Distance-based analysis of dynamical systems and time series by optimal transport

Muskulus, M.

Appendix A

Distances

Life is like a landscape. You live in the midst of it but can describe it only from the vantage point of distance.

Charles Lindbergh

In this appendix we collect and discuss background information about distances and their statistical analysis. Section A.1 reviews the mathematical foundation and culminates in the characterization of the conditions under which a reconstruction of distances by points in an Euclidean space is possible. Section A.2 discusses how to obtain such reconstructions in practice and introduces various diagnostic measures that help to assess their quality. Section A.3 discusses the statistical analysis of distances and describes linear discriminant analysis in the reconstructed functional space, leave-one-out crossvalidation and permutation tests for group effects.

A.1 Distance geometry

The content of this section is developed in more detail in the standard monograph of Blumenthal (1953) and the article of Havel et al. (1983).

A.1.1 Distance spaces

An *abstract space* is a set of elements S, called *points*, that are endowed with a *topology*. The latter embodies a relation of nearness that results from defining certain subsets as *open*. A topology on S is then a collection $\mathcal{T}$ of all open subsets of S, such that the empty set $\emptyset$ and S are in $\mathcal{T}$, the union of any collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$, and the intersection of any finite collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$.

The main use of a topology is to allow for the definition of limits of sequences of elements. A sequence $(p_1, p_2, \dots)$ of elements $p_i \in S$ has a *limit* $p \in S$ if and only if for each integer $n \in \mathbb{N}$ there exists an open set $U_n \in \mathcal{T}$ such that $p \in U_n$ and $p_m \in U_n$, for all $m \geq n$, which is written as $\lim_{i \to \infty} p_i = p$.

Abstract spaces are too general in practice, since they do not need to have unique limits. For example, endowing a space S with the trivial topology $\mathcal{T} = \{\emptyset, S\}$, every point $p \in S$ is the limit of every sequence. Therefore, we will only consider the subset of abstract spaces that are also Hausdorff spaces. These have the following additional property (restriction): If $p \neq q$ for two points $p, q \in S$, then there exist

open sets $U_p, U_q \in \mathcal{T}$ such that $p \in U_p, q \in U_q$ and $U_p \cap U_q = \emptyset$. Since Hausdorff spaces separate their points, they are also called separated spaces.

Although the above notions are necessary for the study of functions on S, in particular, to define the concept of continuity, as a basis for making measurements in a space S additional structure is needed. This will again be axiomatically prescribed.

Definition 6. A *distance space* is an abstract set S together with a *distance* $d : S \times S \to D$ from an abstract *distance set* D.

We write $d(p, q) \in D$ for the value of the distance between two points $p, q \in S$. The most important case are numerical distances:

Definition 7. A distance space is called *semimetric* if (i) $D \subseteq \mathbb{R}_+$, (ii) $d(p, q) = d(q, p)$, and (iii) $d(p, q) = 0$ if and only if $p = q$.

Here $\mathbb{R}_+ = \{x \in \mathbb{R} | x \geq 0\}$ is the set of all non-negative real numbers. We can express the last two conditions in Definition 7 by saying that distances in a distance space are *symmetric* and *positive definite*, or simply by saying that they are *semimetric*.

Definition 8. The distance $d : S \times S \to D$ is *continuous* at $p, q \in S$, if for any two sequences $(p_n)_{n\geq 0}$ and $(q_n)_{n\geq 0}$ with limits $\lim_{n\to\infty} p_n = p$ and $\lim_{n\to\infty} q_n = q$, we have that $\lim_{n\to\infty} d(p_n, q_n) = d(p, q)$.

Continuous distances impose a certain regularity on distance spaces:

Theorem 3 (Blumenthal (1953))**.** A distance space with a continuous distance is Hausdorff.

Although $d(p, q) = 0$ if and only if $p = q$, there nevertheless still exists a potential anomaly in that two distinct points of a semimetric space may be joined by an arc of zero length:

Example 2 (Blumenthal)**.** Let $S = [0, 1]$ be the points of the unit interval and define the distance $d(x, y) = (x - y)^2$ for all points $x, y \in S$. Topologically, this space is equivalent to the space obtained by replacing d by the Euclidean distance $|x - y|$, so its character as a continuous line segment is unchanged, i.e., S is an arc.

Consider the sequence of polygons P_n with vertices

$$0, 1/2^n, 2/2^n, \ldots, (2^n - 1)/2^n, 1.$$

Each pair of consecutive vertices has distance $1/2^{2n}$ and since there are 2^n such pairs, the "length" of P_n is $1/2^n$. In the limit that $n \to \infty$, the length of S approaches zero. □

This anomaly results from the great freedom offered by the distance function, whose values are independent of each other, in the sense that the distance between

any pair of points does not depend on the distance between any other pair. Considering the simplest case of only three points, with three mutual distances, the following property is suggested from a closer look at Euclidean space:

Postulate 1 (Triangle inequality)**.** If p, q, r are any three points of a semimetric space, then

$$d(p,q) \leq d(p,r) + d(r,q). \tag{A.1}$$

Definition 9. A semimetric space in which the triangle inequality holds is called a *metric space*. The distance function of a metric space is called a *metric*.

Remark 6. The triangle inequality can be motivated differently. Let $(a,b),(c,d)$ be two pairs of ordered points in a semimetric space, and define $d(a,c)+d(b,d)$ as the distance of the pairs. When is this distance *uniformly continuous*? By this we mean that for each $\epsilon > 0$ there exists a number $\delta(\epsilon) > 0$ such that for all pairs $(a,b),(c,d)$ the property $d(a,c)+d(b,d) < \delta(\epsilon)$ implies $|d(a,b) - d(c,d)| < \epsilon$.

The *easiest way* to satisfy this requirement is if $|d(a,b) - d(c,d)| \leq d(a,c) + d(b,d)$, since then $\delta(\epsilon)$ may be taken to be equal to ϵ. But if this holds, then consideration of the pair $(a,b),(c,c)$ shows that this implies the triangle inequality, $d(a,b) \leq d(a,c) + d(c,b)$.

On the other hand, if the triangle inequality holds, then

$$|d(a,b) - d(c,d)| \leq |d(a,b) - d(b,c)| + |d(b,c) - d(c,d)| \leq d(a,c) + d(b,d),$$

where the first inequality arises from the triangle inequality of the modulus function, $|a+b| \leq |a| + |b|$.

Note that uniform continuity of a semimetric does not imply the triangle inequality in general. □

Example 3 (The n-dimensional Euclidean space E_n)**.** The points of E_n are all ordered n-tuples $(x_1, x_2, \ldots, x_n)$ of real numbers. The distance is defined for each pair of elements $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ by

$$d(x,y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}.$$

The triangle inequality follows from the Cauchy-Schwartz inequality. □

Example 4 (The n-dimensional spherical space S_n)**.** The points of S_n are all ordered $(n+1)$-tuples $x = (x_1, x_2, \ldots, x_{n+1})$ with $||x||^2 = \sum_{i=1}^{n+1} x_i^2 = 1$. Spherical distance is defined for each pair of elements x, y to be the smallest nonnegative number $d(x,y)$ such that

$$\cos(d(x,y)) = \sum_{i=1}^{n+1} x_i y_i.$$

This is an example of a geodesic (shortest-arc) distance. □

Example 5 (The Hausdorff metric)**.** A metric space M is *bounded* provided there exists a constant $K > 0$ such that $d(x, y) < K$ for all elements $x, y \in M$. Let X be the set of all closed, non-empty subsets of a bounded metric space M. Define

$$d(A, B) = \sup_{a \in A} \left(\inf_{b \in B} d(a, b) \right).$$

The function d is not a metric distance since it is not symmetric in general. Moreover, $d(A, B) = 0$ implies $\inf_{b \in B} d(a, b) = 0$ for all $a \in A$, such that $a \in \operatorname{cl} B = B$. Thus $B \subseteq A$, but in general $d(A, B) = 0$ does not imply that $A = B$. Both these shortcomings are fixed by symmetrizing d, and the resulting metric is called the Hausdorff metric:

$$d_H(A, B) = \max[d(A, B), d(B, A)].$$

To prove the triangle inequality, note that if $d(A, B) < \rho$, then $\inf_{b \in B} d(a, b) < \rho$ for all elements $a \in A$, so there exists $a \in A, b \in B$ such that $d(a, b) < \rho$. Let now A, B, C be three distinct elements of S and put $d(A, B) = \rho$ and $d(B, C) = \sigma$. For each $\epsilon > 0$ we have that $d(B, C) < \sigma + \epsilon$, and there exists $b \in B, c \in C$ such that $d(b, c) < \sigma + \epsilon$. Analogously, from $d(A, B) < \rho + \epsilon$ there exists $a \in A$ such that $d(a, b) < \rho + \epsilon$. Since M is metric,

$$d(a, c) \leq d(a, b) + d(b, c) < \rho + \sigma + 2\epsilon.$$

From this it follows that $d(A, C) \leq \rho + \sigma = d(A, B) + d(B, C)$. Similarly, it follows that $d(C, A) \leq d(C, B) + d(B, A)$. Together, the two relations

$$d_H(A, B) + d_H(B, C) \geq d(A, C)$$
$$d_H(A, B) + d_H(B, C) \geq d(C, A)$$

imply that $d_H(A, B) + d_H(B, C) \geq d_H(A, C)$, i.e., the triangle inequality in S. □

A.1.2 Congruence and embeddability

Topology was originally defined as the study of invariants of homeomorphisms, i.e., continuous functions with continuous inverses. Since homeomorphisms form a group, topology fits the definition of a geometry in the way of Felix Klein, as the study of invariants under a selected group of transformations.

The subgroup of homeomorphisms for which the distance of two points is an invariant is the group of *congruences*, and the resulting geometry is referred to as *distance geometry* (or metric topology).

Definition 10. If $p, q \in S$ and $p', q' \in S'$ for two metric spaces S, S' (with distances d, d'), then p, q are *congruent* to p', q' if and only if $d(p, q) = d'(p', q')$. Two subsets

P, Q of the same or different metric spaces are congruent provided there exists a map $f : P \to Q$ such that each pair of points from P is mapped onto a congruent point-pair of Q.

The relation of congruence is symmetric, reflexive and transitive, and therefore constitutes an equivalence relation.

We now consider the *subset problem*: What are necessary and sufficient conditions that an arbitrary distance space must satisfy in order that it may be congruent with a subset of a member of a prescribed class of spaces? In particular we will be interested in isometric embeddings of a finite set of points into Euclidean space E_n.

Definition 11. A set S is *congruently embeddable* (embeddable, for short) in a semimetric space T if S is congruent to a subset of T. A set S is *irreducibly embeddable* in E_n if it is embeddable in E_n, but not in any nontrivial subspace.

Definition 12. The *Gram matrix* of a set of vectors $\{x_i \mid 0 \leq i \leq N\}$ from an inner-product space is the matrix G of inner-products $G_{ij} = \langle x_i, x_j \rangle$. The *metric matrix* of a finite set of N points from a semimetric space, with respect to a reference point (indexed as the 0-th point), is the $(N \times N)$ matrix M with entries

$$M_{ij} = \frac{1}{2}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2), \tag{A.2}$$

where $d_{ij} = d(x_i, x_j)$ is the value of the semimetric for the points indexed by i and j.

In Euclidean space, as a consequence of the *law of cosines*

$$d(x_i, x_j)^2 = d_(x_0, x_i)^2 + d_(x_0, x_j)^2 - 2\langle x_i, x_j \rangle \tag{A.3}$$

in the plane containing each triple x_0, x_i, x_j of points, the metric matrix corresponds to the matrix of scalar products relative to the reference point x_0, with entries $M_{ij} = \langle x_i - x_0, x_j - x_0 \rangle$. It is also clear that the Gram matrix is positive semidefinite; in fact, that each positive semidefinite matrix can be realized as the Gram matrix of a set of vectors. This characterization carries over to the metric matrix, which solves the subset problem for Euclidean spaces:

Theorem 4 (Havel et al. (1983))**.** A configuration of $N+1$ points in a semimetric space is irreducibly embeddable in E_n, for some $n \leq N$, if and only if the corresponding metric matrix from any point is positive semidefinite of rank n. The eigenvalues of this matrix are then the (second) moments of the distribution of points along the n principal coordinate axes, and the eigenvectors, scaled by the square-roots of the corresponding eigenvalues, are the principal coordinate axes of the Euclidean configuration.

Proof. If the points are irreducibly embeddable in E_n, let $(x_0, \ldots, x_N)$ (where $x_i \in E_n$) be any family of vectors that represent them. The vectors x_i are then necessarily linearly independent. The metric matrix (with respect to the 0-th point, without loss of generality) is equal to the Gram matrix of the family $(x_1 - x_0, \ldots, x_N - x_0)$ in E_n, thus positive semidefinite and of rank n (since linear independence does not change under translation). The statement about the principal axes and the eigenvalues follows from the well-known identification of covariances with scalar products (Rodgers and Nicewander, 1988), such that the eigendecomposition of the Gram matrix defines the principal axes.

Conversely, if the $(N \times N)$ metric matrix M (with respect to the 0-th point, without loss of generality) is positive semidefinite of rank n, it can be diagonalized by an orthogonal transformation Y:

$$\Lambda = Y^t M Y. \tag{A.4}$$

The matrix Λ contains n positive eigenvalues and $N - n$ zeros on the diagonal (ordered by decreasing size, without loss of generality), and scaling the eigenvectors by their roots, a matrix $X = \Lambda^{1/2} Y$ is obtained such that $M = X^t X$. The columns of X are the coordinates of the N original points in E_n, centered on the 0-th point (at the origin). It is clear that the eigenvectors define the principal axes of X. □

This theorem solves the embedding problem for a finite set of points. The reference point is identified with the origin of the Euclidean space, and the coordinates of the points are uniquely reconstructed up to symmetries of the eigenspaces (reflections for eigenvalues with multiplicity one, subspace rotations for eigenvalues with larger multiplicities). In practice, these remaining degrees of freedom are fixed by the details of the numerical method used to diagonalize the metric matrix. It is also customary to choose the center of mass as reference point. A simple calculation shows how to obtain the corresponding metric matrix.

Theorem 5 (Havel et al. (1983))**.** The distance to the center of mass of each point i of a configuration of N points in a Euclidean space is given in terms of the remaining distances by

$$d_{0i}^2 = \frac{1}{N} \sum_{j=1}^{N} d_{ij}^2 - \frac{1}{N^2} \sum_{k>j}^{N} d_{jk}^2. \tag{A.5}$$

Let $1_N = (1, 1, \ldots, 1)^t$ be the $(N \times 1)$-vector consisting of ones. Define the *centering operator* $J = I - \frac{1}{N} 1_N 1_N^t$. A short calculation shows that the corresponding metric matrix is obtained by its action on the matrix of squared distances D^2 (with entries $D_{ij}^2 = d_{ij}^2$) of a given family of N points,

$$M = -\frac{1}{2} J D^2 J^t. \tag{A.6}$$

This operation is usually called *double-centering*. In Section A.2 it will be used to derive representations of reduced dimensionality $n \ll N$ from a given set of distances between N points.

For completeness, we end this section with an important result that characterizes embeddability of a space in terms of finite subsets.

Definition 13. A semimetric space T has *congruence order* k with respect to a class $\mathcal{S}$ of spaces provided each space $S \in \mathcal{S}$ is embeddable in T whenever any k-subset $\{x_0, \ldots, x_{k-1} \mid x_i \in S\}$ has that property.

Theorem 6 (Havel et al. (1983))**.** The Euclidean space E_n has congruence order $n+3$ with respect to the class of all semimetric spaces.

In fact, an even stronger property holds:

Theorem 7. A semimetric space S is irreducibly embeddable in E_n if S contains a $(n+1)$-set of points irreducibly embeddable in E_n such that every $(n+3)$-subset of S containing it is embeddable in E_n.

A.2 Multidimensional scaling

The previous section discussed when points with given distances can be realized by an embedding in some Euclidean space. In practice, we are rarely presented with this ideal situation and distances are usually contaminated by noise and discretized, and we cannot expect to find zero eigenvalues numerically. Moreover, it is often *a priori* unclear whether a set of measured distances admits a Euclidean representation at all. If this were impossible, negative eigenvalues will occur in the diagonalization of the metric matrix. Since these can also arise by numerical instabilities and errors in the distances, it can be difficult to decide whether a Euclidean representation is warranted.

The techniques of *multidimensional scaling* therefore focus on the reduction of dimension, and diagnostic measures are used to quantify the goodness of reconstruction.

Similar to principal component analysis, the reduction of dimension is achieved by restricting to the first $n \leq N$ principal axes in Theorem 4. We need to distinguish between the distances actually measured between all N systems, represented by a $(N \times N)$ matrix of squared distances D^2, and the Euclidean distances of a point configuration reconstructed to represent them, represented by a $(N \times N)$ matrix of squared distances Δ^2. Recall that the Frobenius norm of a matrix A is the root sum-of-squares,

$$||A|| = \left(\sum_{ij} |A_{ij}|^2 \right)^{1/2} . \tag{A.7}$$

Box 11. Why reconstruct distances in Euclidean space?

The alternative would be to consider reconstructions in more general metric spaces, e.g., spaces endowed with a Minkowski norm, or to consider nonmetric reconstructions, where the order relations between the distances are preserved as much as possible. In fact, there are good reasons why we only consider reconstructions of points in Euclidean space here:

- The connection between Euclidean norm and scalar products:
  Since Euclidean norm is a quadratic form, we can transform distances into scalar products. These we can consider values of a kernel function, and pattern analysis by kernel methods becomes possible.
- The efficiency of metric multidimensional scaling:
  Metric solutions are easy to calculate by linear algebra.
- The intuitiveness of Euclidean space:
  Euclidean space is simply the space with which we are most familiar with.

Of course, Euclidean distance has additional beneficial properties, e.g., invariance under rotations.

It induces a distance $d(A, B) = ||A - B||$ between two matrices.

Definition 14. The (*raw*) *stress* of a reconstructed configuration is

$$\sigma_{\mathrm{r}}(D^2, \Delta^2) = \frac{1}{2}||D^2 - \Delta^2||^2 = \frac{1}{2}\sum_{ij}\left(D_{ij}^2 - \Delta_{ij}^2\right)^2. \tag{A.8}$$

In the specific context of classical multidimensional scaling raw stress it is also known as the *strain* of a configuration.

Theorem 8 (Gower (1966),Havel et al. (1983))**.** The $(N \times N)$ symmetric matrix of rank n that best approximates any given $(N \times N)$ symmetric matrix of higher rank, in the sense of minimizing the Frobenius distance, is obtained by setting all but the n eigenvalues of largest magnitude to zero (and transforming back).

Recall the eigendecomposition $Y\Lambda Y^t = M$ (A.4), where Λ is a diagonal matrix of eigenvalues sorted by decreasing value, and Y is an orthogonal matrix whose rows contain the respective eigenvectors. Let Λ_n be the diagonal $(n \times n)$ matrix that contains only the largest $n \leq N$ eigenvalues of Λ, and Y_n be the matrix consisting of the first k columns of Y. Then the $(N \times n)$ coordinate matrix of *classical* (or *metric*) *multidimensional scaling* is given by $X_n = Y_n \Lambda_n^{1/2}$. Note that we have assumed here that the magnitude of negative eigenvalues is smaller than the magnitude of the n-th largest (positive) eigenvalue, i.e., we have assumed that errors and misrepresentations of distances are relatively small.

This representation of distances in Euclidean space minimizes the strain and leads to a nested solution: The coordinates in X_{n-1} are the same as the first $n-1$ coordinates of X_n (up to symmetries of the eigenspaces). It is called the *functional* or *behavior representation* of the distances Δ.

A.2.1 Diagnostic measures and distortions

The raw stress (A.8) has the disadvantage that it depends on the global scale of the distances. The following "badness-of-fit" measure is a scale-invariant diagnostic that quantifies the fraction of the sum-of-squares misrepresentation error that is not accounted for by the distances.

Definition 15 (Borg and Groenen (2005))**.** The *normalized stress* of a reconstructed configuration is

$$\sigma_n(D^2, \Delta^2) = \frac{\sum_{ij} \left(D_{ij}^2 - \Delta_{ij}^2\right)^2}{\sum_{ij} D_{ij}^2}. \tag{A.9}$$

The value of $1-\sigma_n(D^2, \Delta^2)$ is the fraction of distances explained in the Euclidean configuration, i.e., a *coefficient of determination*. Being a global statistic, σ_n is sensitive to outliers, i.e., points with an unusually large misrepresentation error. These can be identified by assessing the local misrepresentation error, and the following two diagnostic measures accomplish this.

Definition 16. The *Shepard diagram* of a reconstructed configuration is the diagram obtained by plotting the $N(N-1)/2$ distances Δ_{ij} of the Euclidean configuration against the measured distances D_{ij}. The (*normalized*) *maximal misrepresentation error* is given by

$$\sigma_{\max} = \frac{\max_{ij} \left(D_{ij}^2 - \Delta_{ij}^2\right)^2}{\frac{1}{N^2} \sum_{ij} D_{ij}^2}. \tag{A.10}$$

Definition 17. The (*normalized*) *stress per point* of the i-th point in a reconstructed configuration, consisting of N points, is given by

$$\sigma_n^i(D^2, \Delta^2) = \frac{\frac{1}{N} \sum_j \left(D_{ij}^2 - \Delta_{ij}^2\right)^2}{\sum_{ij} D_{ij}^2}. \tag{A.11}$$

Whereas the Shepard diagram visualizes the goodness-of-fit of all distances and can be useful to detect anisotropic distortions in the representation, the stress per point allows to detect suspect points or outliers that should be studied more closely. Raw stress per point, defined as in (A.11) but without the normalization in the denominator, can be conveniently visualized in a reconstructed configuration by plotting circles around each point, with area equal to the average stress of each point.

Note that the definitions have been given for symmetric distance matrices; in the case of (small) asymmetries these need to be changed accordingly.

We conclude this overview of the most important diagnostic measures with two examples.

Example 6. Figure A.1 shows three two-dimensional reconstructions of $N = 50$ points randomly distributed along the unit circle. In the left panel the configuration was obtained by classical multidimensional scaling when the distance matrix was calculated from Euclidean distances. Circles were used to depict the values of raw stress per point. The reconstruction is almost perfect, with misrepresentation errors on the order of the numerical accuracy, i.e., with $\sigma_{\max} \approx 10^{-34}$. This is reflected in the Shepard diagram (left panel of Figure A.2), which shows an almost diagonal line.

When the distance matrix is calculated from geodetic distances (Example 4), misrepresentation errors are introduced. The corresponding configuration is shown in the middle panel of Figure A.1. Stress per point is distributed relatively evenly among all points, with the largest errors accruing where the least points were present, and accounts for about 2 percent of the sum-of-square error ($\sigma_n \approx 0.02$). The Shepard diagram (middle panel of Figure A.2) shows that most distances are slightly overrepresented, whereas a few of the largest distances are underestimated. Note that both eigenvalues were positive (not shown). Changing the reconstruction dimension does also not allow for much leeway in improving the reconstruction. In one dimension the misrepresentation error is very large ($\sigma_n \approx 0.30$), whereas for larger dimensions it is also slightly larger than in two dimensions (left panel of Figure A.3, solid curve). For dimensions above about $N/2$, the first negative eigenvalue is encountered.

The right panels of Figure A.1 and Figure A.2 show results for a reconstruction from Euclidean distances that were contaminated with noise (normal, with unit variance). The misrepresentation error is again distributed relatively evenly, but the shape of the configuration has seriously deteriorated due to the large amount of noise. Its influence can be seen in the Shepard diagram, which shows that errors in the distances are distributed randomly. The dependence on reconstruction dimension (Figure A.3, grey curve) is not qualitatively changed, only shifted to larger errors.

Example 7. A different example is provided by the reconstruction of a torus, shown in Figure A.4. Since the line element of the standard torus, embedded as a two-dimensional surface in three dimensions, can only be evaluated numerically, we resort to the simpler representation of the torus as the quotient of the plane under the identification $(x, y) \sim (x + 2\pi, y) \sim (x, y + 2\pi)$. The torus $T^2 \simeq S^1 \times S^1$ is then identified by a square with opposite boundaries identified. This is a natural representation

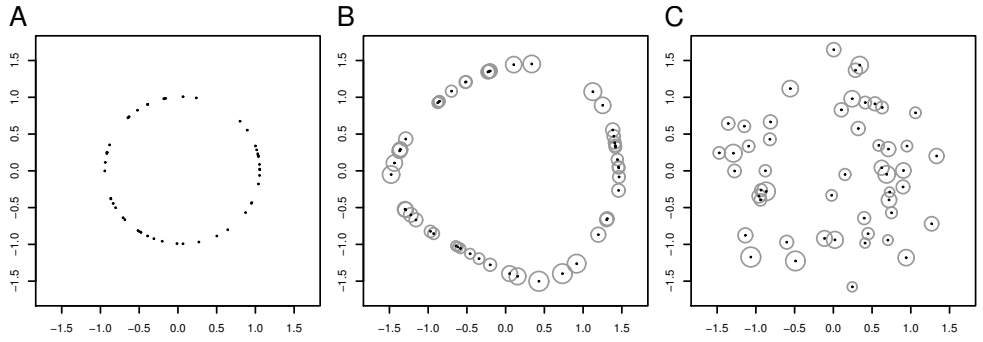

Figure A.1: Reconstruction of the one-dimensional circle S^1 by classical multidimensional scaling from $N = 50$ random samples. A: S^1 with the induced Euclidean metric. B: S^1 with its intrinsic, geodesic metric. C: S^1 with the induced metric, but Gaussian noise (unit variance) added to the distances. Radius of circles indicates (raw) stress-per-point.

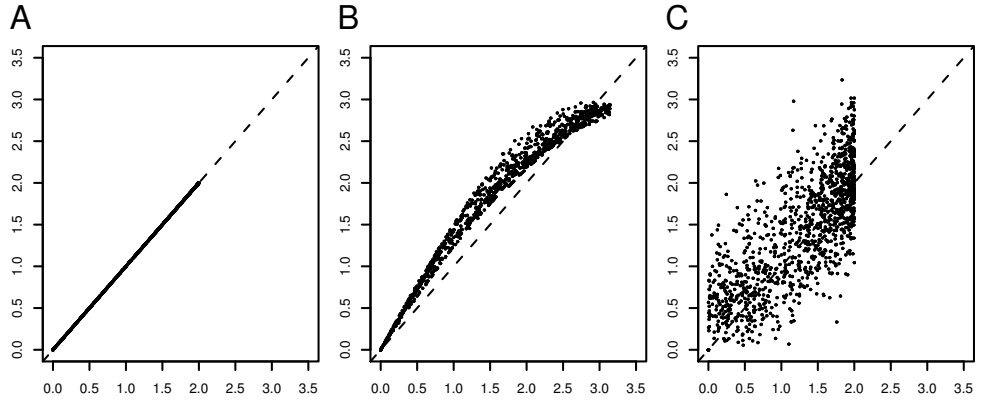

Figure A.2: Shepard diagrams for the reconstructions in Fig. A.1, depicting distances in the reconstructed configuration (vertical axis) against original distances (horizontal axis). A: S^1 with induced Euclidean metric. B: S^1 with intrinsic, geodesic metric. C: S^1 with induced metric under Gaussian noise.

for two simultaneously measured phases, with geodetic distance

$$\left((\min(|y_1 - x_1|, 2\pi - |y_1 - x_1|))^2 + (\min(|y_2 - x_2|, 2\pi - |y_2 - x_2|))^2 \right)^{1/2} \quad \text{(A.12)}$$

between two points $(x_1, x_2), (y_1, y_2) \in T^2$. The left panel of Figure A.4 shows the reconstructed configuration for Euclidean distances, the middle panel the configuration for the geodesic distance, and the right panel was obtained for Euclidean

distances under random noise (normal, unit variance).

The two-dimensional configuration of the geodesic distances approximates a square, with points in its interior exhibiting the largest misrepresentation error. Globally, about 15 percent of the distances cannot be accounted for in this representation ($\sigma_n \approx 0.15$), which drops to a mere 2 percent if the samples are reconstructed in four dimensions. The systemic distortions in the two-dimensional case can be clearly seen in the Shepard diagram (middle panel of Figure A.5), whereas a four-dimensional reconstruction closely approaches the original distances (right panel). The right panel of Figure A.3 shows the normalized stress against the reconstruction dimension (solid curve). The minimal stress is achieved for about four dimensions, and then rises again slightly due to numerical errors.

Reconstruction from the Euclidean distances under noise leads to similar changes as in Example 6. The misrepresentation error shows the same qualitative behavior with respect to the reconstruction dimensionality, only shifted to a higher level (right panel in Figure A.3, grey curve).

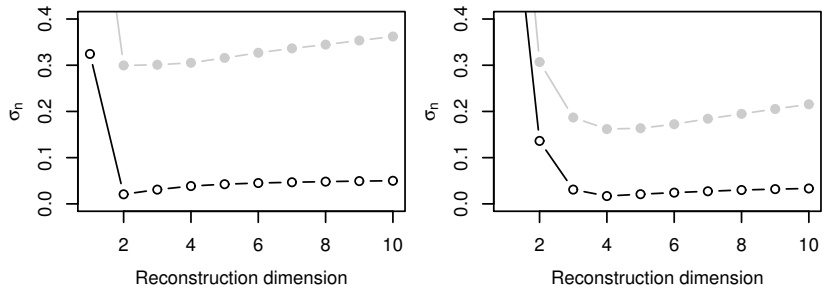

Figure A.3: Normalized stress for different reconstruction dimensions for distances without (dark) and under Gaussian noise (gray). A: S^1 with intrinsic, geodesic metric. B: T^2 with intrinsic, geodesic metric. The configurations were reconstructed from $N = 50$ random points each.

These examples show that stress diagrams as in Figure A.3 can be used to decide which dimension is optimal for the reconstruction from a given distance matrix, and whether misrepresentation errors might be caused by random noise or by systematic distortions due to an intrinsically different geometry. Whereas the effects of noise cannot be reduced by increasing the reconstruction dimension, this is possible (to a great extent) for non-Euclidean distances.

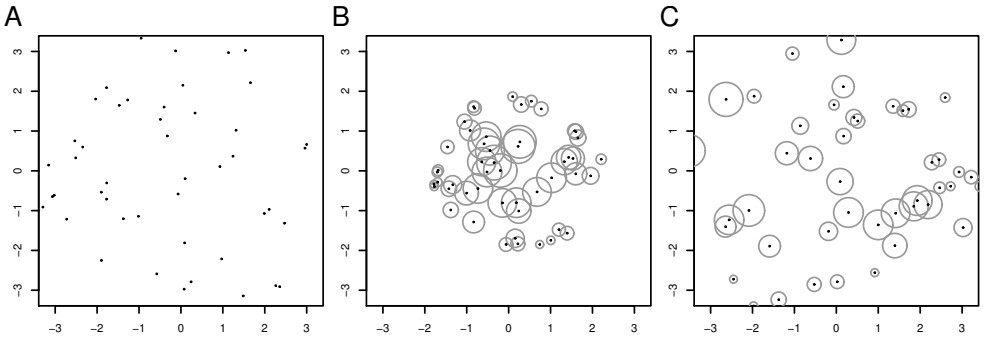

Figure A.4: Reconstruction of the two-dimensional torus T^2 by classical multidimensional scaling from $N = 50$ random samples. A: T^2 with the induced Euclidean metric. B: T^2 with its intrinsic, geodesic metric. C: T^2 with induced metric, but Gaussian noise (unit variance) added to the distances. Radius of circles indicates (raw) stress-per-point.

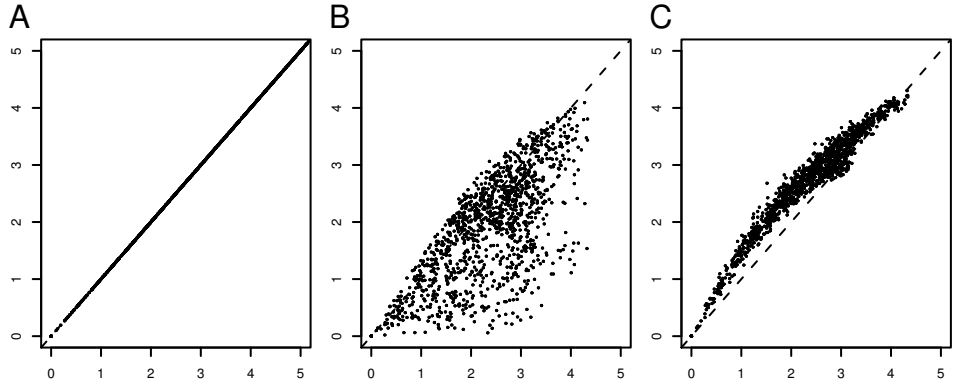

Figure A.5: Shepard diagrams for reconstruction of the torus T^2. A: T^2 with the Euclidean metric in 2D. B: T^2 with its intrinsic metric in 2D. C: T^2 with its intrinsic metric in 4D.

A.2.2 Violations of metric properties and bootstrapping

Different from the effect of non-Euclidean geometries, the influence of noise in measured distances can and usually does destroy the metric properties, for sufficiently large noise levels. Such violations of metric properties are also interesting conceptually. Many bivariate measures commonly used (e.g., in electrophysiology, see Section 6.3) do not respect metric properties, and it is instructive to consider what effect this does have on dissimilarity matrices, where we use the word *dissimilarity* to denote a bivariate measure that does not necessarily fulfill metric properties. Moreover,

these violations occur when the distances are resampled (see below) to reduce bias in measurements, or to improve the speed of computations.

Reflexivity

Reflexivity is the property that the *self-distances* $d(x, x)$ are zero. Conceptually, this embodies the notion of identity, and measures that do not result in reflexive dissimilarities are problematic. The reason is, of course, that such dissimilarities cannot be interpreted in terms of points, but would need to be realized as extended objects — if this is consistently possible at all. Therefore, reflexivity should not be destroyed by even the effect of measurement noise, but since the numerical calculation of distances can introduce round-off errors, reflexivity can be violated in practice. The usual solution is to simply force the diagonal terms of dissimilarity matrices to zero, but there is a situation in which self-distances naturally occur and contain valuable information.

Up to now we have been assuming implicitly that measurements made on systems are ideal, in the sense that the system's behavior is captured in its totality. In practice this is barely the case, however, since measurements are finite and should always be considered approximations of a system. If we consider generalized measurements (Section 1.1) that result in probability measures, these measures are empirical and might differ from the true probability measure that would be obtained under ideal conditions. The variability inherent in these probability measures can be estimated, however, by bootstrapping the empirical measures. Thereby, a random sample (with replacement) is drawn from the measure under consideration, usually of the same size as the original observations on which that measure is based, and interpreted as another (empirical) probability measure. Repeating this process a number of times, a set of probability measures is obtained that represent the variability of the unknown, underlying probability measure. Although this is not an unbiased method, since it takes an empirical realization of a probability measure as its point of departure, such *bootstrapping* obtains an approximation of the original measure that is valid to a great extent, i.e., with largely reduced statistical error (Efron, 1981; Davison and Hinkley, 1997).

A second advantage of resampling the measurements is that one can choose a *smaller* sample size. Although this invariably increases the variance, and a larger number of bootstrap replications is needed to obtain the same reduction in bias, it may speed up computations enormously. We will therefore routinely use this device for the involved calculations of the optimal transportation distances (confer Sections 2.5,3.6.3,6.4). In practice, it will result in not a single distance between two systems, but rather in a set of bootstrap replicates of numerical distances. We will then take the mean of these as an estimate of the "true" distance between two systems.

A special case occurs with the self-distances $d(x, x)$, however, since distances can only be nonnegative. The magnitude of the self-distances under resampling is there-

fore an indication of the *numerical resolution* of our distance measure. Systems that are closer than the average self-distance cannot be resolved properly and appear to be distinct in actual calculations, and distances between two distinct systems should be considered to be influenced by statistical errors of the same order. This state of affairs can also not be remedied by subtracting a constant from all distances, since this might destroy the triangle inequality (see below). It is important to keep this qualification in mind.

Symmetry

Violations of symmetry, where $d(x, y) \neq d(y, x)$ can arise by noise or resampling error (see above), but might also indicate directionality effects. These again lead to representations of systems as extended objects (confer Figure 6.1 in Chapter 6), which is undesirable for further analysis. In the first case, the accepted method is to simply average out the asymmetries. Given a dissimilarity matrix D, it can be decomposed into a symmetric part $S = \frac{1}{2}(D + D^t)$ and an antisymmetric part $A = \frac{1}{2}(D - D^t)$, such that

$$D = A + S. \tag{A.13}$$

The symmetric part S is then used as an estimate of the underlying true distances. However, if the antisymmetric part A is not of negligible size relative to S, this hints at the influence of directionality. General dissimilarity measures (see Section 6.3 for examples) might measure the flow of information between two systems, or the strength of influence one system exerts upon another, which are genuinely asymmetric effects. Due to the decomposition (A.13), however, it is possible to treat the symmetric and antisymmetric part independently. This problem is therefore alleviated to a great extent. Treatment of the antisymmetric part is further discussed in Section 6.2.2, for completeness.

Triangle inequality

The triangle inequality is basic to a representation in Euclidean space. As before, violations of this property hint at directionality effects and suggest that systems might need to be represented by extended objects (Figure 6.1). It is the most common violation for many dissimilarity measures, since reflexivity and symmetry are often easy to accomplish, whereas the triangle inequality is a nontrivial geometric constraint. Violations of the triangle inequality are therefore important conceptually, since they suggest that a geometric representation might be unsuitable. If the triangle inequality is not fulfilled, it is not possible to compare more than two systems in a sensible (multivariate) way without introducing additional, spurious effects that are undesirable. However, adding a constant $c > 0$ to all distances (from a finite set), the triangle

Box 12. How to publish distance matrices?
When publishing research results obtained from or with (measured) distance matrices, the following information should ideally be also given:

- Was the triangle inequality fulfilled? If not, how large was the maximal violation?
- Were all eigenvalues nonnegative? If not, how large was the negative eigenvalue of largest magnitude? How many positive and negative eigenvalues were there?
- Were all diagonal entries zero? If not, how large was the largest diagonal element?

inequality can always be enforced, since for large enough c the equation

$$d(x, y) \leq d(x, z) + d(z, y) + c, \tag{A.14}$$

will be fulfilled.

Bootrapping the distances can break the triangle inequality, and to ensure multivariate comparability we will use the smallest possible constant in (A.14) to fix this, if needed. Of course such violations of the triangle inequality need to be reported.

A.3 Statistical inference

Although multidimensional scaling has a long history, statistical inference about reconstructed point configurations is seldomly encountered in the literature (but see (Anderson and Robinson, 2003)). In this section we will therefore advocate and describe the main methods of statistical analysis used in the rest of this thesis. The starting point for most methods considered here is the reconstructed point configuration of N systems, i.e., their representation as N vectors in a Euclidean space E_n, where $n \leq N$. We call this the *behavior* or *functional space* of the systems. This representation allows for the use of multivariate analysis methods. We are particularly interested in the task of classifying distinct groups of systems. More precisely, we will consider *supervised* classification, in which the true group assignments of all points are assumed to be known perfectly. Let there be $g \in \mathbb{N}$ distinct groups $G_1, \ldots, G_g$, and let the true group label of a point $x \in E_n$ be given by an indicator variable $z = (z_1, \ldots, z_g)$, such that $z_i = 1$ if $x \in G_i$ and $z_i = 0$ if $x \notin G_i$. Let $(x_1, \ldots, x_N)$ denote the points from E_n representing the N systems under study. Denote by $\lambda = (\lambda_1, \ldots, \lambda_N)$ the labelling, such that $g_i = k$ if and only if $z_k = 1$ for the point x_i.

A.3.1 Multiple response permutation testing

The first question about the representation $(x_1, \ldots, x_N)$ of N systems from a priori known g groups is whether this representation does carry information on the group structure, and to what extent.

To assess this, we employ a permutation hypothesis test. Under the null hypothesis of no difference with regard to group association, the labelling λ can be permuted randomly. As a test statistic, we will use the weighted mean of within-group means of pairwise distances among groups. Let $(N_1, \ldots, N_g)$ be the sizes of the g groups, then this is given by

$$\delta_\lambda = \sum_{k=1}^{g} \frac{N_k / \sum_l N_l}{N_k(N_k-1)/2} \sum_{\substack{i<j \\ \lambda_i=\lambda_j=k}} D_{ij}, \tag{A.15}$$

conditional on the group labelling λ and the pairwise distances D_{ij}. Under the null hypothesis the test statistic δ will be invariant under permutations $\pi\lambda$ of the group labelling, and the significance probability of this test is the fraction of values of $\delta_{\pi\lambda}$ obtained that are smaller than the value δ_λ for the original labelling λ:

$$p = \frac{\#\{\delta_{\pi\lambda} < \delta_\lambda\}}{m+1}, \tag{A.16}$$

where m is the number of permutations. Considering all distinct $\binom{N!}{N_1!N_2!\cdots N_g!}$ permutations will be often infeasible, so the value of p is estimated by considering a large enough number of random permutations (typically on the order of 10^5 or larger). This test is called a *multiple response permutation procedure* (MRPP).

Similar as in analysis of variance, the *chance-corrected within-group agreement*

$$A = 1 - \frac{\delta_\lambda}{\mathbb{E}\delta_{\pi\lambda}}, \tag{A.17}$$

where $\mathbb{E}\delta_{\pi\lambda}$ is approximated by the mean of δ under all permutations π considered, is a coefficient of determination that quantifies how much of the group structure is "explained" by the distances.

It is important to stress the difference between these two diagnostic measures. Whereas a small p-value indicates that the structure of the distances is significantly dependent on the group association, it might still be the case (and often will be in practice) that the size of this effect, as measured by A, is rather small. To this extent, the value of A indicates the signal-to-noise ratio of the distances.

The MRPP test is calculated from the distance information only, and can therefore be performed for both the original distance matrix D, and additionally for the distance matrix Δ of the reconstructed points $(x_1, \ldots, x_N)$ that is subject to misrep-

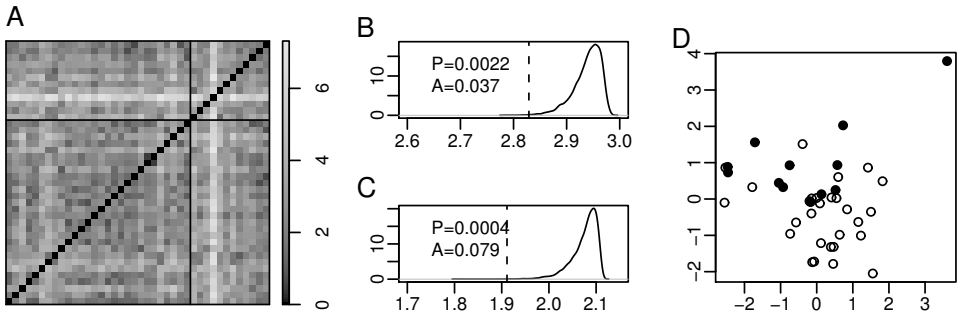

Figure A.6: Example: MRPP test in the Pima dataset. A: Distance matrix between all subjects ($N_1 = 28$ with no diabetes, $N_2 = 12$ with diabetes; separated by dark lines). B: Distribution of MRPP statistic δ for the distances in (A). C: Distribution of δ for the distances of the reconstructed two-dimensional configuration. D: Reconstructed configuration (dark circles: diabetes, open circles: no diabetes).

resentation errors. Both of these tests are of value in practice. The first shows the extent to which measured distances capture the group structure, the second shows how much of this is still preserved under reconstruction. Thereby, it can be judged whether a Euclidean representation is adequate.

Example 8. Let us illustrate the MRPP test with the Pima dataset from the R package MASS (Venables and Ripley, 1999). This dataset contains information collected by the US National Institute of Diabetes and Digestive and Kidney Diseases on diabetes in women of Pima Indian heritage. We will use five variables from the first 40 entries of the training data Pima.tr: plasma glucose concentration, blood pressure, body-mass-index, diabetes pedigree function, and age. The outcome (diabetes or not) is known, with $N_1 = 28$ subjects showing no symptoms of diabetes, and $N_2 = 12$ being diagnosed with diabetes. Distances between subjects were calculated by first centering and scaling the predictor variables to unit variance, and then taking Euclidean distance in the five-dimensional space. Figure A.6 shows the distance matrix, the reconstructed point configuration in two-dimensions, and the distribution of the MRPP statistic δ for both sets of distances. Interestingly, the within-group agreement A of the reconstructed configuration is twice as large as for the original distances, indicating that dimension reduction can improve the classificatory contrast.

A.3.2 Discriminant analysis

In the distance-based approach, discrimination of systems is achieved from their representation in Euclidean space E_n. We advocate the use of robust and conceptually

simple analysis methods, and have therefore chosen *canonical discriminant analysis* as our method of choice for the classification of systems. Canonical discriminant functions are *linear* combinations of variables that best separate the mean vectors of two or more groups of multivariate observations relative to the within-group variance. They are variously known as canonical variates or discriminant coordinates in the literature and generalize the linear discriminant analysis of Fisher (1936) (for the case of $g = 2$ groups). For this reason, the term *linear discriminant analysis* (LDA) is also used for the analysis described here.

Let B_0 be the covariance matrix of the group-wise distributions,

$$B_0 = \frac{1}{g-1} \sum_{i=1}^{g} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^t, \tag{A.18}$$

where $\bar{\mu} = \frac{1}{g} \sum_{i=1}^{g} \mu_i$ is the pooled mean[1]. In practice this will be approximated by the sample between-groups covariance matrix on $g - 1$ degrees of freedoms,

$$B = \frac{1}{g-1} \sum_{i=1}^{g} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^t, \tag{A.19}$$

where $\bar{x}_i$ is the sample mean of the i-th group, and $\bar{x} = \frac{1}{g} \sum_{i=1}^{g} \bar{x}_i = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the mean over the whole sample. The matrix B_0 (and therefore also B) is of rank $b_0 \leq g - 1$, where $b_0 = g - 1$ if and only if the group-means $\mu_1, \ldots, \mu_g$ are linearly independent.

Let Σ_0 be the within-group covariance matrix. The main assumption here is that this is equal for each group (*homoscedasticity* assumption), and it is estimated by the pooled within-group sample covariance matrix. Let $X = [x_1 \cdots x_N]^t$ be the $(N \times n)$ matrix of coordinates and let M be the $(g \times n)$ matrix of group means. Define the $(N \times g)$ matrix of group indicators Z by $Z_{ij} = 1$ if $x_i \in G_i$, and $Z_{ij} = 0$ otherwise. Then

$$\Sigma = \frac{1}{N-g}(X - ZM)^t(X - ZM) \quad \text{and} \quad B = \frac{1}{g-1}(ZM - 1_N \bar{x}^t)^t(ZM - 1_N \bar{x}^t) \tag{A.20}$$

are the two sample covariance matrices in matrix notation.

There exist $r = \min(n, g - 1)$ canonical variates (discriminant "scores"), and for the coordinates in X these are defined by

$$S = XA, \tag{A.21}$$

where $A = [a_1 \cdots a_r]$ is a $(n \times r)$ *scaling* matrix, such that a_1 maximizes the ratio

[1] Using coordinates derived from double centering clearly $\bar{\mu} = 0$, but we prefer to exhibit the general case here.

(generalized Rayleigh quotient)

$$\frac{a_1^t B a_1}{a_1^t \Sigma a_1}. \tag{A.22}$$

The scaling acts on the right, since the coordinates X are in row-order. For $k = 2, \ldots, r$, the variate a_k maximizes the ratio (A.22) subject to the orthogonality constraint $a_k^t \Sigma_0 a_h = 0$ (for $h = 1, \ldots, k-1$). To compute A, choose a preliminary scaling XA_1 of the variables such that they have the identity as their within-group correlation matrix. This is achieved by taking the principal components with respect to Σ, normalized by their variance. On the rescaled variables XA_1, the maximization of (A.22) reduces to the maximization of $a^T B a$ under the constraint $||a|| = 1$. The latter is solved by taking a to be the (normalized) eigenvector of B corresponding to the largest eigenvalue. The eigenvectors corresponding to the next $g - 2$ largest eigenvalues supply the other $g-2$ canonical variates, which are orthogonal as required. In practive we use the lda function in the standard R package MASS (Venables and Ripley, 1999, Ch. 11.4), which employs singular value decomposition (SVD) to find the eigenvectors. Note that this code, as is standard in multivariate analysis, rescales the different coordinates in the reconstruction space E_n to unit variance prior to calculation of the canonical variates. This is one reason why cross-validation (Section A.3.3) is so important: This standardization allows coordinates which contribute very little to the distances (between systems) to influence the discrimination on equal terms with coordinates that contribute much more to the distances. For small sample sizes N the discrimination could then be based on fitting the "noise" in the distances, rather than the "signal".

The allocation of a vector x to a group G_i can be achieved in a number of ways. The simplest way is to choose the group to which the point x has smallest distance. However, this distance should consider the statistical properties of the underlying group conditional distribution, i.e., its spread around its center point. It is therefore common to measure the distance between a vector x and the i-th group, with mean μ_i and covariance matrix Σ, by their Mahalanobis distance,

$$\left((x - \mu_i)^t \Sigma^{-1} (x - \mu_i)\right)^{1/2}. \tag{A.23}$$

If we *assume* that the distribution of the i-th class is multivariate normal with mean μ_i and covariance matrix Σ, then this corresponds to *maximum a posteriori* classification, up to the prior distribution. In detail, the Bayes rule that minimizes the overall misclassification error (under equal misallocation costs) is given by

$$r(x) = i \quad \text{if} \quad \pi_i f_i(x) \geq \pi_j f_j(x) \quad (j = 1, \ldots, g; j \neq i), \tag{A.24}$$

where $\pi = (\pi_1, \ldots, \pi_g)$ is the prior distribution of groups and f_i is the group-conditional probability density of the i-th group. The prior distribution is in practice ap-

proximated by the relative group sizes, and $f_i(x) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i))$. It is more convenient to work in terms of the log-likelihood, which is given by

$$L_i = -\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i) + \log|\Sigma| + \log\pi_i. \tag{A.25}$$

Subtracting the constant terms, this simplifies to the maximalization of

$$L_i = x^t\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + \log\pi_i. \tag{A.26}$$

In the coordinate system defined by the canonical covariates, the within-group variance is trivial, such that on these variables the Mahalanobis distance is just $||x-\mu_i||$. The log-likelihood further simplifies to

$$L_i = x^t\mu_i - \frac{1}{2}||\mu_i||^2 + \log\pi_i. \tag{A.27}$$

The a posteriori probabilities of group membership are then given by

$$\frac{\exp(-(x^t\mu_i - \min_j x^t\mu_j))}{\sum_k \exp(-(x^t\mu_k - \min_j x^t\mu_j))}. \tag{A.28}$$

Let us summarize. The canonical variates are defined in terms of seoond-order statistical properties (means and covariances) between and within groups of (normalized) coordinates. The main assumption is that the covariances for each group are equal (homoscedasticity assumption). In particular, it is not needed to assume that the group conditional distributions are multivariate normal. Under this assumption, however, the allocation rule (A.24) is optimal, if the total misallocation error is to be minimized. The reasons we routinely employ this normal based classification are summarized in Box 13.

A.3.3 Cross-validation and diagnostic measures in classification

For small to medium sized datasets encountered here, cross-validation of classification accuracies is achieved by a leave-one-out method. This proceeds in the following steps:

1. For the k-th sample point ($k = 1, \ldots, N$) we remove its distance information from the set of original distances D_{ij}, leading to a new $(N-1)$-by-$(N-1)$ matrix of squared distances $D^2_{(k)}$.

2. We reconstruct a Euclidean configuration $X_n^{(k)}$ in n dimensions by metric multidimensional scaling of $D^2_{(k)}$.

Box 13. Why use the homoscedastic normal-based allocation rule?

- Why parametric classification: Although non-parametric alternatives exist, these are much more involved and cannot be routinely used for small sample sizes.
- Why homoscedasticity: Estimation of multivariate covariance matrices is difficult for small sample sizes. The assumption of homoscedasticity allows to only estimate one covariance matrix in place of many, thereby improving the stability of the estimate.
- Why normal based allocation: The multivariate normal model is flexible and computationally efficient, and it is relatively robust. Even if the true distribution is not normal, its approximation by a normal distribution (second-order approximation) is often close, if the distribution has finite second moments and is not too irregular otherwise.

3. We train the classifier on $X_n^{(k)}$, i.e., we estimate the group means and covariance matrix from $X_n^{(k)}$.
4. We estimate the coordinates x' of the i-th sample point in the coordinate system defined by $X_n^{(k)}$ by minimizing an error criterion (Trosset and Priebe, 2008).
5. We predict the group membership of the coordinates x' by the normal-based rule. Additionally, we store the discriminant scores of x'.
6. The above is repeated for all N points. The total number of correct predictions results in the cross-validated accuracy.

The accuracy estimates obtained thereby are almost unbiased. The only parameter needed is the reconstruction dimension $n \leq N$. We will usually determine this by considering each possible choice of $1 \leq n \leq N'$ up to some maximum dimension $N' \leq N$ and choosing the dimension n' that maximizes the cross-validated classification accuracy. Note that this introduces a certain selection bias into the accuracies, but this cannot be avoided for small datasets, and should in fact be negligible.

The cross-validated discriminant scores obtained by the above method provide us with additional diagnostic information. Note however, that these scores are biased due to the different scaling invoked at each step. The main problem here is, that the geometric orientation of the discriminant functions can and will often be different for the distinct $X_n^{(k)}$. For two groups, the sign of the discriminant scores can change, but this problem can be largely avoided: Since the original group membership is known, discriminant scores with the wrong sign can be corrected. Thereby, only a slight bias occurs, as the origin of the coordinate system of the $X_n^{(k)}$ depends on the

points. The discriminant scores will therefore be slightly inaccurate and should be considered with care. As often in statistics, outliers in the data can lead to unexpected results, and it at this point where this could potentially happen.

In a classification task with two groups the classification is achieved by fixing a numerical threshold and predicting all scores to the left of it as *negatives*, and all scores to the right as *positives*. Varying the classification threshold, the number of correctly predicted negatives and positives will change. This can be conveniently visualized in a *receiver-operator-characteristic*, which allows to derive additional diagnostic measures (Hanley and McNeil, 1982).

Let TP denote the number of correctly predicted positives, let FP denote the number of incorrectly predicted positives, and likewise TN and FN for the negatives. The *true positive rate* (TPR) and the *false positive rate* (FPR) are defined by

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \text{and} \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \tag{A.29}$$

respectively. Note that $\mathrm{TP}+\mathrm{FN}$ is the number of positives (known a priori) in the dataset, and $\mathrm{FP}+\mathrm{TN}$ the number of negatives. In the context of a diagnostic test the true positive rate TPR is interpreted as the *sensitivity*, and $1-\mathrm{FPR}$ is interpreted as the *specificity*. The receiver-operator-characteristic depicts the relationship between TPR and FPR.

Example 9. For the Pima dataset of Example 8, classification results are shown in Figure A.7. Receiver-operator characteristics of both the original data (A) and its optimal Euclidean reconstruction (D) are given. The accuracies (both resubstiution and cross-validated) for the reconstruction indicate that resubstitution accuracies tend to overestimate the classification success (B, in gray) for larger reconstruction dimensions. The cross-validated accuracies (B, in black) result in a realistic picture, never rising above the accuracy 0.85 of the original data. Interestingly, for the optimal reconstruction in two dimensions (maximal accuracy), the cross-validated accuracy is almost identical to the resubstitution accuracy, as are the receiver-operator-characteristics (D). Again, this indicates that the distance-based classification can improve classification.

A.3.4 Combining classifiers

In some cases of interest there exists more than one type of measurements of a given family of systems and we will briefly discuss two situations here: (i) if more than one distance matrix is available, and (ii) if more than one classification rule is available.

The first case can arise naturally in the framework of optimal transportation distances (Chapter B), since these distances form a parametric family. Similar to the Minkowski distances $||x-y||_p = (\sum_i |x_i - y_i|^p)^{1/p}$, different distances (for distinct values of $p \geq 1$) stress slightly different aspects of the underlying geometry.

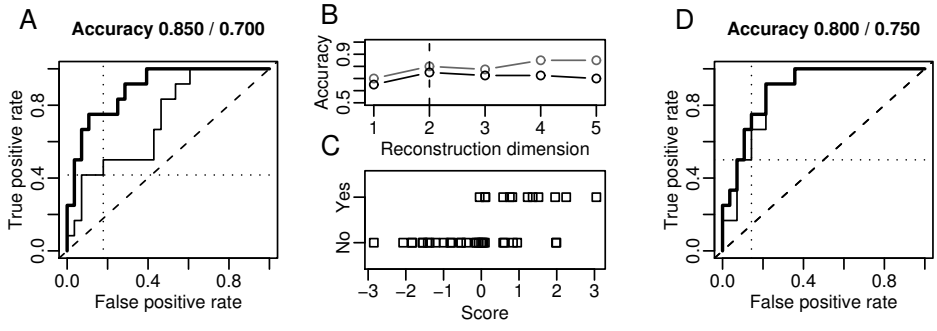

Figure A.7: Classification in the Pima dataset. A: Receiver-operator-characteristic for discriminating negatives (no diabetes) from positives (diabetes). Dark line: resubstitution accuracies. Light line: cross-validated accuracies. The optimal normal-based classification boundary is indicated (stippled lines), leading to the accuracies indicated (above plot). B: Accuracies (grey: resubstitution, dark: cross-validated) against reconstruction dimensions. C: Cross-validated discriminant scores for the optimal two-dimensional reconstruction. D: Corresponding receiver-operator-characteristic.

We encounter a different instance of this problem in Chapter 3, where two distinct time series are available for each subject. There, we will simply normalize both time series to zero mean and unit variance, and combine them into a vector-valued time series. This eventually leads to a multivariate probability distribution from which a single distance matrix is computed.

We recommend to combine distinct measurements into a single distance for practical reasons. Note that squared dissimilarities are additive in the reconstructed Euclidean space, and in the context of multidimensional scaling so-called *three-way scaling* exploits this property, allowing to weight the contributions of distinct distance matrices (Arabie et al., 1987). Since these methods are computationally involved, they will not be considered further here.

For the second situation, there exists a large literature on voting procedures that allow to combine distinct classifiers, and even optimal training rules for this meta-decision problem (Tax et al., 2000).