11 Clustering, Distance Methods and Ordination

Further reading. The paper Kaufmann and Whiteman (1999) applies cluster analysis to wind patterns in the Grand Canyon Region.

11.1 Introduction

Broadly speaking, cluster analysis involves categorization: dividing a large group of observations into smaller groups so that the observations within each group are relatively similar, i.e. they possess largely the same characteristics, and the observations in different groups are relatively dissimilar.

We shall discuss some additional displays based on certain measures of distance and suggested step-by-step rules (algorithms) for grouping objects (variables or items). Searching the data for a structure of "natural" groupings is an important exploratory technique. Groupings can provide an informal means for assessing dimensionality, identifying outliers, and suggesting interesting hypotheses concerning relationships.

Grouping, or clustering, is distinct from the classification methods discussed in Chapter 10:

- Classification pertains to a known number of groups, and the operational objective is to assign new observations to one of these groups;
- Cluster analysis is a more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities).

Example (Weather Report, p. 1-6)**.** We want to find the locations with similar sunshine duration and precipitation characteristics.

11.2 Similarity Measures

Most efforts to produce a rather simple group structure from a complex data set require a measure of "closeness" or "similarity". There is often a great deal of subjectivity involved in the choice of a similarity measure. Important considerations include the nature of the variables (discrete, continuous, binary), scales of measurement (nominal, ordinal, interval, ratio), and subject matter knowledge.

When items (units, cases) are clustered, proximity is usually indicated by some sort of distance. By contrast, variables are usually grouped on the basis of correlation coefficients or like measures of association.

11.2.1 Distances and Similarity Coefficients for Pairs of Items

Definition 11.2.1. Generally a distance measure $d(P, Q)$ between two points P and Q satisfies the following properties, where R is any other intermediate point:

$$\begin{aligned} &d(P, Q) = d(Q, P) \\ &d(P, Q) > 0, \text{ if } P \neq Q \\ &d(P, Q) = 0, \text{ if } P = Q \\ &d(P, Q) \leq d(P, R) + d(R, Q). \end{aligned}$$

Now we define distance measures d that are often used in clustering. Let $\boldsymbol{x}' = (x_1, \ldots, x_p)$ and $\boldsymbol{y}' = (y_1, \ldots, y_p)$. Then

- Euclidean distance: $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})'(\boldsymbol{x} - \boldsymbol{y})}$.
- Statistical distance: $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})'\mathbf{A}(\boldsymbol{x} - \boldsymbol{y})}$. Ordinarily, $\mathbf{A} = \mathbf{S}^{-1}$, where $\mathbf{S}$ contains the sample variances and covariances. However, without prior knowledge of the distinct groups, these sample quantities cannot be computed. For this reason, Euclidean distance is often preferred for clustering.
- Minkowski metric: $d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^m\right)^{1/m}$. For $m = 1$, $d(\boldsymbol{x}, \boldsymbol{y})$ measures the "city-block" distance between two points in p dimensions. For $m = 2$, $d(\boldsymbol{x}, \boldsymbol{y})$ becomes the Euclidean distance. In general, varying m changes the weight given to larger and smaller differences.

When items cannot be represented by meaningful p-dimensional measurements, pairs of items are often compared on the basis of presence or absence of certain characteristics. Similar items have more characteristics in common than do dissimilar items. Therefore we use binary variables, which assume the value 1 if the characteristic is present and the value 0 if the characteristic is absent.

Let us arrange the frequencies of matches and mismatches for items i and k in the form of a contingency table: In this table, a represents the frequency of 1–1 matches, b

Table 11.1: Contingency table.

		Item k		
		1	0	Totals
Item i	1	a	b	$a + b$
	0	c	d	$c + d$
Totals		$a + c$	$b + d$	$p = a + b + c + d$

is the frequency of 1–0 matches, and so forth.

There exist a lot of similarity coefficients, defined in terms of the frequencies in Table 11.1. Table 11.2 shows a few examples:

Table 11.2: Similarity coefficients for clustering items.

$\frac{a+d}{p}$	:	equal weights for 1-1 and 0-0 matches
$\frac{2a}{2a+b+c}$	:	no 0-0 matches in numerator or denominator. Double weight for 1-1 matches
$\frac{a}{b+c}$	:	ratio of matches to mismatches with 0-0 matches excluded.

11.2.2 Similarities and Association Measures for Pairs of Variables

Thus far, we have discussed similarity measures for items. In some applications, it is the variables, rather than the items, that must be grouped. Similarity measures for variables often take the form of sample correlation coefficients. When the variables are binary, the data can again be arranged in the form of a contingency table.

11.3 Hierarchical Clustering Methods

Hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions:

- Agglomerative hierarchical methods start with the individual objects. Thus, there are initially as many clusters as objects. The most similar objects are first grouped, and these initial groups are merged according to their similarities.
- Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two subgroups such that the objects in one subgroup are "far from" the objects in the other. These subgroups are the further divided into dissimilar subgroups. The process continues until there are as many subgroups as objects – that is, until each object forms a group.

The results of both agglomerative and divisive methods may be displayed in the form of a two-dimensional diagram known as a dendrogram. As we shall see, the dendrogram illustrates the mergers or divisions that have been made at successive levels.

In this section we shall concentrate on agglomerative hierarchical procedures and, in particular, linkage methods. Linkage methods are suitable for clustering items, as well as variables. We shall discuss, in turn, single linkage (minimum distance or nearest neighbor), complete linkage (maximum distance or farthest neighbor), and average linkage (average distance). The merging of clusters under the three linkage criteria is illustrated in Figure 11.1.

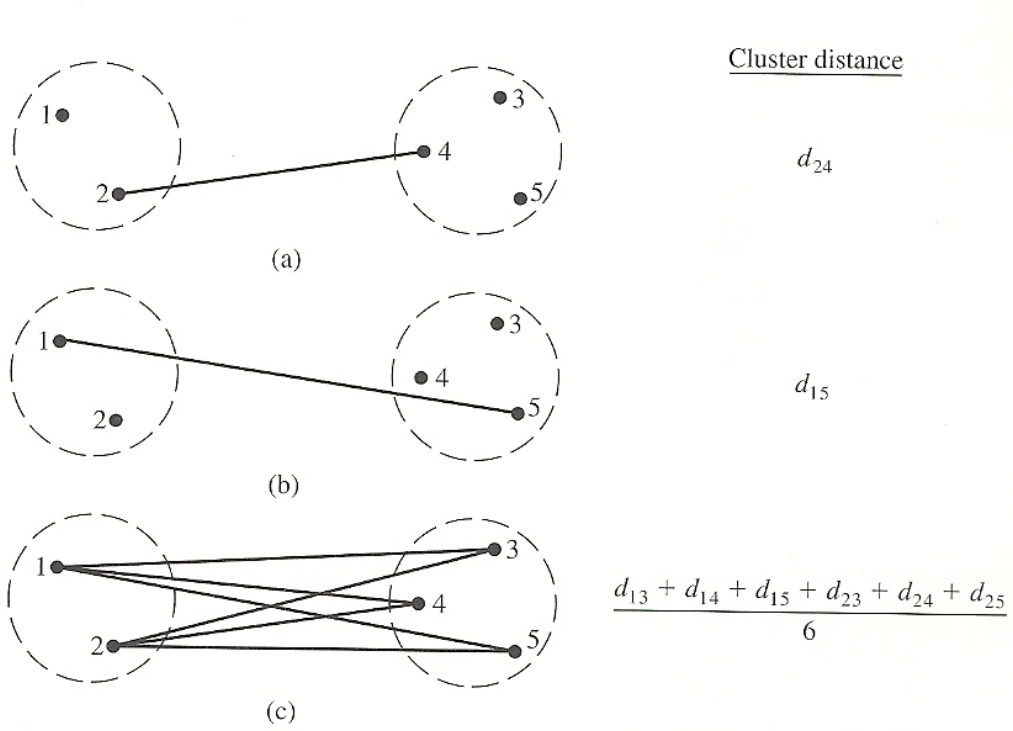

Figure 11.1: Intercluster distance (dissimilarity) for (a) single linkage, (b) complete linkage, and (c) average linkage. Source: Johnson and Wichern (2007).

The following are the steps in the agglomerative hierarchical clustering algorithm for grouping N objects (items or variables):

1. Start with N clusters, each containing a single entity and an $N \times N$ symmetric matrix of distances (or similarities) $\mathbf{D} = \{d_{ik}\}$.

2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{UV}.

3. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by

   - deleting the rows and columns corresponding to clusters U and V and

   ($\star$) adding a row and column giving the distances between cluster (UV) and the remaining clusters.

4. Repeat steps 2 and 3 a total of $N - 1$ times. All objects will be in a single cluster after the algorithm terminates. Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

In step ($\star$) there are different possibilities of defining distances:

- Single linkage (minimum distance): groups are formed from the individual entities by merging nearest neighbors, where the term nearest neighbor connotes the smallest distance or largest similarity.

  Initially, we must find the smallest distance in $\mathbf{D} = \{d_{ik}\}$ and merge the corresponding objects, say, U and V, to get the cluster (UV). For step $(\star)$ of the general algorithm, the distances between (UV) and any other cluster W are computed by

  $$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}.$$

- Complete linkage (maximum distance): Complete linkage clustering proceeds in much the same manner as single linkage clustering, with one important exception: At each stage, the distance between clusters is determined by the distance between the two elements, one from each cluster, that are most distant. Thus, complete linkage ensures, that all items in a cluster are within some maximum distance of each other.

  The general agglomerative algorithm again starts by finding the minimum entry in $\mathbf{D} = \{d_{ik}\}$ and merging the corresponding objects, such as U and V, to get cluster (UV). For step $(\star)$ of the general algorithm, the distances between (UV) and any other cluster W are computed by

  $$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}.$$

- Average linkage (average distance between all pairs of items): Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.

  For step $(\star)$ of the general algorithm, the distances between (UV) and any other cluster W are computed by

  $$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W},$$

  where d_{ik} is the distance between object i in the cluster (UV) and object k in the cluster W, $N_{(UV)}$ and N_W are the number of items in cluster (UV) and W, respectively.

- Ward's method: Ward's method is based on minimizing the "loss of information" from joining two groups. This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion.

  1. For a given cluster k, let ESS_k be the sum of the squared deviations of every item in the cluster from the cluster mean (centroid). If there are currently K clusters, define ESS as the sum of the ESS_k.
  2. At each step in the analysis, the union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS are joined.

Initially, each cluster consists of a single item, and, if there are N items, $\text{ESS}_k = 0$ for all $k = 1, \ldots, N$. At the other extreme, when all the clusters are combined in a single group of N items, the value of the ESS is given by

$$\text{ESS} = \sum_{j=1}^{K} (\boldsymbol{x}_j - \overline{\boldsymbol{x}})'(\boldsymbol{x}_j - \overline{\boldsymbol{x}}),$$

where $\boldsymbol{x}_j$ is the multivariate measurement associated with the jth item and $\overline{\boldsymbol{x}}$ is the mean of all items.

- Centroid method: Pruscha (2006) p. 299.

Remark. The different methods of distances have different characteristics.

- Simple linkage: it tends to be extremely myopic. An object will be added to a cluster so long as it is close to any one of the other objects in the cluster, even if it is relatively far from all the others. Thus, single linkage has a tendency to produce long, stringy clusters and nonconvex cluster shapes.
- Complete linkage: it tends to produce convenient and homogeneous groupings but it can be highly sensitive to outliers in the data.
- Average linkage: Compromise between single and complete linkage.
- Ward's method: seeks to join the two clusters whose merger leads to the smallest within-cluster sum of squares (i.e. minimum within-group variance). It has a tendency to produce equal-sized clusters that are convex and compact.

Example (Weather Report, p. 1-6)**.** We go back to the data set from Section 1.3.4. We want to find the weather stations with similar sunshine and precipitation characteristics. Figure 11.2 shows the clusters for the sunshine duration and precipitation with single linkage and the combination of both variables for single and complete linkage.

Example. MeteoSchweiz publishes several interesting working papers on special topics. In Begert (2008) Cluster analysis plays an important role to classify the different temperature and precipitation regions within the Swiss National Basic Climatological Network.

Remark. Some final comments concerning hierarchical clustering methods:

- Since there is no provision for a reallocation of objects that may have been "incorrectly" grouped at an early stage, the final configuration of clusters should always be carefully examined to see whether it is sensible.
- For a particular problem, it is a good idea to try several clustering methods and, within a given method, a couple different ways of assigning distances.

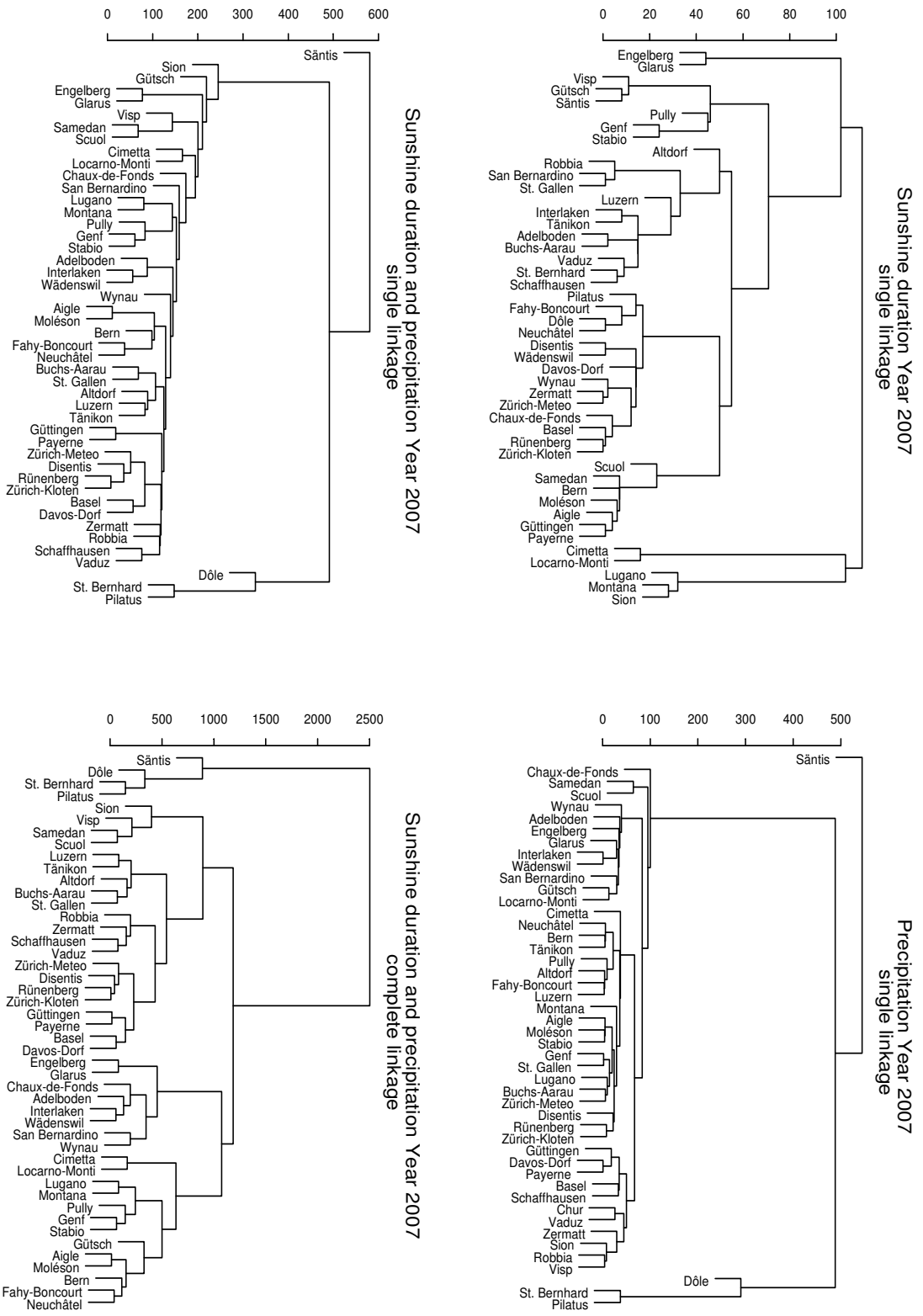

Figure 11.2: Cluster analysis for sunshine duration and precipitation for Swiss stations in the year 2007. Data set: Weather Report, p. 1-6

- The stability of a hierarchical solution can sometimes be checked by applying the clustering algorithm before and after small errors (perturbations) have been added to the data units. If the groups are fairly well distinguished, the clusterings before and after perturbation should agree.

- Some data sets and hierarchical clustering methods can produce inversions. An inversion occurs when an object joins an existing cluster at a smaller distance than that of a previous consolidation.

11.4 Nonhierarchical Clustering Methods

Nonhierarchical clustering techniques are designed to group items, rather than variables, into a collection of K clusters. The number of clusters, K, may either be specified in advance or determined as part of the clustering procedure. Because a matrix of distances does not have to be determined, and the basic data do not have to be stored during the computer run, nonhierarchical methods can be applied to much larger data sets than can hierarchical techniques.

Nonhierarchical clustering methods start from either

- an initial partition of items into groups or

- an initial set of seed points, which will form the nuclei of clusters. Good choices for starting configurations should be free of overt biases. One way to start is to randomly select seed points from among the items or to randomly partition the items into initial groups.

11.4.1 K-means Method

The K-means method assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of these three steps:

1. Partition the items into K initial clusters or specify K initial centroids (seed points).

2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. Distance is usually computed using Euclidean distance. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

3. Repeat step 2 until no more reassignments take place.

Remark. The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points.

Remark. To check the stability of the clustering, it is desirable to rerun the algorithm with a new initial partition. Once clusters are determined, intuitions concerning their interpretations are aided by rearranging the list of items so that those in the first cluster appear first, those in second cluster appear next, and so forth.

Remark. There are strong arguments for not fixing the number of clusters, K, in advance, including the following:

- If two or more seed points inadvertently lie within a single cluster, their resulting clusters will be poorly differentiated.
- The existence of an outlier might produce at least one group with very disperse items.
- Even if the population is known to consist of K groups, the sampling method may be such that data from the rarest group do not appear in the sample. Forcing the data into K groups would lead to nonsensical clusters.

In cases in which a single run of the algorithm requires the user to specify K, it is always a good idea to rerun the algorithm for several choices.

Remark. Discussions of other nonhierarchical clustering procedures are available in Pruscha (2006) p. 305f.

11.5 Clustering based on Statistical Models

The popular clustering methods discussed earlier in this chapter are intuitively reasonable procedures but that is as much as we can say without having a model to explain how the observations were produced. Major advances in clustering methods have been made through the introduction of statistical models that indicate how the collection of $p \times 1$ measurements $\boldsymbol{x}_j$, from the N objects, was generated. The most common model is one where cluster k has expected proportion p_k of the objects and the corresponding measurements are generated by a probability density function $f_k(\boldsymbol{x})$. Then, if there are K clusters, the observation vector for a single object is modeled as arising from the mixing distribution

$$f_{mix}(\boldsymbol{x}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}), \qquad \text{where each } p_k \geq 0 \text{ and } \sum_{k=1}^{K} p_k = 1.$$

This distribution $f_{mix}(\boldsymbol{x})$ is called a mixture of the K distributions $f_1(\boldsymbol{x}), \ldots, f_K(\boldsymbol{x})$ because the obeservation is generated from the component distribution $f_k(\boldsymbol{x})$ with probability p_k. The collection of N observation vectors generated from this distribution will be a mixture of observations from the component distributions.

The most common mixture model is a mixture of multivariate normal distributions where the k-th component $f_k(\boldsymbol{x})$ is the $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ density function.

For further details see Johnson and Wichern (2007), pp. 703–706.

11.6 Multidimensional Scaling

Multidimensional scaling techniques deal with the following problem: For a set of observed similarities (or distances) between every pair of N items, find a representation of the items in few dimensions such that the interitem proximities "nearly match" the original similarities (or distances).

Example (Weather Report, p. 1-6)**.** We consider the three variables which describe the differences of sunshine duration, temperature and precipitation from the norm. We are interested to find groups of Swiss stations with similar patterns. For the result see Figure 11.3.

It may not be possible to match exactly the ordering of the original similarities. Consequently, scaling techniques attempt to find configurations in $q \leq N-1$ dimensions such that the match is as close as possible. The numerical measure of closeness is called the stress.

Remark. There are two methods:

- Nonmetric multidimensional scaling: arrange the N items in a low-dimensional coordinate system using only the rank orders of the $N(N-1)/2$ original similarities (distances), and not their magnitudes.
- Metric multidimensional scaling (or principal coordinate analysis): the actual magnitudes of the original similarities (distances) are used to obtain a geometric representation in q dimensions.

For further details see Johnson and Wichern (2007), pp. 706–715.

11.7 Biplots for viewing Sampling Units and Variables

Further reading. In Gabriel (1971) the method of biplots was mentioned for the first time, in Gabriel (1972) the plots where used then as a tool to analyze the monthly rainfall of fifty-five stations in Israel.

A biplot is a graphical representation of the information in an $n \times p$ data matrix. The bi- refers to the two kinds of information contained in a data matrix:

- the information in the rows pertains to samples or sampling units
- the information in the columns pertains to variables.

A two-dimensional plot of the sampling units can be obtained by graphing the first two principal components. The idea behind biplots is to add the information about the variables to the principal component graph.

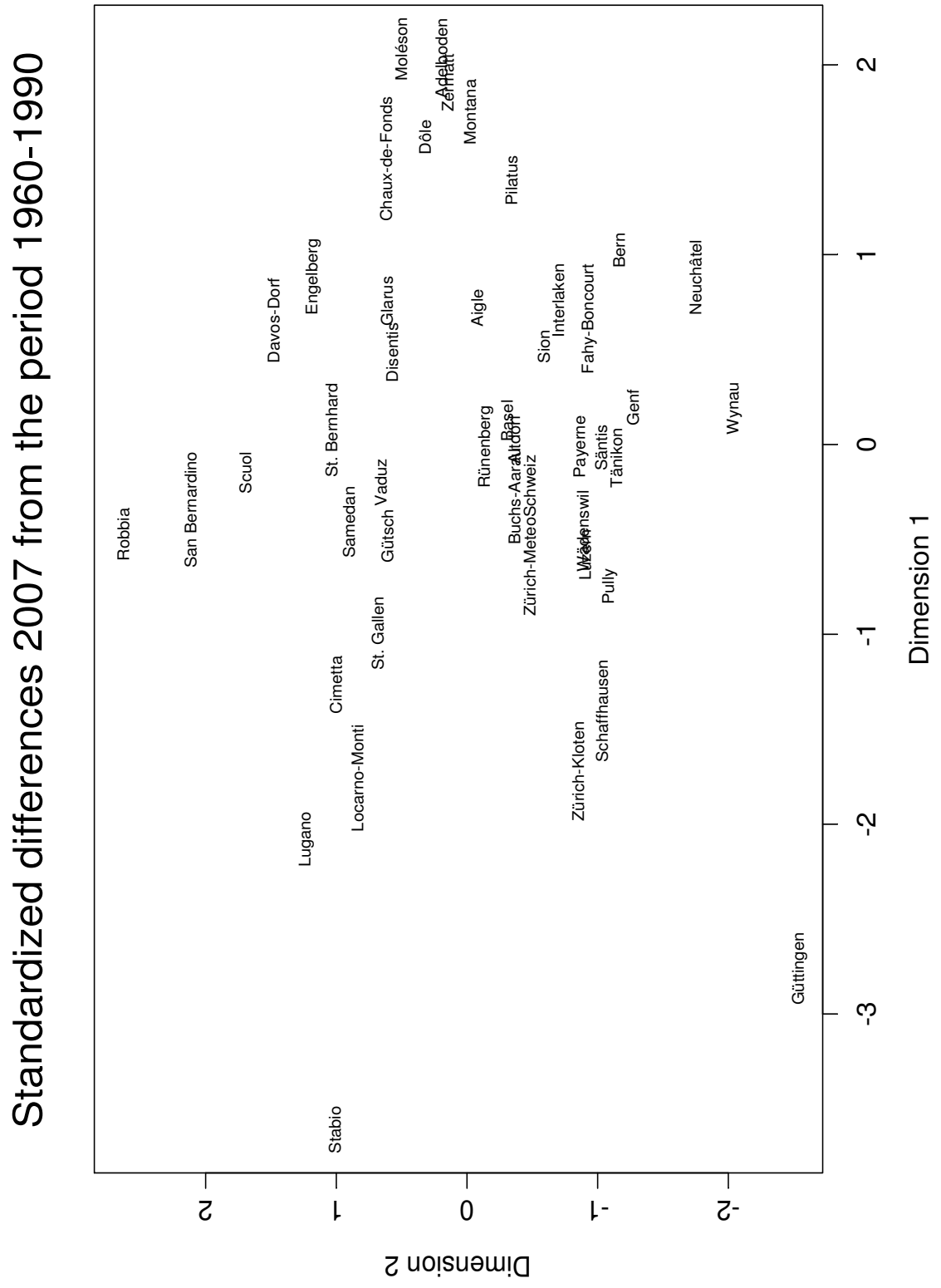

Figure 11.3: Multidimensional scaling for Swiss stations in the year 2007. Data set: Weather Report, p. 1-6.

Example (Weather Report, p. 1-6)**.** Consider the data set in Section 1.3.4. Figure 11.4 shows a biplot, where the variables altitude, sunshine duration, mean temperature, heating degree days, sum of precipitation and days of precipitation are included in the analysis.

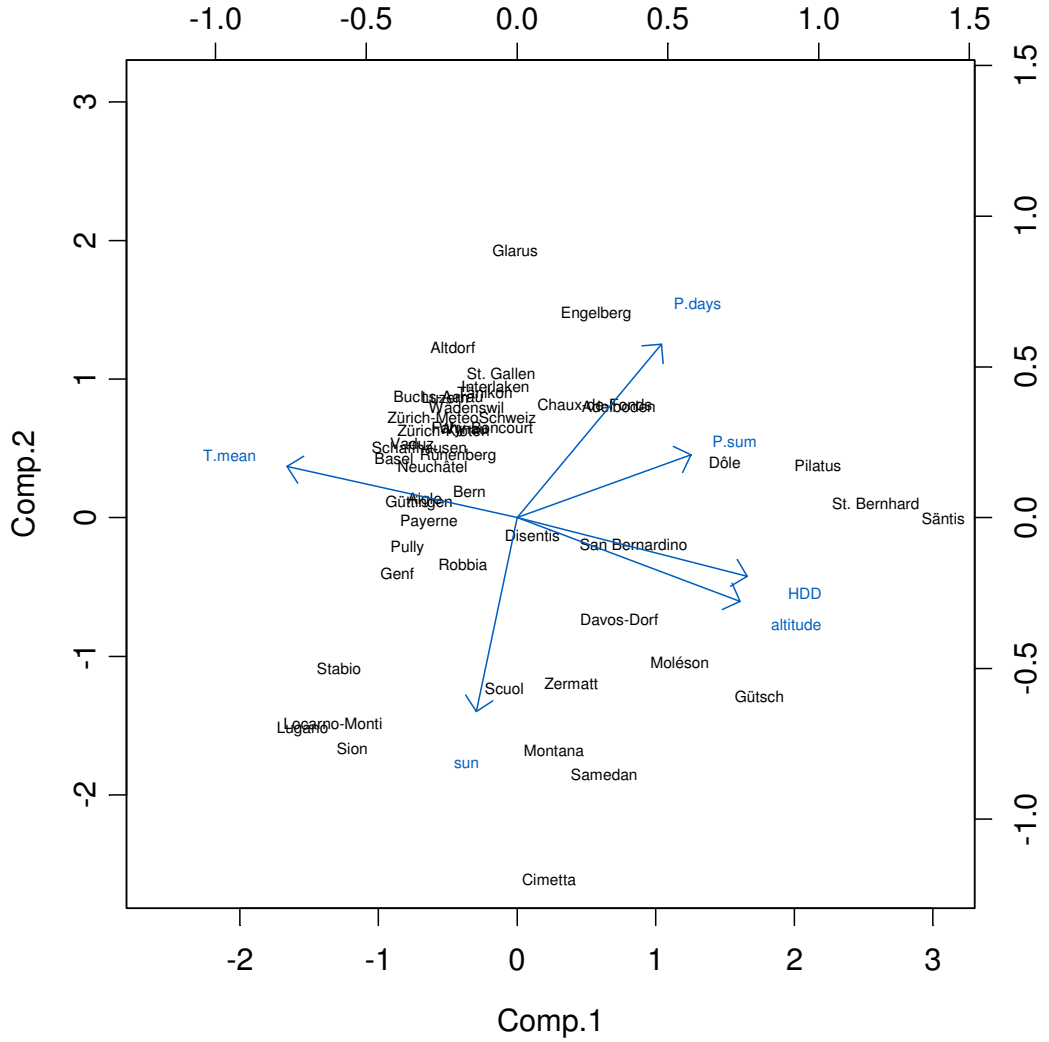

Figure 11.4: Biplot for Swiss stations in the year 2007. Data set: Weather Report, p. 1-6.

For further details see Johnson and Wichern (2007), pp. 726–732.