

Chapter 2

Exploratory Analysis of Ranking Data

2.1 Descriptive Statistics

Descriptive statistics present an overall picture of ranking data. Not only do they provide a summary of the ranking data, but they are also often suggestive of the appropriate direction to analyze the data. Therefore, it is suggested that researchers consider descriptive analysis prior to any sophisticated data analysis.

We begin with a single measure of the popularity of an object. It is natural to use the mean rank attributed to an object to represent the central tendency of the ranks. The mean rank $\mathbf{m} = (m_1, \dots, m_t)'$ is defined as the t -dimensional vector in which the i th entry equals

$$m_i = \sum_{j=1}^{t!} n_j v_j(i)/n,$$

where $v_j, j = 1, 2, \dots, t!$ represents all possible rankings of the t objects, n_j is the observed frequency of ranking j , $n = \sum_{j=1}^{t!} n_j$, and $v_j(i)$ is the rank score given to object i in ranking j .

Apart from the mean ranks, the pairwise frequencies, that is, the frequency with which object i is more preferred (i.e., ranked higher with a smaller rank score) than object j , for every possible C_2^t object pairs (i, j) , are also often used. These pairwise frequencies can be summarized in a matrix called a *pair matrix* P in which the (a, b) th entry equals

$$P_{ab} = \sum_{j=1}^{t!} n_j I(v_j(a) < v_j(b)),$$

where $I(\cdot)$ is the indicator function. Note that P_{ab}/n represents the empirical probability that object a is more preferred than object b . In addition to mean ranks and pairwise frequencies, one can look more deeply into ranking data by studying the so-called “marginal” distribution of the objects. A *marginal matrix*, specifically for this use, is the $t \times t$ matrix M in which the (a, b) th entry equals

$$M_{ab} = \sum_{j=1}^{t!} n_j I[v_j(a) = b].$$

Note that M_{ab} is the frequency of object a being ranked b th. Marden (1995) called it a marginal matrix because the a th row gives the observed marginal distribution of the ranks assigned to object a and the b th column gives the marginal distribution of objects given the rank b .

Example 2.1. The function `destat` in the R package `pmr` computes three types of descriptive statistics of a ranking data set, namely mean ranks, pairs, and marginals. Here, we will use Sutton’s Leisure Time data (Table A.1) for illustration. The data set `leisure.black` in the `pmr` package contains the rank-order preference of spending leisure time with (1: male; 2: female; 3: both sexes) by 13 black women. By using the R code `destat(leisure.black)`, the function `destat` produces the following mean rank vector, pair matrix, and marginal matrix (Fig. 2.1):

From the above descriptive statistics, we can see that the object “3: both sexes” is clearly most preferred by the black females, and there is no strong preference between the other two objects.

```
> library("pmr")
> data(leisure.black)
> destat(leisure.black)

$mean.rank
[1] 2.308 2.462 1.231

$pair
  [,1] [,2] [,3]
[1,]  0   7   2
[2,]  6   0   1
[3,] 11  12   0

$mar
  [,1] [,2] [,3]
[1,]  2   5   6
[2,]  0   7   6
[3,] 11   1   1
```

Fig. 2.1 Sutton’s leisure time data: descriptive statistics

2.2 Visualizing Ranking Data

Visualization techniques for ranking data have drawn the attention of many researchers. Some of them are basically adopted from classical graphical methods for quantitative data while some are tailor-made for ranking data only. In this section, we will briefly review various graphical visualization methods and discuss the similarities and differences among them. Essentially, when a graphical method is developed for displaying ranking data, we would like this method to help answer the following questions:

1. What is the typical ranking of the t objects (the general preference)?
2. To what extent is there an agreement among the judges (the dispersion)?
3. Are there any outliers among the judges and/or the objects?
4. What are the similarity and dissimilarity among the objects?

Note that when the size of the ranking data is large (e.g., $t \geq 8$ or $n \geq 100$), it is practically impossible to reveal the abovementioned pattern and characteristics by merely looking at the raw data or by using some simple descriptive statistics such as the means and standard deviations of the ranks. In this section, we will focus on several major visualization methods—permutation polytopes, multidimensional scaling (MDS) and unfolding (MDU), and multidimensional preference analysis (MDPREF). For other visualization methods, see the monograph by Marden (1995).

2.2.1 *Permutation Polytope*

To display a set of rankings, it is not advisable to use traditional graphical methods such as histograms and bar graphs because the elements of \mathcal{P} , the set of all possible permutations of the t objects, do not have a natural linear ordering.

Geometrically, rankings of t objects can be represented as points in \mathbb{R}^{t-1} . The set of all $t!$ rankings can then form a convex hull of $t!$ points in \mathbb{R}^{t-1} known as a permutation polytope. The idea of using a permutation polytope to visualize ranking data was first proposed by Schulman (1979) and was considered later by McCullagh (1993a). Thompson (1993a,b) initiated the use of permutation polytopes to display the frequencies of a set of rankings in analogy with histograms for continuous data.

For complete ranking data, frequencies can be plotted on the vertices of a permutation polytope. Based on this polytope, Thompson found that the two most popular metrics for measuring distance between two rankings are the Kendall and Spearman distances which provide natural geometric interpretations of the rankings. More specifically, she showed that the minimum number of edges that must be traversed to get from one vertex of the permutation polytope to another reflects the Kendall distance between the two rankings labeled by the two vertices, whereas the Euclidean distance between any two vertices is proportional to the Spearman distance between the two rankings corresponding to the two vertices.

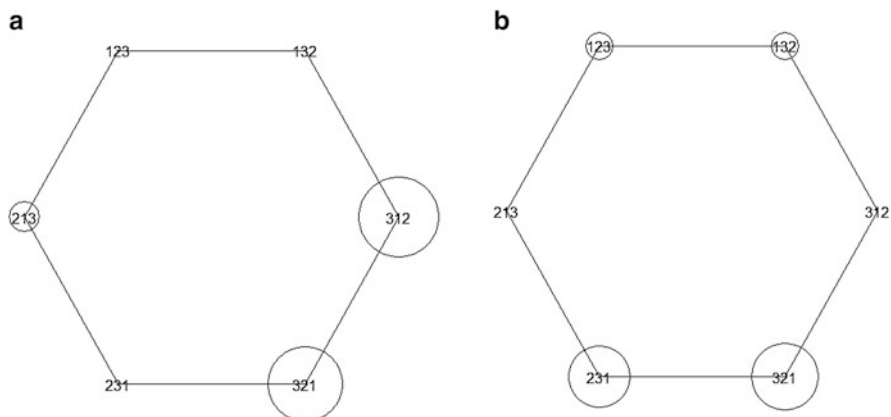


Fig. 2.2 A hexagon displaying the leisure time data set (a) White females (b) Black females

Example 2.2. Note that a ranking of t objects can be represented as a data point located in Euclidean space \mathbb{R}^{t-1} . Therefore, only rankings of three or four objects can be represented in a two-dimensional or three-dimensional graph without losing any information. For instance, ranking data with three objects can be displayed on a hexagon, in which each vertex represents a ranking and each edge connects two rankings a Kendall tau distance of 1 apart. To plot the leisure preferences given by black and white females in Sutton’s Leisure Time data, we can make use of the `rankplot` function in the `pmr` package to produce the hexagons as shown in Fig. 2.2. Note that the area of the circle in each vertex is proportional to the frequency of the corresponding ranking. With a quick glance, we can see that the two rankings (2, 3, 1) and (3, 2, 1) have the largest frequencies for black females, indicating that many black females preferred to spend their leisure with “both sexes” the most while most of white females did not prefer to spend time with “male(s).”

Example 2.3. Consider the case of rankings of four objects. The $24(=4!)$ vertices form a permutation polytope in three dimensions called a truncated octahedron (Thompson 1993b). For illustration, a truncated octahedron of the four-object ranking data `big4yr` (the relative ranking of four teams in the English Premier League (EPL), namely Arsenal (1), Chelsea (2), Liverpool (3), and Manchester United (4), from 1992–1993 season to 2012–2013 season) is plotted in Fig. 2.3.

It has eight hexagonal faces and six square faces. Each face has its interpretation. Four of the hexagons refer to the rankings where a particular object is ranked first, the other four refer to the rankings where a particular object is ranked last, and the six square faces refer to the rankings where two particular objects are ranked among the top two. The hexagon face with Manchester United (4) ranked first implying that Manchester United was the best team in the EPL over many seasons.

In contrast to a complete ranking, a partial ranking is represented by a permutation of t *nondistinct* numbers. For example, the top 2 partial ranking (2, –, –, 1)

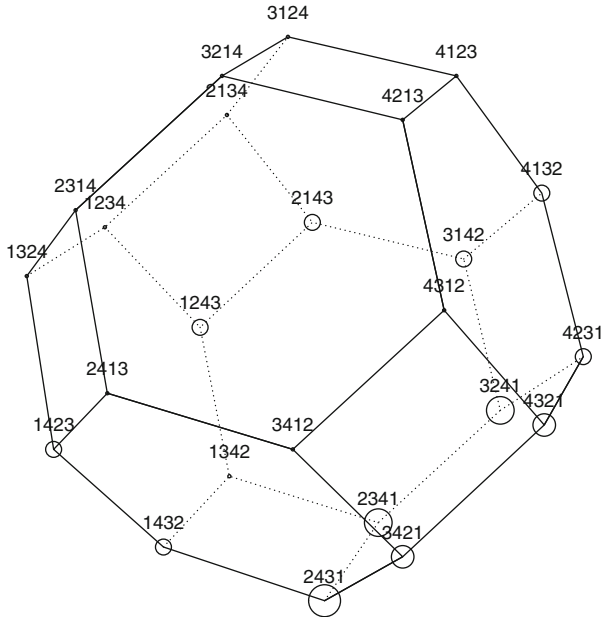


Fig. 2.3 A truncated octahedron displaying the rankings of the Big Four EPL teams

can be represented by $(2, 3.5, 3.5, 1)$. Therefore, the permutation polytope is not applicable to represent partial rankings. To tackle this problem, Thompson (1993a) defined a generalized permutation polytope as the convex hull of points in \mathcal{P} whose coordinates are permutations of t nondistinct numbers. The frequencies of a set of partial rankings can be plotted in a natural way on the vertices of a generalized permutation polytope. Similar to permutation polytopes, generalized permutation polytopes induce a new and very reasonable extension of Kendall and Spearman distances for top q partially ranked data.

Owing to the fact that the generalized permutation polytope on which the frequencies are displayed is inscribed in a sphere in a $(t - 1)$ -dimensional subspace of \mathcal{P} , it is difficult to visualize all the points on a polytope in a high-dimensional space for $t \geq 5$. Thompson (1993a) proposed an approach to explore high-dimensional polytopes by examining the three-dimensional faces and portions of the four-dimensional faces. However, because drawing permutation polytopes is fairly difficult, this visualization method is not so commonly used.

2.2.2 *Multidimensional Scaling and Unfolding*

Multidimensional scaling is a collection of graphical methods for representing data which are in the form of similarities, dissimilarities, or other measures of

“closeness” between each pair of objects. Unless the data are already in that form a transformation on the raw data is required in order to obtain all pairs of objects in their (dis)similarity measurements. See Tables 1.1 and 1.2 of Cox and Cox (2001) for various transformations to calculate (dis)similarity measures for quantitative and binary data. The basic idea behind MDS is to search for a low-dimensional space, usually Euclidean, in which each object is represented by a point in the space, such that the distances between the points in the space “match” as well as possible with the original (dis)similarities. Applications of MDS can be found in fields such as behavioral science, marketing, and ecology.

The starting point consists of an $n \times n$ nonnegative symmetric matrix $\Delta = (\delta_{ij})$ of dissimilarities among n observations (e.g., products, people, or species) such that δ_{ij} indicates the perceived dissimilarity between observations i and j . The goal of MDS is, given Δ , to find n points $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ in a low-dimensional space such that the distances between the points approximate the given dissimilarities:

$$\delta_{ij} \approx \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}.$$

Various approaches have been developed to determine the low-dimensional points \mathbf{x}_i 's. One typical approach is to formulate MDS as an optimization problem, where the n points $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are found by minimizing some loss function, commonly called *stress*, for instance,

$$\max_{x_1, \dots, x_n} \sum_{i < j} (\delta_{ij} - d_{ij})^2,$$

where $d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}$.

In the context of ranking data, Kidwell et al. (2008) suggested computing the dissimilarity between any two complete or partial rankings on t objects by Kendall distance proposed by Alvo and Cabilio (1995a) (see Chap. 3 for more details) and then applied MDS to find an embedding of a data set of n rankings assigned by n judges in a two- or three-dimensional Euclidean space.

Example 2.4. Consider a movie rating data set containing 72,979 possibly incomplete and tied rankings of 55 movies made by 5,625 raters who visited the web site MovieLens (Resnick et al. 1994) in 2000. We calculate the distance matrix by using a normalized version of the Kendall distance:

$$d^* = \frac{d_{orig}^* - m^*}{M^* - m^*} \quad (2.1)$$

where d_{orig}^* is the Kendall distance, and M^* and m^* are the maximum and minimum values of the Kendall distance as defined later in Lemma 3.5.

Applying two-dimensional MDS to the distance matrix, we obtain a scatterplot of 5,625 points for the movie raters. However, the points are too densely clustered that the scatterplot is ineffective to visualize the patterns of the ranking data. Kernel smoothing is therefore used to produce a heat map for better identification of different clusters of movie raters (Fig. 2.4).

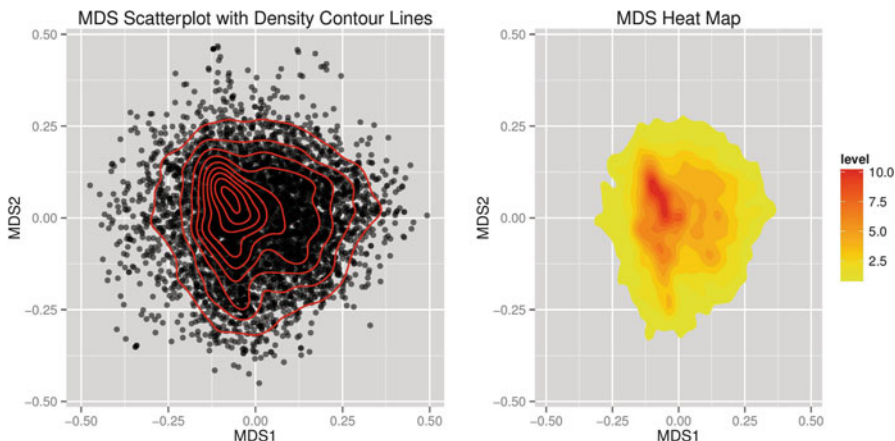


Fig. 2.4 MDS scatterplot and heat map for the movie rating data

The movie-viewing population initially appears to be a single large cluster. A closer look identifies a dichotomy within the cluster where preferences vary from the left to the right. The right half of the cluster appears to appreciate action films with sci-fi elements—the *Star Wars* series and *The Matrix*. In the left half, the upper part mainly consists of people interested in romance, *Casablanca* and *The Graduate*, while the people in the lower part enjoy serious drama movies more—*Seven Samurai* and *To Kill a Mockingbird*.

Among the MDS techniques, the unfolding technique, formulated by Coombs (1950), is a typical variant designed for representing ranking data. Unlike the abovementioned MDS that only visualizes a set of judge points in a low-dimensional Euclidean space, the unfolding technique attempts to visualize a set of points in a low Euclidean space with both judges and objects being represented by the points in the same space. The points are sought so that the ranked order of the distances from a *judge* point to the object points matches as “close” as possible with the ranking assigned by the judge.

Mathematically, suppose that the j th judge is represented by the point $\mathbf{x}_j = (x_{j1}, \dots, x_{jd})'$ in \mathbb{R}^d ($j = 1, \dots, n$) and the i th object is represented by the point $\mathbf{y}_i = (y_{i1}, \dots, y_{id})'$ in the same space ($i = 1, \dots, t$). Assume that the degree of preference of the i th object given by the j th judge is measured by the Euclidean distance, d_{ij} , between \mathbf{x}_j and \mathbf{y}_i , where

$$d_{ij} = \sqrt{(\mathbf{y}_i - \mathbf{x}_j)'(\mathbf{y}_i - \mathbf{x}_j)}. \quad (2.2)$$

The smaller the value of d_{ij} , the more preferable for the j th judge is the i th object.

The problem of multidimensional unfolding (MDU) is to find \mathbf{x}_j 's and \mathbf{y}_i 's such that the distances d_{ij} 's match as much as possible with the ranks of objects given by the judges. In other words, this can be viewed as MDS for a rectangular dissimilarity matrix $\mathbf{\Delta}$ whose (i, j) entry represents the rank of object i assigned by judge j . Various methods are available to tackle this problem. See Chap. 8 of Cox and Cox

(2001) and Chaps. 14–16 of Borg and Groenen (2005) for detailed explanations on these methods.

Example 2.5. When $d = 1$, we have the so-called *unidimensional unfolding* for which objects and judges are represented by points on a straight line. For example, suppose there are two judges J_1 and J_2 who ranked the four objects A, B, C, and D and their rankings are

	First	Second	Third	Fourth
Judge J_1	A	B	C	D
Judge J_2	C	D	B	A

Then the unfolding result for this example is represented by the top line in Fig. 2.5. From the figure, we can see that for both judges, the distances from judge point (J_1 or J_2) to the four object points have the same ranking as his/her original ranking of the objects. It is interesting to note that when the line is folded from one side to the other side at any judge point, the judge’s rankings can be observed and hence the name *unfolding* is termed. For instance, the folded line in Fig. 2.5 reveals that judge J_1 prefers A the most, B the second, C the third, and D the least.

However, the ranking of the distances from a judge’s point to all the objects and the judge’s ranking cannot guarantee to be perfectly matched for every judge. For example, it is impossible to place a point for judge J_3 who ranked the objects as DABC in Fig. 2.5.

A widely used approach of solving MDS problems is called SMACOF (Scaling by MAjorizing a COMplicated Function) which minimizes stress by means of majorization. de Leeuw and Mair (2009) extended the basic SMACOF theory to cover more types of data structures including MDU and they developed the `smacof` package in R.

Example 2.6. Consider the Big Four data (Table A.2) in which the seasonal rankings of four EPL teams, Arsenal, Chelsea, Liverpool, and Manchester United, from the 1992–1993 season to 2012–2013 season. Applying SMACOF to this 4×21 matrix, we obtain the unfolding solution in Fig. 2.6.

In Fig. 2.6, the configuration plot shows the coordinates of the Big Four teams and the seasons jointly. We can see that the Big Four teams are specially located at four distinct configurations with Manchester United located at the center and

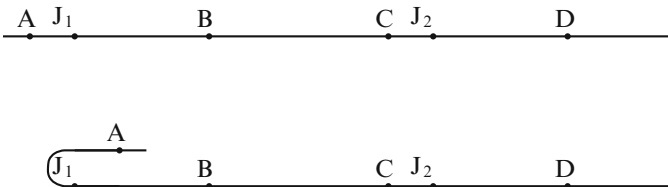


Fig. 2.5 Unidimensional unfolding

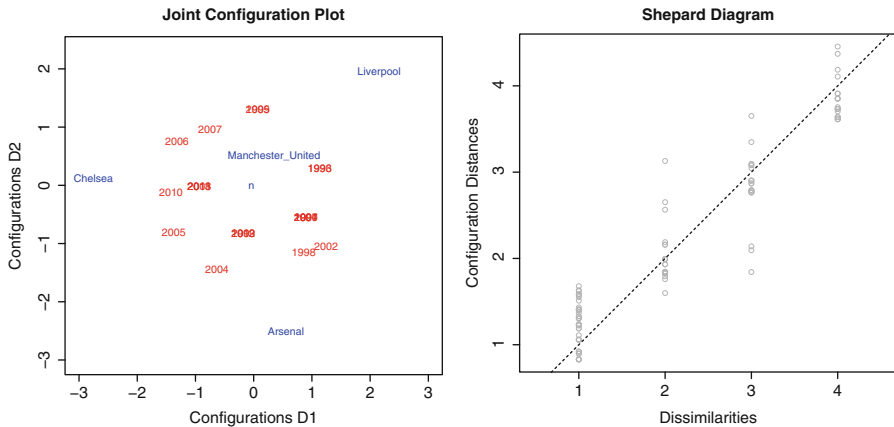


Fig. 2.6 Joint configuration plot and Shepard diagram for the Big Four data

surrounded by the seasons. This is reasonable since out of 21 seasons, Manchester United ranked first in 15 seasons. Examining the configurations of teams and seasons, most of the seasons in the 1990s are on the right of the graph where Liverpool and Arsenal are located, and the seasons in the recent years are on the left and Chelsea and Manchester United are the nearest teams. This indicates that Chelsea had its increasing ability over Liverpool and Arsenal in the recent decade.

The Shepard diagram in Fig. 2.6 shows a scatterplot of the reconstructed distances obtained from the SMACOF solution against the original dissimilarities (i.e., ranks) and an isotonic regression fitted to the points in the graph. If all reconstructed distances lie on the fitted line, then the dissimilarities would be perfectly reproduced by the MDS solution. From Fig. 2.6, the Shepard diagram shows some lack of fit, particularly for ranks 2 and 3 in a few seasons.

2.2.3 Multidimensional Preference Analysis

Similar to MDU, MDPREF (Carroll 1972) displays the relationship between judges and their respective perceptions on objects by reducing the dimensionality of the data, while retaining their main features as much as possible. Instead of using points as in the unfolding technique, each judge is now represented by a vector in a low-dimensional space while each object is represented by a point in the same space. The vectors and points are so chosen that the projections of the object points onto the judge vector indicate the rankings of the judge as closely as possible.

More specifically, let $\mathbf{y}^{(j)} = (y_{j1}, \dots, y_{jt})'$, $j = 1, \dots, n$ be a $t \times 1$ vector of anti-ranks for t objects given by judge j . Also, let $\mathbf{Y} = [\mathbf{y}^{(1)} \dots \mathbf{y}^{(n)}]'$ be a collection of the rankings given by n judges. Notice that

$$y_{ji} = t + 1 - \delta_{ij}, i = 1, \dots, t, j = 1, \dots, n,$$

where δ_{ij} is the dissimilarity (rank) measure used in the MDU in the previous section. Since each row of \mathbf{Y} is just a particular permutation of the integers $1, 2, \dots, t$, the mean of the ranks in any row equals $\frac{t+1}{2}$. For the sake of convenience, each row of \mathbf{Y} is centered by $\frac{t+1}{2}$, resulting in a new matrix, \mathbf{Y}_c :

$$\mathbf{Y}_c = \mathbf{Y} - \frac{t+1}{2} \mathbf{1}_n \mathbf{1}'_t$$

where $\mathbf{1}_m$ is a $m \times 1$ vector with all entries being 1.

The problem of MDPREF is to factorize \mathbf{Y}_c as

$$\mathbf{Y}_c = \mathbf{G} \mathbf{H}', \quad (2.3)$$

where \mathbf{G} and \mathbf{H} are $n \times d$ and $t \times d$ matrices, respectively. By comparing entries on both sides of (2.3), we obtain

$$y_{ij} - \frac{t+1}{2} = \mathbf{g}'_i \mathbf{h}_j,$$

where \mathbf{g}_i and \mathbf{h}_j are the rows of \mathbf{G} and \mathbf{H} , respectively, and we have a geometric representation of \mathbf{Y}_c in terms of d -dimensional vectors.

Suppose the rectangular matrix \mathbf{Y}_c has rank $r (\leq \min(n, t-1))$. A natural approach to determine \mathbf{G} and \mathbf{H} is to use the singular value decomposition:

$$\mathbf{Y}_c = \mathbf{P} \mathbf{\Lambda} \mathbf{Q}',$$

where \mathbf{P} is an $n \times r$ orthogonal matrix of rank r , \mathbf{Q} is an $t \times r$ orthogonal matrix of rank r , $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$, and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$ are the positive eigenvalues of \mathbf{Y}_c . Here, \mathbf{P} , \mathbf{Q} , $\mathbf{\Lambda}$ can also be obtained using the spectral decomposition, since $\mathbf{Y}_c \mathbf{Y}'_c = \mathbf{P} \mathbf{\Lambda}^2 \mathbf{P}'$ and $\mathbf{Y}'_c \mathbf{Y}_c = \mathbf{Q} \mathbf{\Lambda}^2 \mathbf{Q}'$.

Using only the $d (\leq r)$ largest eigenvalues, \mathbf{Y}_c can be approximated by

$$\hat{\mathbf{Y}}_c = \mathbf{P}_d \mathbf{\Lambda}_d \mathbf{Q}'_d, \quad (2.4)$$

where \mathbf{P}_d , \mathbf{Q}_d denote the matrices consisting of the first d columns of \mathbf{P} and \mathbf{Q} , respectively, and $\mathbf{\Lambda}_d = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$. In fact, it was shown by many researchers that the approximation is the least squares solution of the problem of minimizing

$$\text{trace}[(\mathbf{Y}_c - \mathbf{X})(\mathbf{Y}_c - \mathbf{X})'] = \sum_{i=1}^n \sum_{j=1}^t (y_{ij} - \frac{t+1}{2} - x_{ij})^2$$

among all $n \times t$ matrices \mathbf{X} of rank d or less.

A d -dimensional MDPREF solution includes the following steps: (a) label object i in \mathbb{R}^d by a point represented by the i th row of \mathbf{Q}_d , and (b) label judge j on the same graph represented by the j th row of $\mathbf{P}_d \mathbf{\Lambda}_d$, by an arrow drawn from the origin on the graph. To give a better graphical display, the length of the judge vectors can be scaled to fit the position of the objects. It is not difficult to see that the perpendicular

projection of all t object points onto a judge vector will closely approximate the ranking of the t objects by that judge if the current MDPREF solution fits the data well. Otherwise, we may look for a higher-dimension solution.

Example 2.7. Let us revisit the data on ranking the Big Four EPL teams (Table A.2). Applying MDPREF to this Big Four data gives the two-dimensional MDPREF solution as shown in Fig. 2.7.

```
> mdpref(big4yr,rank.vector=T)

$item
      [,1]      [,2]
[1,] -0.1208  1.0514
[2,] -1.3187 -1.9222
[3,] -1.1504  1.2467
[4,]  2.5898 -0.3759

$ranking
      [,1] [,2] [,3] [,4] [,5]      [,6]      [,7]
[1,]    3    4    2    1    1  0.83970  0.5716
[2,]    2    4    3    1    1  1.00136  0.5254
[3,]    4    3    2    1    1  0.65162 -0.1316
[4,]    3    4    2    1    1  0.83970  0.5716
[5,]    2    4    3    1    1  1.00136  0.5254
[6,]    1    4    3    2    1  0.57576  0.8630
[7,]    2    3    4    1    1  0.97494 -0.2240
[8,]    2    4    3    1    1  1.00136  0.5254
[9,]    2    4    3    1    1  1.00136  0.5254
[10,]   1    4    2    3    1 -0.01150  1.2467
[11,]   2    3    4    1    1  0.97494 -0.2240
[12,]   1    2    4    3    1 -0.06435 -0.2521
[13,]   2    1    4    3    1 -0.25244 -0.9554
[14,]   4    1    3    2    1  0.01150 -1.2467
[15,]   4    2    3    1    1  0.62519 -0.8810
[16,]   3    2    4    1    1  0.78685 -0.9272
[17,]   4    3    2    1    1  0.65162 -0.1316
[18,]   3    1    4    2    1  0.17317 -1.2929
[19,]   3    2    4    1    1  0.78685 -0.9272
[20,]   2    3    4    1    1  0.97494 -0.2240
[21,]   3    2    4    1    1  0.78685 -0.9272

$explain
[1] 0.7232
```

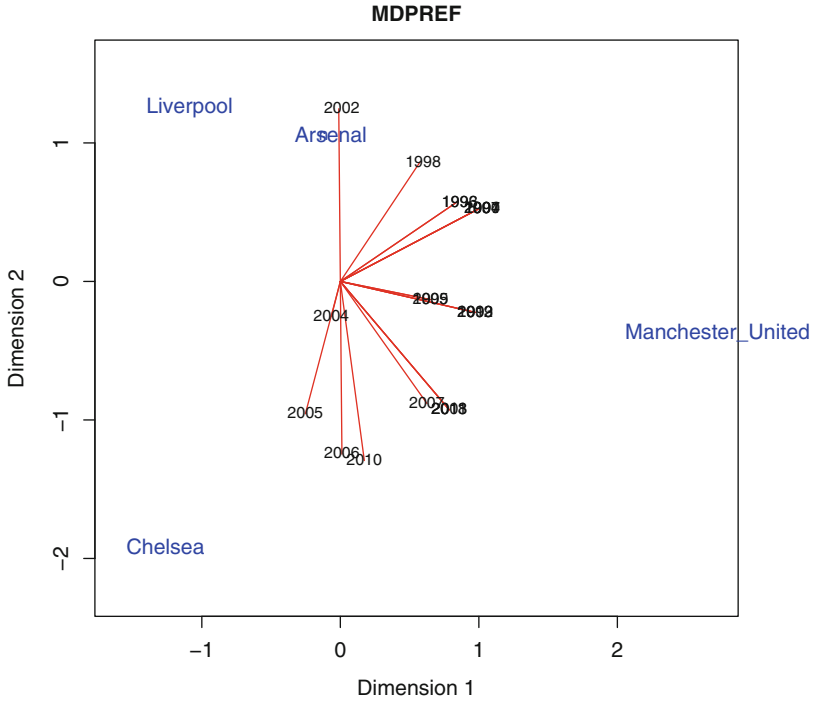


Fig. 2.7 The two-dimensional MDPREF solution for the Big Four data

The coordinates of the objects and season rankings and the proportion of variance explained by the first two dimensions are stored in the values `$item`, `$ranking`, and `$explain`, respectively. It can be seen that two dimensions can explain about 72.32% of the total variance. Figure 2.7 shows the two-dimensional MDPREF solution. The first dimension can be interpreted as the overall preference of the four teams. Manchester United performed best, whereas Liverpool and Chelsea performed poorly. The second dimension represents the contrast between Chelsea and the two teams, Liverpool and Arsenal. Examining their rankings in Table A.2, we see that Chelsea had its increasing performance over Liverpool and Arsenal in the recent decade.

Note that the season vectors point to the direction of the best performing team. For example, the ranking (3, 2, 4, 1) was observed in three seasons in 2008, 2011, and 2013, meaning that Manchester United ranked first, then Chelsea second, Arsenal third, and finally Liverpool the last. They all have the same season vector pointing to (0.79, -0.93). By projecting the four team points to the season vector in Fig. 2.7, we obtain correctly the ordering Manchester United > Chelsea > Arsenal > Liverpool.

2.2.4 Comparison of the Methods

So far, we have described three methods of visualizing ranking data. Each method has its distinct features. Permutation polytopes provide the frequency distribution of the observed rankings. Unfolding and MDPREF attempt to reveal relationships among the judges and the objects such as clustering of objects and unusual judges.

Note that permutation polytopes do not make any assumptions while unfolding and MDPREF assume that the ranking assigned by each judge can be represented in terms of the ordering of distances and projections, respectively.

Despite their features and assumptions, each of them has their own weaknesses. Permutation polytopes become difficult to interpret for large numbers of objects, say more than 5. Unfolding and MDPREF rely on the validity of assumptions made in each method which may not be true in general. Nevertheless, unfolding and MDPREF can often provide a very informative graph which gives a spatial representation of objects and judges and from which the between-object and between-judge relationships could then be identified (see Cohen and Mallows 1980).

2.3 Tests for Randomness on Ranking of a Set of Objects

When we say that selecting a ranking of t objects is completely random, we mean all possible rankings of t objects have the same probability (i.e., a uniform distribution) of being selected. In that case, the expected frequencies of each ranking of t objects in a data set of size n should be $n/t!$, and the standard chi-square goodness-of-fit test can be applied to test for uniformity. However, it is not always applicable when $t!$ is too large as compared to n , because we may encounter the problem of observed frequencies on some rankings being less than 5. If such a case occurs, Marden (1995) suggested using mean ranks, pairs, or marginals (see Sect. 2.1), instead of ranking proportions, to test for randomness.

Under the null hypothesis H_0 of randomness, the expected mean ranks, pairs, and marginals should be $(t + 1)/2$, $0.5n$, and n/t , respectively. Under H_0 , the test statistics based on the sample mean ranks, pairs, and marginals are

$$\text{Mean rank: } \frac{12n}{t(t+1)} \sum_{j=1}^t \left(m_j - \frac{t+1}{2}\right)^2$$

$$\text{Pairs: } 12n \left\{ \sum_{a>b}^t (P_{ab} - 0.5)^2 - \frac{1}{t+1} \sum_{j=1}^t \left(m_j - \frac{t+1}{2}\right)^2 \right\}$$

$$\text{Marginals: } n(t+1) \sum_{a>b}^t \left(M_{ab} - \frac{1}{t}\right)^2$$

which follow a χ^2 distribution with degrees of freedom $t - 1$, $\binom{t}{2}$, and $(t - 1)^2$, respectively.

Example 2.8. To assess for randomness for Sutton's Leisure Time data based on the mean ranks, the following R code could be used:

```
> data(leisure.black)
> de1<-destat(leisure.black)
> mean<-rep(2,3)
> chi<-12*13*sum((de1$mean.rank - mean)^2)/3/4
> chi
[1] 11.69
> dchisq(chi,2)
[1] 0.001445
```

The value of the test statistic is 11.69 which has a p-value of 0.0014. This indicates that black females had uneven preference on the three choices.

Example 2.9. The above χ^2 tests can also be extended to evaluate the agreement or diversity between two populations of judges who ranked the same set of objects. We may test for the agreement of leisure time preferences between black females and white females based on the marginals. The R code used is given as follows:

```
> de1 <- destat(leisure.black)
> data(leisure.white)
> de2 <- destat(leisure.white)
> chisq.test(cbind(as.vector(de1$mar),
                  as.vector(de2$mar)))
```

Pearson's Chi-squared test

```
data: cbind(as.vector(de1$mar), as.vector(de2$mar))
X-squared = 27.22, df = 8, p-value = 0.0006479
```

The value of the test statistic is 27.22 which has a p-value of 0.00065. It is evident that black and white females did not agree on their preference with which sex they prefer to spend their leisure time. Note that the marginal matrix has only $(t-1)^2 = 4$ df and hence the proper degrees of freedom of the chi-square test here should be 4. However, this will make p-value even smaller and hence this does not affect our conclusion.

Chapter Notes

Alvo and Ertas (1992) extended MDPREF to visualize rankings obtained from more than one population. Yu and Chan (2001) and Leung (2003) developed a probabilistic extension of MDPREF and MDU, respectively, so that statistical inference on model parameters can be made. Other graphical representation techniques include Ye and McCullagh (1993), Han and Huh (1995), Baba (1986), and Hirst and Naes (1994). For examining agreement or diversity among three or more populations of judges, see Chap. 4 and Marden (1995) for some distance-based methods and MANOVA-like methods.