observation is taken as a *row*. .

Sampling If the observations were selected from a total population of N p-vectors, then a *random sample* would mean that any subset of n vectors from N was chosen with probability $\frac{1}{\begin{pmatrix} n \\ N \end{pmatrix}}$ and each ordering of the n vectors occurred with probability $\frac{1}{n!}$. In general, a *random sample* is a sample that has the properties of such a sample for $N >> n$.

The most widely used standard is to store the data in an $n \times p$ matrix, denoted $\mathbf{x}$, where

$$\mathbf{x} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \underline{x}_1^t \\ \underline{x}_2^t \\ \vdots \\ \underline{x}_n^t \end{pmatrix}. \tag{10.1}$$

10.2 One Way Representations of Data Matrices: Andrews Curves

When considering a one way representation of a two dimensional data matrix, one can represent either the n units, or the p variables. Each variable, may be represented by an appropriate curve or solid pattern that highlights the similarities or disssimilarities between the constructions.

One example is the method of *Andrews Curves*. For each unit (or p-variate observation) i of the data matrix, set

$$f_i(t) = \frac{1}{\sqrt{2}} x_{i1} + \sum_{j=1}^{[p/2]} x_{i,2j} \sin(jt) + \sum_{j=1}^{[p/2]} x_{i,2j+1} \cos(jt) \quad t \in [-\pi, \pi].$$

Properties The Andrews curve satisfies the following properties:

1. Let $\overline{f} = \frac{1}{n} \sum_{i=1}^{n} f_i(t)$, then

$$\overline{f}(t) = \frac{1}{\sqrt{2}} \overline{x}_{.1} + \sum_{j=1}^{[p/2]} \overline{x}_{.2j} \sin(jt) + \sum_{j=1}^{[p/2]} \overline{x}_{.,2j+1} \cos(jt) \quad t \in [-\pi, \pi].$$

2. This function representation preserves the Euclidean distance between the variables. That is, if

$$d_{ij}^2 = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2$$

then

$$\frac{1}{\pi} \int_{-\pi}^{\pi} (f_i(t) - f_j(t))^2 dt = d_{ij}^2.$$

3. Suppose $(X_1, \ldots, X_p)$ are independent variables, each with variance σ^2, then for each i,

$$\mathrm{Var}(f_i(t)) = \begin{cases} \frac{\sigma^2}{2} p & p \quad \text{odd} \\ \frac{\sigma^2}{2}(p-1) + \sigma^2 \cos^2(pt) & p \qquad \text{even.} \end{cases}$$

The following features should be noted:

- An outlier appears as single Andrews' curves that looks different from the rest.
- A subgroup of data is characterised by a set of simular curves.
- The order of the variables plays an important role for interpretation.
- For more than 20 observations we may obtain a bad "signal-to-ink-ratio", i.e., too many curves are overlaid in one picture.

10.3 Subspace Projections

The data matrix $\mathbf{x}$ described by Equation (10.1) of p quantitative measurements on n units may be described either in the *object space* or the *variable space*, as defined below.

Definition 10.1 (Object Space)**.** *Let $\underline{\overline{x}} = \frac{1}{n}\sum_{j=1}^{n} \underline{x}_j$ denote the sample mean vector. The* object space *is the p dimensional space with origin at $\underline{\overline{x}}$.*

Multivariate analysis studies how the variables relate to each other; their covariance and correlation. Centralising around the sample average helps to keep this in view. When studying object space, n points in $\mathbb{R}^p$ are considered, labelled $(\underline{y})_{j=1}^{n}$, where $\underline{y}_j = \underline{x}_j - \underline{\overline{x}}$. The *distance* between unit i and j in object space is given by the Euclidean distance;

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(y_{ik} - y_{jk})^2}.$$

Definition 10.2 (Variable Space)**.** *Let $\overline{x}_{.k} = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$, the sample average for variable k. Consider the vectors $\underline{z}_k = \underline{x}_k - \overline{x}_{.k}\mathbf{1} \in \mathbb{R}^n$, $k = 1, \ldots, p$, where $\mathbf{1} = (1, \ldots, 1)^t \in \mathbb{R}^n$. These vectors are all perpendicular to $\mathbf{1}$. The* variable space *is defined as the space spanned by these vectors. The variable space is therefore a space of dimension less than or equal to p, embedded in the $n-1$ dimensional subspace of $\mathbb{R}^n$ perpendicular to the vector $\mathbf{1}$.*

In the *variable space*, the scalar product c_{kl} between $\underline{z}_k$ and $\underline{z}_l$ is given by

$$c_{kl} = \sum_{i=1}^{n} z_{ik} z_{il}.$$

The quantity $s_{kl} = \frac{1}{n-1}c_{kl}$ is defined as the sample covariance between variate k and variate l. In the exercises, it is proved that this is an unbiased estimator of the population covariance. The *sample correlation* between these variables is defined as

$$\cos(\alpha_{kl}) = r_{kl} := \frac{c_{kl}}{\sqrt{c_{kk}c_{ll}}},$$

where α_{kl} is the *angle* between vector $\underline{z}_k$ and $\underline{z}_l$.

Note: It should be clear (Exercise **??** Page **??**) that the projection of the vector $\underline{z}_k$ onto the one dimensional subspace of $\mathbb{R}^n$ spanned by the vector $\underline{z}_l$ is simply the linear regression of $\underline{z}_k$ onto $\underline{z}_l$;

$$r_{kl}\sqrt{\frac{c_{kk}}{c_{ll}}}\underline{z}_l.$$

Definition 10.3. *Set $S_{kl} = \frac{1}{n-1}c_{kl}$. The matrix S is the* sample covariance *matrix of the data matrix $\mathbf{X}$. The matrix R with entries r_{kl} is the* sample correlation *matrix.*

Remark When the n observations are considered in object space, their respective distances from each other may be represented by the $n \times n$ matrix $(d_{ij})_{1\leq i\leq n, 1\leq j\leq n}$. Considered in *variable space*, the observations lead to the $p \times p$ correlation and covariance matrices and the $p \times p$ matrix of angles α_{kl}, all representing the similarity between the variables.

Lemma 10.4. *The matrices S and R are non-negative definite.*

Proof Consider any p vector $\underline{a}$. Then

$$\underline{a}^t S \underline{a} = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{kl} a_k a_l z_{ik} z_{il} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\sum_{k} a_k z_{ik} \right)^2 \geq 0.$$

$$\underline{a}^t R \underline{a} = \sum_{i=1}^{n} a_k a_l \frac{z_{ik}}{\sqrt{c_{kk}}} \frac{z_{il}}{\sqrt{c_{ll}}} = \sum_{i=1}^{n} \left(\sum_{k=1}^{p} \frac{z_{ik} a_k}{\sqrt{c_{kk}}} \right)^2 \geq 0.$$

□

10.4 Distances and Proximity Matrices

When the p variables are numerical and observations of continuous random variables, the *distance* between unit i and j in object space may be given by the Euclidean distance;

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (y_{ik} - y_{jk})^2}.$$

Data sets often also give information in the form of *categorical* variables and it is useful to be able to incorporate both numerical and categorical variables. Also, there is a common problem of *missing data*; for an observation i, the datum x_{ik} may be missing for some, but not all, values of k.

The following measure of distance between observations is known as *Gower's dissimilarity*: Let

$$\delta_{ijk} = \begin{cases} 1 & x_{ik}, x_{jk} \quad \text{can be compared} \\ 0 & \text{otherwise} \end{cases}$$

$$s_{ijk} = 0 \qquad \text{if} \qquad \delta_{ijk} = 0.$$

If either x_{ik} or x_{jk} are missing, then both $\delta_{ijk} = 0$ and $s_{ijk} = 0$. For $\delta_{ijk} = 1$, let

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{\max_{a,b} |x_{ak} - x_{bk}|}$$

if variable k is a quantitive variable and

$$s_{ijk} = \begin{cases} 1 & x_{ik} = x_{jk} \\ 0 & \text{otherwise} \end{cases}$$

if variable k is categorical. Gower then constructs a *distance* by:

$$d_{ij} = \sum_k \frac{s_{ijk}\delta_{ijk}}{\sum_k \delta_{ijk}}.$$

If greater weight is attached to some of the variables, this can be modified using weights;

$$d_{ij} = \sum_k \frac{w_k s_{ijk}\delta_{ijk}}{\sum_k w_k \delta_{ijk}}.$$

Constructing a 'Virtual' data set from distances There are situations where the data matrix $\mathbf{x}$ is not given, but instead the distance matrix (d_{ij}) is given. The following discussion describes how to construct a virtual data matrix $\mathbf{x}$, which preserves the correct distances.

Let $\mathbf{x}$ be an $n \times p$ data matrix with entries x_{ij} and let $H_n = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t$ where $\mathbf{1}$ denotes an n-vector where each entry is 1. Then it is an easy computation to see that

$$(H_n\mathbf{x})_{ij} = x_{ij} - \overline{x}_{.j}.$$

Set

$$Q = H_n\mathbf{x}(H_n\mathbf{x})^t,$$

then it is clear that

$$Q_{ij} = \sum_{k=1}^{p}(x_{ik} - \overline{x}_{.k})(x_{jk} - \overline{x}_{.k}).$$

Set $y_{ij} = x_{ij} - \overline{x}_{.j}$. If the distance d_{ij} is the Euclidean distance, then

$$\begin{aligned} d_{ij}^2 &= \sum_k (y_{ik} - y_{jk})^2 \\ &= \sum_k y_{ik}^2 + \sum_k y_{jk}^2 - 2\sum_k y_{ik}y_{jk} \\ &= Q_{ii} + Q_{jj} - 2Q_{ij}. \end{aligned}$$

Note that $Q_{ij} = Q_{ji}$ and that $\sum_{i=1}^n Q_{ij} = 0$ for each j. It follows that

$$2n\sum_i Q_{ii} = \sum_{i,j} d_{ij}^2,$$

$$Q_{ii} = \frac{1}{n}\sum_k d_{ik}^2 - \frac{1}{2n^2}\sum_{i,j} d_{ij}^2$$

$$Q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_{k=1}^n d_{kj}^2 - \frac{1}{n}\sum_{k=1}^n d_{ik}^2 + \frac{1}{n^2}\sum_{ij} d_{ij}^2\right). \tag{10.2}$$

If the data matrix is not given, but instead the distances $(d_{ij})_{(i,j)\in\{1,\ldots,n\}}$, then a matrix Q may be constructed using the formula given by Equation (10.2). The matrix constructed in this way is clearly symmetric and can be diagonalised as

$$Q = P\Lambda P^t,$$

where P is orthonormal and Λ is diagonal. If the matrix $(d_{ij})_{(i,j)\in\{1,\ldots,n\}^2}$ is a distance, in the sense that it is symmetric, the entries are non negative and $d_{ij} \leq d_{im} + d_{mj}$ for all $(i,j,m) \in \{1,\ldots,n\}^3$, then $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_n)$, $\lambda_1 \geq \ldots \geq \lambda_n = 0$. Let $\sqrt{\lambda_j}$ denote the positive square root of λ_j and let $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}\ldots,\sqrt{\lambda_n})$.

Definition 10.5 (Data Matrix obtained by Metric Scaling)**.** *Let*

$$\mathbf{x} = P\Lambda^{1/2},$$

then $\mathbf{x}$ *is the data matrix corresponding to* (d_{ij}) *obtained by metric scaling.*

Recall that the situation considered here is where the original data is not given; rather, the analyst has been presented with a matrix of distances between the original data points. The 'data matrix' obtained in this manner will preserve the distances between the original data.

Remarks

1. Metric scaling only works if the matrix Q is non negative definite (i.e. positive semi definite). This holds if and only if the input matrix (d_{ij}) satisfies the triangle inequality;

$$d_{ij} \leq d_{im} + d_{mj} \qquad \forall (i,j,k) \in \{1,\ldots,n\}^3.$$

2. By construction, the data matrix $\mathbf{x}$ obtained in this way is already centred; $\mathbf{x} = H\mathbf{x}$.

3. Since $\mathbf{x} = H\mathbf{x}$, it follows that $\text{rank}(Q) \leq n-1$, at least one eigenvalue is zero. If

$$\frac{\lambda_1 + \ldots + \lambda_m}{\lambda_1 + \ldots + \lambda_{n-1}}$$

is sufficiently large for some $m < n-1$, then the data matrix can be constructed from the first m columns of P by taking $\mathbf{x}$ as the $n \times m$ matrix with entries $\mathbf{x}_{ij} = P_{ij}\sqrt{d_j}$ $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,m\}$.

10.5 Measuring and Testing Multivariate Distances

Often in multivariate analysis, the n observations are not an observed random sample from a single population, but rather come from m different populations. Often, the aim is *classification*; to decide, based on the p-variate observation, which population the observation belongs to.

Consider m populations (for example, 7 different types of dog), where p features (variables) are measured (for example, p different bones within the body may be considered and the length of each measured for each animal). Suppose that $n = n_1 + \ldots + n_m$, where n_b denotes the number of different animals from population b, for each population $b = 1, \ldots, m$. Let x_{abc} denote the observation: observation a, population b, variable c. Suppose you are given an observation, but you are not told which population the observation comes from. As a first step for making a guess, it is useful to have a measure of distance between the various populations.

10.5.1 Penrose and Mahalanobis Distance

Penrose Distance Let $n = \sum_{b=1}^{m} n_b$ denote the total number of observations and let

$$s_c^2 = \frac{\sum_{b=1}^{m}(n_b - 1)s_{bc}^2}{n - m}.$$

The observed *Penrose distance* between two populations α and β is defined as

$$p_{\alpha,\beta} = \frac{1}{p}\sum_{k=1}^{p}\frac{(\bar{x}_{.,\alpha,k} - \bar{x}_{.,\beta,k})^2}{s_k^2}.$$

Formal tests, of whether or not an observed Penrose distance is significantly different from zero, may be carried out under distributional assumtions. If it is assumed that the the observations x_{abc} are from *independent* random variables X_{abc}, where

$$X_{abc} \sim N(\mu_{bc}, \sigma_c^2)$$

(that is, the variables are normal and for variate c, the population variance is the same for each population $b = 1, \ldots, m$), then the distribution of

$$P_{\alpha,\beta} = \frac{1}{p}\sum_{k=1}^{p}\frac{(\overline{X}_{.,\alpha,k} - \overline{X}_{.,\beta,k})^2}{S_k^2}$$

under the null hypothesis that $\underline{\mu}_{\alpha,.} = \underline{\mu}_{\beta,.}$ may be computed.

The Mahalanobis Distance The Penrose distance does not take into account correlations between the variables. The Mahalanobis distance is a modification of the Penrose distance that takes into account possible correlations. If the independence assumption holds, then the Penrose distance is better, because there are fewer parameters involved. Let $\underline{X}_a$ denote a random vector that models population a, with $\mathbb{E}[\underline{X}_a] = \underline{\mu}_a$ and $\mathbf{C}(\underline{X}_a) = C$ (the notation $\mathbf{C}$ is used to denote a covariance matrix), where C is the same for *each* population $a = 1, \ldots, m$. Let $\bar{x}_{aj}$ denote the jth component of the vector $\underline{\bar{x}}_a$, the sample average from population a. Let S denote the pooled estimate of the covariance matrix and let $V = S^{-1}$. The Mahalanobis distance between two populations α and β is *defined* as

$$D_{\alpha\beta} = \sum_{r=1}^{p}\sum_{s=1}^{p}(\bar{x}_{.,\alpha,r} - \bar{x}_{.,\beta,r})V_{rs}(\bar{x}_{.,\alpha,s} - \bar{x}_{.,s\beta}) = (\underline{\bar{x}}_\alpha - \underline{\bar{x}}_\beta)^t V(\underline{\bar{x}}_\alpha - \underline{\bar{x}}_\beta).$$

To test whether the sample Mahalanobis distance, computed from the sample means and sample covariance matrix is statistically significant, one uses Hotelling's T^2 distribution; under the null hypothesis (of no difference),

$$\frac{n_a + n_b - p - 1}{(n_a + n_b - 2)p}\frac{n_a n_b}{n_a + n_b}D_{ab} \sim F_{p,n_a+n_b-p-1}.$$

Note that there are $p(p+1)/2$ terms to be estimated in the covariance matrix for the Mahalanobis distance, while there are only p variances to be estimated for the Penrose distance. Therefore, if there is reason to believe that an independence assumption gives an accurate model, the Penrose distance is a better measure of distance; rather many observations are required to obtain the whole matrix S^{-1} with accuracy.

Example 10.6 (Egyptian Skull Data)**.**

The data set on Egyptian skulls, found in `skulls.dat` on the course home page gives the measurements X_1 = maximum breadth, X_2 = basibregmatic height, X_3 = basialveolar length and X_4 = nasal height. The data is for a total of 150 skulls, 30 from each of 5 groupings; -4000 Early Predynastic, -3300 Late Predynastic, -1850 12th and 13th Dynasties, -200 Ptolemaic Period, 150 Roman Period.

Firstly, the sample mean vector for $(X_1, X_2, X_3, X_4)^t$ is computed for each period, and the pooled covariance matrix. That is, firstly S_a, the sample covariance matrix for period a is computed for each of the 5 periods and then

$$S = \frac{\sum_{a=1}^{5} 29 S_a}{145}.$$

Here S is a 4×4 covariance matrix, with the sample variances along the diagonal.

The Penrose distances may now be computed directly; to compute the Mahalanobis distances, the inverse S^{-1} is required. These distances turn out to be:

Penrose

	I	II	III	IV	V
I	–				
II	0.023	–			
III	0.216	0.163	–		
IV	0.493	0.404	0.108	–	
V	0.736	0.583	0.244	0.066	–

Mahalanobis

	I	II	III	IV	V
I	–				
II	0.091	–			
III	0.903	0.729	–		
IV	1.881	1.594	0.443	–	
V	2.697	2.176	0.911	0.219	–

Due to the change of scale (the Penrose is divided by a $1/p$) it does not make sense to compare the *absolute* values of these distances, but the *ratios* should be comparable, giving the change between one group and another. The ratio of the $I \to II$ and $I \to V$ distance is $0.736/0.023 = 32.0$ for the Penrose and $2.697/0.091 = 29.6$ for the Mahalanobis measure; the results are similar. □

10.6 Classical Scaling and Distance Geometry

Suppose we have n points $X_1, \ldots, X_n \in \mathbb{R}^r$ and we compute an $n \times n$ proximity matrix Δ with entries

$$\delta_{ij} = \|X_i - X_j\|.$$

If Euclidean distances are used, then:

$$\delta_{ij}^2 = \|X_i\|^2 + \|X_j\|^2 - 2(X_i, X_j).$$

Let

$$b_{ij} = (X_i, X_j) = \frac{1}{2}(\delta_{ij}^2 - \delta_{i0}^2 - \delta_{j0}^2)$$

where $\delta_{i0}^2 := \|X_i\|^2$. Then, summing over i and j gives:

$$\frac{1}{n}\sum_i \delta_{ij}^2 = \frac{1}{n}\sum_i \delta_{i0}^2 + \delta_{j0}^2$$

$$\frac{1}{n}\sum_j \delta_{ij}^2 - \delta_{i0}^2 + \frac{1}{n}\sum_j \delta_{j0}^2$$

$$\frac{1}{n^2}\sum_{i,j} \delta_{ij}^2 = \frac{2}{n}\sum_i \delta_{i0}^2$$

and, letting $a_{ij} = -\frac{1}{2}\delta_{ij}^2$, we get:

$$b_{ij} = a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}.$$

The notation is:

$$a_{i.} = \frac{1}{n}\sum_j a_{ij}^2 \qquad a_{.j} = \frac{1}{n}\sum_i a_{ij}^2 \qquad a_{ii} = \frac{1}{n^2}\sum_{ij} a_{ij}^2.$$

Let A denote the matrix with entries a_{ij} and B matrix with entries b_{ij} then A and B are related through

$$B = HAH \qquad H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'.$$

The matrix B is a 'doubly centred' version of A.

MDS is about dimensionality reduction and we would like to find $Y_1, \ldots, Y_n \in \mathbb{R}^t$ where $t < r$ (referred to as the *principal co-ordinates*). When distances are Euclidean interpoint distances, this is the same as the PCA problem.

In typical classical scaling problems, we are not given the points $X_i \in \mathbb{R}^r$, but rather the proximity matrix Δ. Using Δ, we form A and then B. The idea is to find a matrix B^* with entries b_{ij}^* with rank at most t which minimises

$$\text{tr}((B - B^*)^2) = \sum_{ij}(b_{ij} - b_{ij}^*)^2$$

If $\lambda_1 \geq \ldots \geq \lambda_n$ are the eigenvalues of B, then it turns out that the eigenvalues of B^* are $\lambda_k^* = \max(\lambda_k, 0)$ for $k = 1, \ldots, t$ and 0 otherwise.

The classical scaling algorithm is based on an eigenvalue/vector decomposition of B which produces $Y_1, \ldots, Y_n \in \mathbb{R}^t$, a configuration whose Euclidean interpoint distances d_{ij} satisfy

$$d_{ij}^2 = \|Y_i - Y_j\|^2$$

The solution is not unique; any orthogonal transformation also gives a solution.

Assessing Dimensionality One way of doing this is to look at the eigenvalues of B.The usual strategy is to plot the ordered eigenvalues against dimension and then identify a dimension at which the eigenvalues become 'stable' (i.e. do not change perceptively).

10.6.1 Distance Scaling

Given n items and the $n \times n$ matrix of dissimilarities Δ with entries δ_{ij} we wish to find a function f such that

$$d_{ij} = f(\delta_{ij})$$

gives interpoint distances. The use of 'metric' or 'non-metric' distance scaling depends on the nature of the dissimilarities.

10.6.2 Metric distance scaling

For MDS, the function f is taken as:

$$f(\delta_{ij}) = \alpha + \beta\delta_{ij}$$

where α and β are unknown positive constants. We make the dimension reduction and get points $Y_1, \ldots, Y_n \in \mathbb{R}^t$ and compute d_{ij} the distances between them.We then compute the weighted loss function

$$L_f(Y_1, \ldots, Y_n; W) = \sum_{i<j} w_{ij}(d_{ij} - f(\delta_{ij})^2$$

the parameters α and β are chosen to minimise this. W is a given matrix of weights and the stress is defined as

$$\text{stress} = \sqrt{L_f(Y_1, \ldots, Y_n; W)}.$$

Sammon Mapping The Sammon mapping is a popular choice. Here

$$w_{ij} = \frac{1}{\delta_{ij} \sum_{k<l} \delta_{kl}}.$$

The Sammon mapping preserve the small δ_{ij} and gives them a greater emphasis than the larger δ_{ij}.

Bayesian MDS

We consider the situation where the entries of Δ are tained by measurement error. Let us assume that the measured dissimilarity $\delta_{ij} > 0$ is subject to a Gaussian error, so that

$$\delta_{ij} = \delta^0_{ij} + \epsilon_{ij}$$

where δ^0_{ij} is the true measurement error and $\epsilon_{ij} \sim N(0, \sigma^2)$ (independent of each other). Hence

$$\delta_{ij} \sim N(\delta^0_{ij}, \sigma^2)\mathbf{1}_{\{\delta_{ij}>0\}}.$$

The likelihood of $(\{X_i\}, \sigma^2)$ given Δ is therefore

$$\begin{aligned} L(\{X_i\}, \sigma^2|\Delta) &= \prod_{i<j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\delta_{ij} - \delta^0_{ij})^2}{2\sigma^2}\right\} \left\{1 - \Phi\left(\frac{\delta^0_{ij}}{\sigma}\right)\right\}^{-1} \\ &\propto (\sigma^2)^{-m/2} \exp\left\{-\frac{ESS}{2\sigma^2} - \sum_{i<j} \log\Phi\left(\frac{\delta^0_{ij}}{\sigma}\right)\right\} \end{aligned}$$

where $ESS = \sum_{i<j}(\delta_{ij} - \delta_{ij}^0)^2$ is the error sum of squares and $\Phi(.)$ is the standard Gaussian c.d.f. and $m = \frac{n(n-1)}{2}$, the number of dissimilarities. The second term is the modification of the likelihood due to the truncation.

Now we assume that $X_i \sim N(0, C_{XX})$ where $C_{XX} = \text{diag}(\lambda_1, \ldots, \lambda_r)$. Then the full conditional posterior is:

$$(\sigma^2)^{-m/2}\left(\prod_{j=1}^{r}\lambda_j^{-n/2}\right)\exp\left\{-\frac{Q_1+Q_2}{2} - \sum_{i<j}\log\Phi\left(\frac{\delta_{ij}^0}{\sigma}\right)\right\}$$

where $Q_1 = \frac{ESS}{\sigma^2}$, $Q_2 = \sum_{i=1}^{n}(X_i' C_{XX}^{-1} X_i) = \sum_{j=1}^{r}\frac{1}{\lambda_j}s_j$ are quadratic functions of the $\{X_i\}$ and $s_j = \sum_{i=1}^{n} X_{ij}^2$.

Now assume that the error variance σ^2 has conjugate prior

$$\sigma^2 \sim IG(a, b)$$

inverse Gamma with parameters a and b. That is

$$\pi(\sigma^2) \propto (\sigma^2)^{-(a+1)}e^{-b/\sigma^2} \qquad a, b > 0$$

and we take the prior for $\lambda_j \sim IG(\alpha, \beta_j)$ independently for each j. The joint posterior, given the observed proximity matrix is:

$$p(\{X_i\}, \{\lambda_j\}, \sigma^2|\Delta) \propto (\sigma^2)^{-((m/2)+a+a)}\left(\prod_{j=1}^{r}\lambda_j^{-((n/2)+\alpha+1)}\right)e^{-A}$$

$$A = \frac{Q_1+Q_2}{2} + \sum_{i<j}\log\Phi\left(\frac{\delta_{ij}^0}{\sigma}\right) + \frac{b}{\sigma^2} + \sum_{j=1}^{r}\frac{\beta_j}{\lambda_j}$$

The maximum posterior estimate may be computed using MCMC.

10.6.3 Non-metric Distance Scaling

In non-metric distance scaling, we assume that f is an arbitrary function that satisfies $f(x) \leq f(y)$ whenever $x < y$ for any pairs of dissimilarities x and y. The function f is chosen to preserve the rank of the dissimilarities. We find $Y_1, \ldots, Y_n \in \mathbb{R}^t$ where $t < r$ with distances

$$d_{ij} = \|Y_i - Y_j\|$$

such that the ordering of the distances (from lowest to highest) matches the ordering of the dissimilarities.

Motivation The motivation comes from psychological experiments, where respondents often give extreme answers (extremely good or extremely bad). The ranking is therefore important, but raw distances can give an exaggerated picture.

10.7 The Mantel Randomisation Test

The Mantel test (1967) was introduced to detect space / time clustering of diseases. Suppose that n objects are being studied and suppose that there are observations on two sets of observations. Let M be the $n \times n$ matrix where M_{ij} is the distance between object i and object j based on the first set of variables and let E be a matrix of distances between the objects based on the second set of variables. Mantel's test assesses whether or not the elements in M and E show some significant correlation. Let

$$Z = \sum_{j=2}^{n} \sum_{k=1}^{j-1} M_{jk} E_{jk}.$$

This is compared with observations

$$Z_\sigma = \sum_{j=2}^{n} \sum_{k=1}^{j-1} M_{\sigma(j)\sigma(k)} E_{jk},$$

where σ is a randomly chosen permutation of $(1, \ldots, n)$. The values z_σ are computed for *each* of the $n!$ permutations σ and then it is seen if Z is a 'typical' observation of this distribution (i.e. does it land between the $\frac{\alpha}{2} \times 100$ and $1 - \frac{\alpha}{2} \times 100$ percentiles of this empirical distribution, where α is the significance level?)

For the Egyptian skulls data, the n objects are the n different skulls. To perform a Mantel randomisation test, the two sets of variables are: Set 1 (on which M is based) are the measurements of the skulls and Set 2 (on which E is based) is the single variable, the period from which the skull comes.