View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Serveur académique lausannois

Journal of Classification 28:297-314 (2011)
DOI: 10.1007/s00357-011-9092-x

On the Schoenberg Transformations in Data Analysis: Theory and Illustrations

François Bavaud

University of Lausanne, Switzerland

Abstract: The class of Schoenberg transformations, embedding Euclidean distances into higher dimensional Euclidean spaces, is presented, and derived from theorems on positive definite and conditionally negative definite matrices. Original results on the arc lengths, angles and curvature of the transformations are proposed, and visualized on artificial data sets by classical multidimensional scaling. A distance-based discriminant algorithm and a robust multidimensional centroid estimate illustrate the theory, closely connected to the Gaussian kernels of Machine Learning.

Keywords: Bernstein functions; Conditionally negative definite matrices; Discriminant analysis; Euclidean distances; Huygens principle; Isometric embedding; helix; Kernels; Menger curvature; Multidimensional scaling; Rectifiable curves; Robust centroids; Robust PCA.

1. Introduction

Schoenberg transformations are elementwise mappings of Euclidean distances into new Euclidean distances, embeddable in a higher dimensional space. Their potential in Data Analysis seems evident in view of the omnipresence of Euclidean dissimilarities in Multidimensional Scaling (MDS), Factor Analysis, Correspondence Analysis or Clustering. Yet, despite its respectable age (Schoenberg 1938a), the properties and the very existence of this class of transformations appear to be little known in the Data Analytic community.

The helpful suggestions of two anonymous reviewers are gratefully acknowledged.

Author's Address: F. Bavaud, Department of Computer Science and Mathematical Methods, Department of Geography, University of Lausanne, CH-1015 Lausanne, Switzerland, tel: +41-21-692-3022, e-mail: francois.bavaud@unil.ch.

Published online 7 October 2011

By contrast, non-linear embeddings of original data into higher dimensional feature spaces are familiar in the Machine Learning community, which however bases its formalism upon kernels, which are positive definite (p.d.) matrices, rather than on squared Euclidean distances, which are conditionally negative definite (c.n.d.) matrices with a null diagonal (Section 3.1).

Some aspects of the correspondence between p.d. and c.n.d. matrices are well-known in Data Analysis, and lie at the core of classical MDS (Theorems 1 and 2). Other aspects (Theorem 4), central to the derivation of Schoenberg transformations (Definition 2), are less well-known. Section 2 is a self-contained review of all those results, scattered in the literature, together with their proofs. Section 3 analyzes some of the general properties of Schoenberg transformations, and yields original results about angles, arc lengths and curvatures. Section 4 illustrates the non-linear and spectral properties of the transformations on two artificial data sets – the grid and the rod. Section 5 briefly illustrates data-analytic applications, namely distance-based discriminant analysis and robust centroid estimation. In conclusion, Section 6 proposes to revisit the Machine Learning formalism in terms of Euclidean distances, rather than in terms of kernels.

2. Definitions and Theorems

2.1 Preliminaries

Classical multidimensional scaling (MDS) (e.g. Borg and Groenen 1997) can be performed iff the eigenvalues of the so-called *matrix of scalar products* are non-negative. For concision, we shall refer to such a matrix as *positive definite* – instead of "positive semi-definite".

Vectors are column vectors. I denotes the identity matrix, and $\mathbf{1}$ the unit vector, whose components are all unity. Depending upon context, the "prime" either denotes the transpose of a matrix, or the derivative of a scalar function.

Definition 1. A real symmetric $n \times n$ matrix $C = (c_{ij})$ is said to be

- positive definite (p.d.) if $z'Cz = \sum_{ij}^{n} c_{ij} z_i z_j \geq 0$ for all vectors $z \in \mathbb{R}^n$
- conditionally negative definite (c.n.d) if $z'Cz = \sum_{ij}^{n} c_{ij} z_i z_j \leq 0$ for all $z \in \mathbb{R}^n$ such that $\sum_{i=1}^{n} z_i = 0$.

Consider a *signed distribution* a on n objects, that is a vector obeying $\sum_{i=1}^{n} a_i = 1$, where some components are possibly negative. Consider also the $n \times n$ *centering matrix* $H(a) = I - \mathbf{1}a'$, with components $\delta_{ij} - a_j$. Let

C be a symmetric $n \times n$ matrix, and define the matrix $B(a) = (B_{ij}(a))$ as

$$B(a) = -\frac{1}{2}\, H(a)\, C\, H'(a) \quad . \tag{1}$$

Theorem 1 (Young and Houseolder 1938; Schoenberg 1938b).
For any signed distribution a,

$$B(a) \quad \textit{is p.d.} \qquad \Leftrightarrow \qquad C \quad \textit{is c.n.d.}$$

Proof: First observe that if $B(a)$ is p.d., then $B(\tilde{a})$ is also p.d. for any other signed distribution $\tilde{a}$, in view of the identity $B(\tilde{a}) = H(\tilde{a})B(a)H'(\tilde{a})$, itself a consequence of $H(\tilde{a}) = H(\tilde{a})H(a)$. Also, for any z, $z'B(a)z = -\frac{1}{2}y'Cy$ where the vector $y = H'(a)z$ obeys $\sum_i y_i = 0$ for any z, showing "$\Leftarrow$". Finally, $y = H'(a)y$ whenever $\sum_i y_i = 0$, and hence $y'B(a)y = -\frac{1}{2}y'Cy$, thus demonstrating "$\Rightarrow$".
■

Theorem 2 (Classical MDS). *Let $C = (c_{ij})$ be a symmetric $n \times n$ matrix. Define the associated* null-diagonal *matrix $\hat{C} = (\hat{c}_{ij})$ as $\hat{c}_{ij} = c_{ij} - \frac{1}{2}c_{ii} - \frac{1}{2}c_{jj}$. Then*

$$B(a) = -\frac{1}{2}\, H(a)\, \hat{C}\, H'(a) \qquad \textit{and} \qquad \hat{c}_{ij} = B_{ii}(a) + B_{jj}(a) - 2B_{ij}(a)\,. \tag{2}$$

Moreover, C is c.n.d. iff $\hat{C}$ is c.n.d. In this case, the components $\hat{c}_{ij}$ are "isometrically embeddable in l_2", that is representable as squared Euclidean distances D_{ij} between n objects as

$$D_{ij} \equiv \hat{c}_{ij} = \sum_{\alpha=1}^{p} (x_{i\alpha} - x_{j\alpha})^2, \qquad\qquad i, j = 1, \ldots, n \tag{3}$$

where the object coordinates can be chosen as

$$x_{i\alpha} = \sqrt{\lambda_\alpha(a)}\, u_{i\alpha}(a), \tag{4}$$

where the λ_α are the diagonal components of the diagonal matrix $\Lambda(a)$ and $u_{i\alpha}(a)$ are the components of the orthogonal matrix $U(a)$ occurring in the spectral decomposition $B(a) = U(a)\Lambda(a)U'(a)$.

Proof: The first identity in (2) follows from $H(a)\mathbf{1} = 0$, and the second one from $B_{ii}(a) + B_{jj}(a) - 2B_{ij}(a) = c_{ij} - \frac{1}{2}c_{ii} - \frac{1}{2}c_{jj}$, itself a consequence of the form (1) $B_{ij}(a) = -\frac{1}{2}c_{ij} + \gamma_i + \gamma_j$ for some vector γ. The next assertion follows from $y'Cy = y'\hat{C}y$ whenever $\sum_i y_i = 0$, and identity (3) can be shown to amount to the second identity (2) by direct substitution.
■

On one hand, the p.d. nature of $B(a)$ (Theorem 1) makes its eigenvalues λ_α non-negative. One the other hand, $H'(a)a = 0$ yields $B(a)a = 0$. Hence, at least one eigenvalue is zero and $p \leq n-1$ in (3).

Theorems 1 and 2 show that any p.d. matrix B, or equivalently any c.n.d. matrix C, define a unique set of squared Euclidean distances D between objects (Torgerson 1958; Rao 1964; Gower 1966). The latter can be shown (e.g. from (4)) to obey the celebrated *Huygens principle*, namely (e.g. Benzécri 1973)

$$\sum_{j=1}^{n} a_j D_{ij} = D_{ia} + \Delta_a \qquad\qquad \Delta_a = \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j D_{ij}\,, \qquad (5)$$

where D_{ia} denotes the squared distance between object i (with coordinates x_i) and the *a-barycenter* defined by the coordinates $\bar{x}_a = \sum_j a_j x_j$. Also, $\Delta_a \geq 0$ interprets as the average dispersion of the cloud, provided a is a non-negative distribution representing the relative weights of the objects. In the general case of a signed distribution, Δ_a is still well defined, but can be negative.

The squared Euclidean distance between the barycenters $\bar{x}_a$ and $\bar{x}_b$ associated to two signed distributions a and b can also be shown to satisfy

$$D_{ab} = -\frac{1}{2}\sum_{ij}(a_i - b_i)(a_j - b_j)D_{ij}\,, \qquad (6)$$

which directly demonstrates the c.n.d. nature of D (since $z_i = a_i - b_i$ obeys $\sum_i z_i = 0$). Also, (6) entails (5) with the choice $b_j = \delta_{jk}$ for some k.

Substituting (5) in (1) yields

$$B_{ij}(a) = -\frac{1}{2}(D_{ij} - D_{ia} - D_{ja})\,,$$

which, by the cosine theorem, is the matrix of the *scalar products* between x_i and x_j as measured from the origin $\bar{x}_a$. Low-dimensional factorial reconstructions (that is limiting the sum in (3) to the largest eigenvalues) express a maximum amount of $\mathrm{tr}(B(a)) = \sum_i D_{ia}$. This quantity, without direct interpretation, is proportional to the *uniform* dispersion of the coordinates cloud with respect to the point $\bar{x}_a$. Also, $\mathrm{tr}(B(a))$ is minimum when a is the uniform distribution, a standard choice in classical MDS (e.g. Mardia, Kent, and Bibby 1979).

Concentrating the mass of a on a single existing object, typically the last one, is often proposed for computational convenience. Other prescriptions consider a_i as proportional to the precision of measurement of object i (e.g. Borg and Groenen 1997), or set $a_i = 0$ for objects whose behavior

might influence excessively the overall configuration, as in the treatment of "*supplementary elements*" in Correspondence Analysis (e.g. Benzécri 1973; Lebart, Morineau and Piron 1998; Meulman, van der Kooij and Heiser 2004; Greenacre and Blasius 2006). Other choices such as the circumcenter or the incenter are discussed in Gower (1982). Note that the signed nature of a allows one to define an *external origin* $\bar{x}_a$ lying outside the convex hull of the n points, resulting in $B_{ij}(a) \geq 0$ for all pairs.

As a matter of fact, the choice of the origin a and the choice of the object weights f constitute two *distinct* operations, as made explicit by the following generalization of classical MDS (Cuadras and Fortina 1996; Bavaud 2006, 2009):

Theorem 3 (Weighted MDS). *Consider n weighted objects with positive weights $f_i > 0$ normalized to $\sum_i f_i = 1$, together with a (symmetric, non-negative, null-diagonal) pairwise dissimilarity matrix $D = (D_{ij})$. Let $\Pi = (\pi_{ij}) = diag(f)$, i.e. $\pi_{ij} = f_i \delta_{ij}$. Then D is squared Euclidean iff the matrix of* weighted scalar products

$$K(a) = -\frac{1}{2}\sqrt{\Pi} H(a) D H'(a) \sqrt{\Pi} \qquad \textit{that is} \qquad K_{ij}(a) = \sqrt{f_i\, f_j}\, B_{ij}(a)$$

is p.d. The objects coordinates can be chosen as

$$x_{i\alpha} = \sqrt{\tfrac{\lambda_\alpha(a)}{f_i}}\, u_{i\alpha}(a) \qquad \textit{with} \qquad D_{ij} = \sum_{\alpha=1}^{p} (x_{i\alpha} - x_{j\alpha})^2 \,, \qquad (7)$$

where the eigenvalues $\lambda_\alpha(a)$ and eigenvectors $u_{i\alpha}(a)$ are obtained from the spectral decomposition of $K(a) = U(a)\Lambda(a)U'(a)$. Moreover, the corresponding low-dimensional factorial reconstruction, retaining in (7) only the components α associated with the largest eigenvalues, express a maximum proportion of the total inertia relatively to a, namely

$$tr(K(a)) = \sum_{\alpha=1}^{p} \lambda_\alpha = \sum_i f_i D_{ia} = \Delta_f + D_{fa} \ . \qquad (8)$$

The proof follows from the definitions and Theorem 2 by direct substitution. The last identity is a consequence of (5), and shows in particular the total inertia to be minimum for $a = f$, as expected. When f is uniform, the eigenvalues in Theorems 2 and 3 coincide up to a factor n.

2.2 The Class of Schoenberg Transformations

If $A = (a_{ij})$ and $B = (b_{ij})$ are p.d. matrices of the same order n, so are cA for $c \geq 0$, $(t_i a_{ij} t_j)$ for any vector t (cf. Theorem 3), $A + B$,

AB as well as the element-wise product or *Hadamard product* $A \circ B$ with components $a_{ij}b_{ij}$. The latter result (Schur theorem), can be first proved for rank-one p.d. matrices, and then extended to arbitrary ranks by matrix addition (e.g. Horn and Johnson 1991; Bhatia 2006). Combining those facts, one obtains that the Hadamard integral power $A^{\circ p}$ with components a_{ij}^p (where $p \in \mathbb{N}$) or the Hadamard exponential $\exp(\circ A)$ with components $\exp(a_{ij})$ are p.d. However, $A^{\circ\lambda}$ is generally not p.d. for $\lambda > 0$, unless $\lambda \geq n-2$ (Fitzgerald and Horn 1977). P.d. matrices A such that $A^{\circ\lambda}$ is p.d. for every $\lambda \geq 0$ are called *infinitely divisible*.

P.d. matrices are referred to as *kernels* in the Machine Learning community (e.g. Haussler 1999; Cristianini and Shawe-Taylor 2003; Hofmann, Schölkopf and Smola 2008; and references therein). One of the most popular kernels is the so-called *radial basis function* or *Gaussian kernel* $\exp(-\lambda D_{ij})$.

Theorem 4 (Infinitely Divisible Kernels). *Let $C = (c_{ij})$ be a symmetric matrix, and define $B = \exp(\circ - C)$, that is $b_{ij} = \exp(-c_{ij})$. Then*

$$B \text{ is infinitely divisible} \quad \Leftrightarrow \quad C \text{ is c.n.d.}$$

Proof: (Horn and Johnson 1991, p. 456): Consider the matrix $a_{ij}(\lambda) = (1 - b_{ij}^{\lambda})/\lambda$. If B is infinitely divisible, then $z'A(\alpha)z \leq 0$ for any vector z summing to zero, that is $A(\lambda)$ is c.n.d. for any $\lambda > 0$. Hence $\lim_{\lambda\to 0^+} a_{ij}(\lambda) = -\ln b_{ij}$ is c.n.d., showing "$\Rightarrow$". Conversely, suppose C is c.n.d., and define $F = -H(a)CH'(a)$ where $H(a)$ is the centering matrix of Section 2.1. By Theorem 1, F is p.d., and so is $\exp(\circ F)$. But $\exp(f_{ij}) = \exp(-c_{ij} - \eta_i - \eta_j)$ since $f_{ij} = -c_{ij} - \eta_i - \eta_j$ for some η. Hence $b_{ij} = \exp(-c_{ij}) = \exp(\eta_i)\exp(f_{ij})\exp(\eta_j)$ is of the form $t_i a_{ij} t_j$ with A p.d, and hence p.d. By the same reasoning, $b_{ij}^{\lambda} = \exp(-\lambda c_{ij})$ is p.d. for any $\lambda \geq 0$, since λC is c.n.d. iff C is c.n.d., thus proving "$\Leftarrow$".
■

Corollary 1 (Gaussian Kernel). *Let D_{ij} be a squared Euclidean distance. Then, for any $\lambda \geq 0$, $\exp(-\lambda D_{ij})$ is p.d., and $\tilde{D}_{ij}(\lambda) = 1 - \exp(-\lambda D_{ij})$ is a squared Euclidean distance.*

Proof: The first assertion follows form Theorem 4, and the second from Theorem 2 together with the fact that $\tilde{D}_{ij}(\lambda)$ can easily be shown to be c.n.d. with a null diagonal.
■

More generally, any mixture of $\tilde{D}(\lambda)$ over $\lambda \geq 0$ is a squared Euclidean distance, yielding the following definition and theorem:

Definition 2 (Schoenberg Transformations). A Schoenberg transformation is a function $\varphi(D)$ from $\mathbb{R}^+$ to $\mathbb{R}^+$ of the form (Schoenberg 1938a)

$$\varphi(D) = \int_0^\infty \frac{1 - \exp(-\lambda D)}{\lambda}\, g(\lambda)\, d\lambda\,, \tag{9}$$

where $g(\lambda)\, d\lambda$ is a non-negative measure on $[0, \infty)$ such that $\int_1^\infty \frac{g(\lambda)}{\lambda} d\lambda < \infty$.

Note that (9) entails $\varphi(D) \geq 0$ and $\varphi(0) = 0$ together with

$$\varphi'(D) = \int_0^\infty \exp(-\lambda D)\, g(\lambda)\, d\lambda\,, \tag{10}$$

where $\varphi'(D)$ denotes the *derivative* of $\varphi(D)$.

Theorem 5 (Fundamental Property of Schoenberg Transformations). *Let D be a $n \times n$ matrix of squared Euclidean distances. Define the components of the $n \times n$ matrix $\tilde{D}$ as $\tilde{D}_{ij} = \varphi(D_{ij})$, where $\varphi(D)$ is a Schoenberg transformation. Then $\tilde{D}$ is a squared Euclidean distance.*

It follows from the above that all componentwise transformations of the form $\tilde{D}_{ij} = \varphi(D_{ij})$ transform a squared Euclidean distance into another squared Euclidean distance. In his paper (1938a), Schoenberg indeed proved (Theorem 6, p. 828) that *all* such transformations are given by Definition 2. More precisely, Schoenberg addressed and solved the question of determining the class Φ_m of all the transformations $\tilde{D} = \varphi(D)$ of squared Euclidean distances D, associated with any configuration in $\mathbb{R}^p$, which are isometrically embeddable in an Euclidean space of sufficiently large dimensionality, that is in an Hilbert space $\mathbb{R}^\infty$. By construction, $\Phi_1 \supset \Phi_2 \supset \ldots \supset \Phi_\infty$, and Definition 2 characterizes the class $\Phi_\infty = \cap_{p \geq 1} \Phi_p$. The class Φ_1 is central to Brownian and fractional Brownian motion (e.g. Alpay, Attia, and Levanony 2009), while lower-order classes $\Phi_{p \leq 3}$ are fundamental in Geostatistics (e.g. Christakos 1984) and spatial interpolation (e.g. Micchelli 1986; Stein 1999).

3. Some Properties of the Schoenberg Transformations

3.1 Complete Monotonicity

By construction, $\varphi'(D)$ in (10) coincides with the class of *completely monotonic functions* $f(D)$ obeying $(-1)^n f^{(n)}(D) \geq 0$ (Bernstein 1929). Hence Schoenberg transformations are characterized by $\varphi(D) \geq 0$ with $\varphi(0) = 0$, with positive odd derivatives $\varphi'(D)$, $\varphi'''(D)$, etc., and negative even derivatives $\varphi''(D)$, $\varphi''''(D)$, etc. (see Table 1).

In particular, D^α with $0 < \alpha < 1$ is Euclidean when D is Euclidean (Schoenberg 1937) – or even, for α small enough, when D is a plain dissimilarity (Joly and Le Calvé 1986; see also Critchley and Fichet 1994 for a review on typologies of Euclidean and non-Euclidean dissimilarities).

Table 1. Some Schoenberg transformations.

function $g(\lambda)$		transformation $\varphi(D)$	$*$	$**$
$g_1(\lambda) = \delta(\lambda - a)$	$a \geq 0$	$\varphi_1(D) = \frac{1-\exp(-aD)}{a}$	✓	✓
$g_2(\lambda) = \theta(\lambda \leq \frac{\pi}{2})\ \lambda\ \sin\lambda$		$\varphi_2(D) = \frac{D(D+\exp(-\frac{\pi}{2}D))}{1+D^2}$	✓	✓
$g_3(\lambda) = \exp(-a\lambda)$	$a > 0$	$\varphi_3(D) = \ln(1 + \frac{D}{a})$	$-$	✓
$g_4(\lambda) = \lambda \exp(-a\lambda)$	$a > 0$	$\varphi_4(D) = \frac{D}{a(a+D)}$	✓	✓
$g_5(\lambda) = \frac{a}{\Gamma(1-a)}\lambda^{-a}$	$0 < a < 1$	$\varphi_5(D) = D^a$	$-$	$-$
see Berg et al. (2008)		$\varphi_6(D) = \frac{D^a}{1+D^a}$ $\quad 0 < a < 1$	✓	$-$

*Bounded ** Rectifiable

Also, the identity transformation $\varphi(D) = D$ obtains from $g(\lambda) = \delta(\lambda)$. The latter contribution can be made explicit in the following variant, equivalent to Definition 2:

$$\varphi(D) = b\, D + \int_0^\infty (1 - \exp(-\lambda D))\, d\mu(\lambda)\,,$$

where μ is a non-negative measure on $(0, \infty)$ such that $\int_0^\infty \frac{\lambda}{1+\lambda}\, d\mu(\lambda) < \infty$ and $b \geq 0$.

There exists an important literature about *Bernstein functions* (see e.g. Berg, Mateu, and Porcu 2008; Schilling, Song, and Vondraček 2010; and references therein), defined as the smooth non-negative functions whose first derivatives are completely monotonic. Hence, Schoenberg transformations coincide with the class of Bernstein functions which are zero at the origin, in the same way that Euclidean distances are c.n.d matrices with a null diagonal (Theorem 2).

By construction, Schoenberg transformations are closed under composition, as exemplified by $\varphi_6 = \varphi_4 \circ \varphi_5$ in Table 1.

3.2 Arc Length; Rectifiable and Bounded Transformations

A Schoenberg transformation acts as an anamorphosis between Euclidean spaces: to any initial configuration of points X, with mutual squared Euclidean distances $D(X)$, corresponds a distorted configuration $\tilde{X}$ reconstructible by MDS from $\tilde{D} = \phi(D)$. By construction, the mapping $\tilde{X}(X)$ is unique up to a translation and a rotation.

Consider a smooth curve C whose arc length is parameterized by s, containing two close points at mutual distance Δs. The corresponding distance on the transformed curve $\tilde{C}$ is $\Delta\tilde{s} = \sqrt{\varphi((\Delta s)^2)}$. By l'Hospital's rule, the ratio of the infinitesimal arc lengths is

$$\frac{d\tilde{s}}{ds} = \lim_{\Delta s \to 0} \frac{\sqrt{\varphi((\Delta s)^2)}}{\Delta s} = \sqrt{\varphi'(0)}\,,$$

which might be finite or not. On the other hand, infinitely distant points in the original space might be infinitely distant or not in the transformed space:

Definition 3. The transformation $\varphi(D)$ is said to be

- rectifiable if $\varphi'(0) < \infty$, that is iff $\int_0^\infty g(\lambda)\, d\lambda < \infty$
- bounded if $\varphi(\infty) < \infty$, that is iff $\int_0^\infty \frac{g(\lambda)}{\lambda}\, d\lambda < \infty$.

3.3 Right Angles

Consider a triangle ikj with a right angle in k. Hence $D_{ij} = D_{ik} + D_{jk}$ by Pythagoras' theorem. Yet, in the transformed space, $\tilde{D}_{ij} \leq \tilde{D}_{ik} + \tilde{D}_{jk}$ since $\varphi(D_1 + D_2) \leq \varphi(D_1) + \varphi(D_2)$, which can be demonstrated by integrating $(1 - \exp(-\lambda D_1))(1 - \exp(-\lambda D_2)) \geq 0$ as in (9). That is, *the Schoenberg transformation $\tilde{\alpha}$ of a right angle $\alpha = \pi/2$ is in general acute*. By the cosine theorem,

$$\cos\tilde{\alpha} = \frac{\varphi(D_1) + \varphi(D_2) - \varphi(D_1 + D_2)}{2\sqrt{\varphi(D_1)\varphi(D_2)}} \geq 0 \ . \tag{11}$$

Under uniform linear dilatation of the original right-angled triangle by a factor $\epsilon > 0$, (11) readily yields that $\lim_{\epsilon\to\infty} \tilde{\alpha}(\epsilon) = \pi/3$ whenever φ is bounded, and $\lim_{\epsilon\to 0} \tilde{\alpha}(\epsilon) = \pi/2$ whenever φ is rectifiable.

3.4 Curvature

Straight lines are bent by Schoenberg transformations: think of a rod whose linear distances d between constituents are contracted as, say, $\sqrt{d}$. The curvature in the transformed space can be measured by first considering in the original space three aligned points i, k, j with $d_{ik} = d_{kj} = \epsilon$ and $d_{ij} = 2\epsilon$. The *Menger's curvature* κ is defined as the limit (Blumenthal 1953, p. 75)

$$\kappa = \lim_{\epsilon\to 0} \frac{4\tilde{A}_{ijk}(\epsilon)}{\tilde{d}_{ij}(\epsilon)\, \tilde{d}_{jk}(\epsilon)\, \tilde{d}_{ik}(\epsilon)} \,,$$

where $\tilde{A}_{ijk}$ is the area of the triangle ijk in the transformed space and $\tilde{d}$ denotes the length of the corresponding sides. Heron's formula

$$16\tilde{A}^2_{ijk} = (\tilde{d}_{ij}+\tilde{d}_{jk}+\tilde{d}_{ki})(-\tilde{d}_{ij}+\tilde{d}_{jk}+\tilde{d}_{ki})(\tilde{d}_{ij}-\tilde{d}_{jk}+\tilde{d}_{ki})(\tilde{d}_{ij}+\tilde{d}_{jk}-\tilde{d}_{ki})$$

yields after simplification

$$\kappa^2 = \lim_{\epsilon\to 0} \frac{4\varphi(\epsilon^2) - \varphi(4\epsilon^2)}{\varphi^2(\epsilon^2)} = -\frac{6\,\varphi''(0)}{(\varphi'(0))^2} \geq 0 \,,$$

where l'Hospital's rule has been used twice in the last equality, under the assumption of rectifiability.

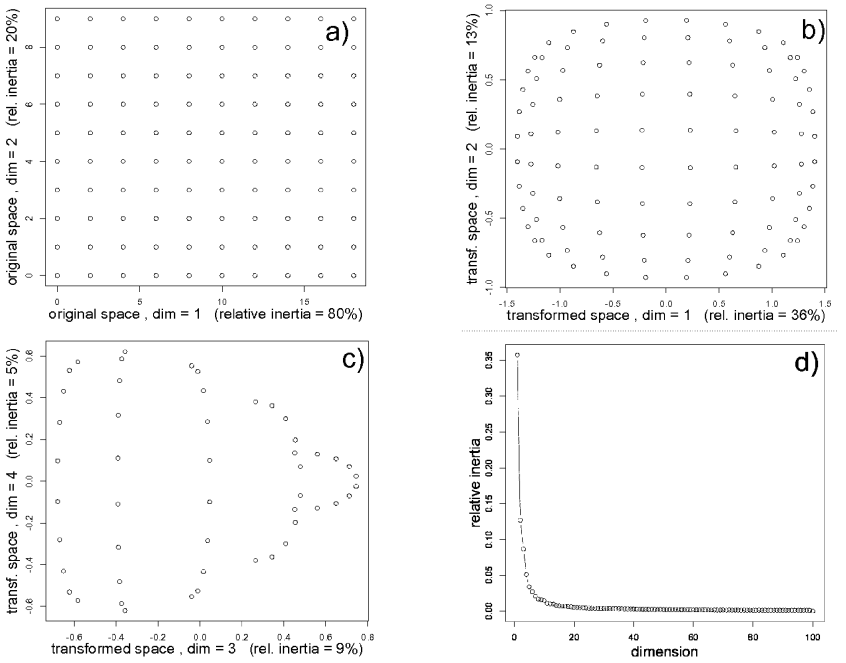

Figure 1. a) Initial configuration, on which the transformation $\varphi(D) = D^{0.4}$ is applied. b) and c) depict the low-dimensional reconstruction of the transformed configuration, obtained by weighted MDS (Theorem 4 where $a = f$ is the uniform distribution. d) Scree graph, proportional to the eigenvalues (8).

4. Illustrations

4.1 Grid

Consider $n = 100$ points forming the bidimensional grid of Figure 1a), on which the transformation $\varphi(D) = D^{0.4}$ is applied. Figures 1b) and 1c) depict the four first dimensions of the transformed configuration, expressing altogether 62% of the total inertia.

4.2 Rod

Figure 2 depicts the low-order projections (b, c, d, e and f) of the non-rectifiable square root transformation $\tilde{D} = \sqrt{D}$ of a quasi-unidimensional rod of $n = 1'000$ points, uniformly generated as $X_1 \sim U(0, 1000)$ and $X_2 \sim U(0, 1)$ (a). As expected, the transformed rod is bent, although the curvature formula of Section 3.4 does not applies here ($\varphi'(0) = \infty$).

The transformation of a line is called "screw line" by Von Neumann and Schoenberg (1941), and "helix" by Kolmogorov (1940) – an adequate terminology in view of Figure 2.

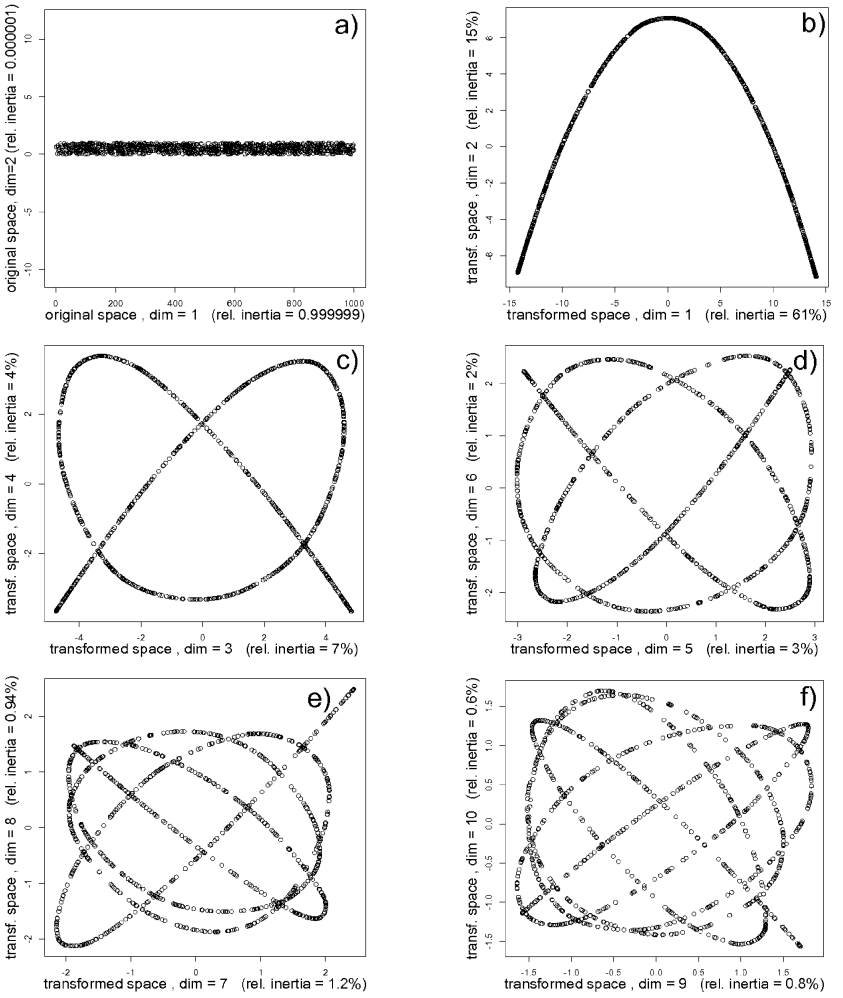

Figure 2. Low-order projections (b, c, d, e and f) of the square root transformation $\tilde{D} = \sqrt{D}$ of a finite rod (a).

The first MDS dimensions turn out to express 61.0%, respectively 15.1% of the relative inertia. Analytic arguments, to be developed in a forthcoming publication, demonstrate the corresponding exact quantities to be $\frac{6}{\pi^2} = 60.8\%$, respectively $\frac{15}{2\pi^2} = 15.2\%$ for a line.

5. Applications

Arguably, all traditional methods in Data Analysis involve, explicitly or not, squared Euclidean distances between observations. Transforming the

latter hence extends the scope of classical methods quite straightforwardly – as briefly illustrated below.

5.1 Distance-based Discriminant Analysis

Consider a collection of objects $i = 1, \ldots, n$ endowed with p-dimensional features, yielding squared Euclidean distances D_{ij} between objects, possibly after standardization and/or orthogonalization of the features (Mahalanobis distances). Also, suppose that each object belongs to a group $g = 1, \ldots m$. An elementary discriminant strategy would consist in assigning each object i to the group g whose centroid is the closest to i, that is to assign i to $\arg\min_g D_{ig}$: this is the linear discriminant prescription of Fisher (1936), successfully applied on the Iris Data ($n = 150$, $p = 4$, $m = 3$) with a percentage of well-classified individuals as high as 97%.

The same strategy is bound to fail with the data of Figure 3 ($n = 150$, $p = 2$, $m = 3$), reaching a percentage of well-classified individuals of 35%, close to the expected value of 33% under random attribution.

However, linear discrimination can be attempted on Schoenberg transformations of the original distances, resulting in the algorithm (see (5)):

Distance-based discriminant algorithm:
1) compute $\tilde{D}_{i\tilde{g}} = \sum_{j=1}^{n} f_j^g \tilde{D}_{ij} - \frac{1}{2}\sum_{j,k=1}^{n} f_j^g f_k^g \tilde{D}_{jk}$, where $\tilde{D}_{ij} = \varphi(D_{ij})$ and where $f_j^g = I(i \in g)/n_g$ (with $n_g = \sum_{j\in g} 1$) is the distribution of objects i in group g
2) assign object i to group $\arg\min_{\tilde{g}} \tilde{D}_{i\tilde{g}}$.

Figure 4 shows the resulting proportion of well-classified individuals, for various one-parameter families of transformations $\varphi(D|a)$. In this data set, the maximum proportion of well-classified individuals reaches 100% for the Gaussian transformation (for $a \geq 0.65$). That is, a sufficiently non-linear Schoenberg transformation succeeds in mapping the initial configuration of Figure 3 in such a way that the three groups can be enclosed in three associated *disjoint* hyperspheres.

Close results are, ever since the nineties, rightly claimed by Machine Learning, where the non-linear, higher-dimensional embedding enabling the linear separation of groups is emblematic (see e.g. Chen, He, and Wang 2007 and references therein). Also, Cuadras, Fortina, and Oliva (1997) have, in another context, proposed the same algorithm – whose conceptual, formal and computational simplicity should be emphasized.

5.2 Robust Estimates of Location; Robust PCA

In one dimension, determining the point a minimizing $\sum_i f_i(x_i - a)^2$ yields the weighted mean; minimizing $\sum_i f_i|x_i - a|$ yields the weighted median. More generally, finding the centroid a minimizing the quantity

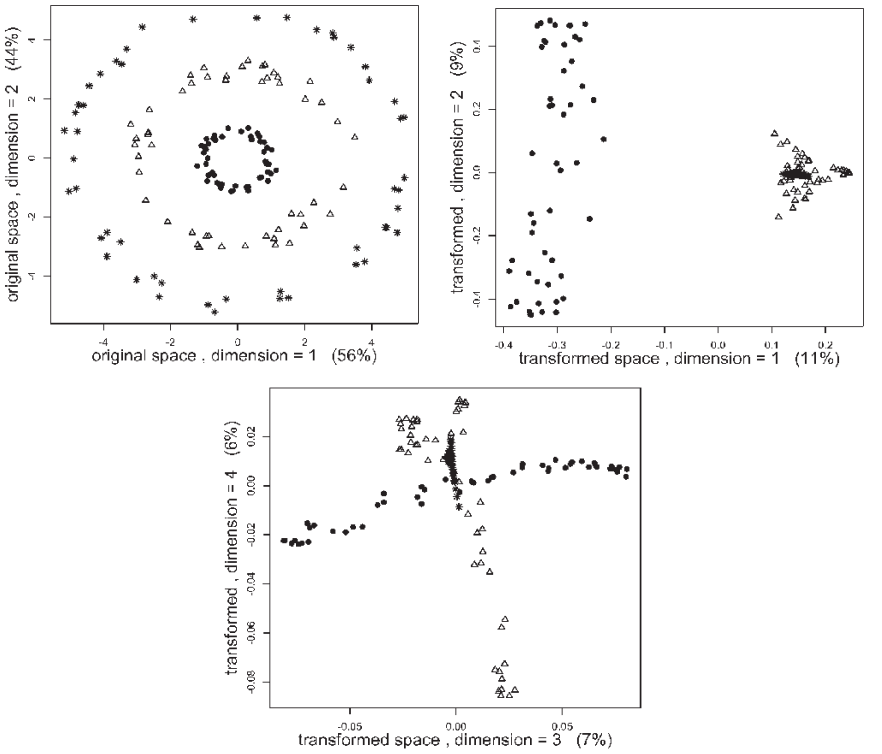

Figure 3. Top left: three groups of 50 individuals each, uniformly generated on concentric circles of radii 1, 3 and 5, with a radial standard deviation of 0.1, 0.3 and 0.2, respectively. MDS reconstruction of the configuration transformed as $\varphi(D) = 1 - \exp(-0.65\ D)$ (see text), in dimensions 1 and 2 (top right) and dimensions 3 and 4 (bottom).

$$\sum_i f_i \varphi(D_{ia}) \quad \text{with} \quad D_{ia} = \|x_i - a\|^2 \ , \ a = \sum_i \alpha_i x_i \quad \text{and} \quad \sum_i \alpha_i = 1 \tag{12}$$

defines a centroid a as the solution of the iterative scheme (see (5) and Bavaud 2011 for details)

$$\alpha_i = \frac{f_i \varphi'(D_{ia})}{\sum_j f_j \varphi'(D_{ja})} \qquad D_{ia} = \sum_j \alpha_j D_{ij} - \frac{1}{2} \sum_{jk} \alpha_j \alpha_k D_{jk} \ . \tag{13}$$

The centroid a is a robust estimate of location, analogous to a M-estimate (e.g. Hampel, Ronchetti, Rousseeuw, and Stahel 1986 and references therein), valid for any dimension: the term $\varphi'(D_{ia})$ downweights

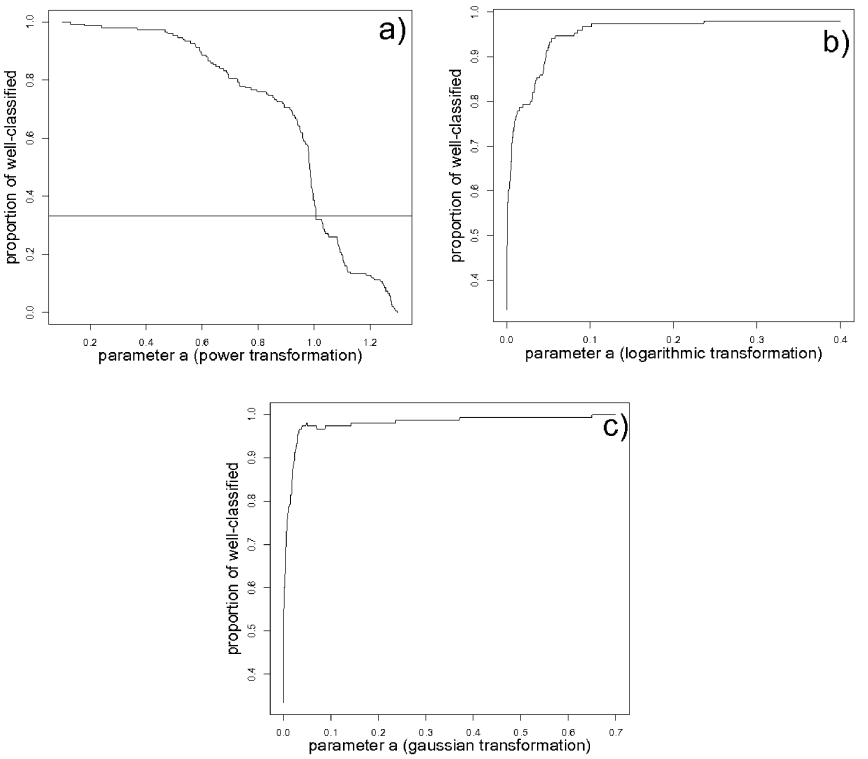

Figure 4. Proportion of well-classified individuals, after Schoenberg transformation of the original data of Figure 3. a) power transformation $\varphi(D) = D^a$; note that $a > 1$ does *not* correspond to a valid transformation, and results in a *decrease* of the proportion below the chance level. b) logarithmic transformation $\varphi(D) = \ln(1+aD)$. c) Gaussian transformation $\varphi(D) = 1 - \exp(-aD)$.

distant observations, and many solutions coexist in general (local minima), in particular when $\varphi'(D)$ is rapidly decreasing.

Figure 5 left depicts the bidimensional dataset `faithful` (Härdle 1991), together with the trajectory of the centroid a resulting from the transform $\varphi(D) = 1-\exp(-\lambda D))$ where $\lambda \in (0, 2.7)$, with initial values a_0 uniformly distributed in the range of values. $\lambda \to 0$ yields the mean $a = (0, 0)$. Increasing λ pushes the centroid towards the center of the NE cluster, or, for $\lambda \geq 0.7$, towards the center of the SW cluster as well. In the limit $\lambda \to \infty$, each observation yields a local minimum; in that respect, (13) outlines a clustering scheme, where λ controls the number of groups.

Distributions α (13) also permit to define a "robust covariance" between two variables x and y as

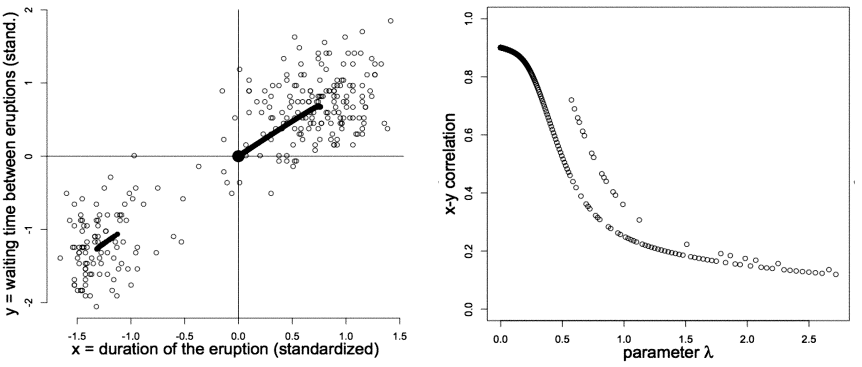

Figure 5. Left: trajectory of the centroid a (solid circles) for various values of $\lambda \in (0, 2.7)$. Right: behavior of the associated "robust correlation", related to the two families of centroids.

$$\mathrm{cov}_\alpha(x, y) = \sum_{i=1}^{n} \alpha_i (x_i - \sum_j \alpha_j x_j)(y_i - \sum_k \alpha_k y_k)$$

whose spectral decomposition defines in turn a "robust PCA scheme", alternative to other proposals found in the literature (e.g. Campbell 1980; Verboven and Hubert 2005; and references therein). Figure 5 right depicts the behavior of the associated "robust correlation", with minima lying in the NE cluster (lower branch) or the SW cluster (upper branch).

6. Conclusion

The Machine Learning literature abounds in algorithms based upon Gaussian and other radial kernels: the procedures exposed in Section 5 exemplify and specify some among many possible applications, aimed at illustrating the operational content of the theory. Higher-order "principled" embeddings, pioneered by the work of Vapnik (1995) and embodied in this article by the class of Schoenberg transformations, are arguably about to be incorporated in standard Data Analysis, to be routinely used in applications, and taught at graduate and undergraduate non-specialized audiences.

The two approaches (kernels versus distances) appear to be equivalent, as illustrated by the results of Section 2. In particular, to the "kernel trick" stating that all the quantities of interest depend upon kernels only (and not upon the object features themselves) corresponds an equally efficient "distance trick", stating that Euclidean distances themselves (and not their underlying coordinates) express all the real quantities of interest, as in (5), (6), or Section 5; see also Schölkopf (2000) and Williams (2002).

Reexpressing the Machine Learning formalism in terms of Euclidean distances, rather than kernels, is hence not only possible, but arguably more intuitive; some related progress on the question "which transformation should be used in which context", so far open, is also expected.

References

ALPAY, D., ATTIA, H., and LEVANONY, D. (2010), "On the Characteristics of a Class of Gaussian Processes Within the White Noise Space Setting", *Stochastic Processes and their Applications, 120*, 1074–1104.

BAVAUD, F. (2006), "Spectral Clustering and Multidimensional Scaling: A Unified View", in *Data Science and Classification*, eds. V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, Heidelberg: Springer, pp. 131–139.

BAVAUD, F. (2009), "Aggregation Invariance in General Clustering Approaches", *Advances in Data Analysis and Classification, 3*, 205–225.

BAVAUD, F. (2011), "Robust Estimation of Location through Schoenberg Rransformations", *submitted*.

BENZÉCRI, J.-P., and collaborators. (1973), "L'analyse des données. 1 : La taxinomie. 2 : L'analyse des correspondances", Paris: Dunod.

BERNSTEIN, S. (1929), "Sur les fonctions absolument monotones", *Acta Mathematica, 52*, 1–66.

BERG, C., MATEU, J., and PORCU, E. (2008), "The Dagum Family of Isotropic Correlation Functions", *Bernoulli, 14*, 1134–1149.

BHATIA, R. (2006), "Infinitely Divisible Matrices", *The American Mathematical Monthly, 113*, 221–235.

BLUMENTHAL, L.M. (1953), "Theory and Applications of Distance Geometry", Oxford: Oxford University Press.

BORG, I., and GROENEN, P.J.F. (1997), *Modern Multidimensional Scaling: Theory and Applications*, Heidelberg: Springer.

CAMPBELL, N.A. (1980), "Robust Procedures in Multivariate Analysis I : Robust Covariance Estimation", *Applied Statistics, 29*, 231–237.

CHEN, D., HE, Q., and WANG, X. (2007), "On Linear Separability of Data Sets in Feature Space", *Neurocomputing, 70*, 2441–2448.

CHRISTAKOS, G. (1984), "On the Problem of Permissible Covariance and Variogram Models", *Water Resources Research, 20*, 251–265.

CRITCHLEY, F., and FICHET, B. (1994), "The Partial Order by Inclusion of the Principal Classes of Dissimilarity on a Finite Set, and Some of Their Basic Properties", in: *Lecture Notes in Statistics. Classification and Dissimilarity Analysis*, ed. B. van Cutsem, Heidelberg: Springer, pp. 5–65.

CRISTIANINI, N., and SHAWE-TAYLOR, J. (2003), "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge: Cambridge University Press.

CUADRAS, C.M., and FORTINA, J. (1996), "Weighted Continuous Metric Scaling", in *Multidimensional Statistical Analysis and Theory of Random Matrices*, eds. A.K. Gupta and V.L. Girko, The Netherlands: VSP, pp. 27–40.

CUADRAS, C.M., FORTINA, J., and OLIVA, F. (1997), "The Proximity of an Individual to a Population with Applications in Discriminant Analysis", *Journal of Classification, 14*, 117–136.

FISHER, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics, 7*, 179–188.

FITZGERALD C. H., and HORN, R.A. (1977), "On Fractional Hadamard Powers of Positive Definite Matrices", *Journal of Mathematical Analysis and Applications, 61*, 633–642.

GOWER, J.C. (1966), "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis", *Biometrika, 53*, 325–338.

GOWER, J.C. (1982), "Euclidean Distance Geometry", *Mathematical Scientist, 7*, 1–14.

GREENACRE, M., and BLASIUS, J. (2006), "Multiple Correspondence Analysis and Related Methods", London: Chapman & Hall.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986), "Robust Statistics – The Approach Based on Influence Functions", New York: Wiley.

HÄRDLE, W. (1991), "Smoothing Techniques with Implementation in S", Heidelberg: Springer.

HAUSSLER, D. (1999), "Convolution Kernels on Discrete Structures", Technical Report, UCSC-CRL-99-10, University of California at Santa Cruz.

HOFMANN, T., SCHÖLKOPF, B., and SMOLA, A.J. (2008), "Kernel Methods in Machine Learning", *Annals of Statistics, 36*, 1171–1220.

HORN, R.A., and JOHNSON, C.R. (1991), "Topics in Matrix Analysis", Cambridge: Cambridge University Press.

JOLY, S., and LE CALVE, G. (1986), "Etude des puissances d'une distance", *Statistique et Analyse des Données, 11*, 29–50.

KOLMOGOROV, A.N. (1940), "The Wiener Helix and Other Interesting Curves in the Hilbert Space", *Doklady Akademii Nauk SSSR, 26*, 115–118.

LEBART, L., MORINEAU, A., and PIRON, M. (1998), *Statistique exploratoire multidimensionnelle*, Paris: Dunod.

MARDIA, K.V., KENT, J.T., and BIBBY, J.M. (1979), *Multivariate Analysis*, New York: Academic Press.

MEULMAN, J.J., VAN DER KOOIJ, A., and HEISER, W.J. (2004), "Principal Components Analysis with Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data", in *The Sage Handbook of Quantitative Methodology for the Social Sciences*, ed. D. Kaplan, pp. 49–70.

MICCHELLI, C. (1986), "Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions", *Constructive Approximation, 2*, 11–22.

VON NEUMANN, J., and SCHOENBERG, I.J. (1941), "Fourier Integrals and Metric Geometry", *Transactions of the American Mathematical Society, 50*, 226–251.

RAO, C.R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research", *Sankhyā A, 26*, 329–358.

SCHILLING, R., SONG, R., and VONDRAČEK, Z. (2010), *Bernstein Functions: Theory and Applications, Studies in Mathematics, 37*, Berlin: de Gruyter.

SCHOENBERG, I.J. (1937), "On Certain Metric Spaces Arising From Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space", *The Annals of Mathematics, 38*, 787–793.

SCHOENBERG, I.J. (1938a), "Metric Spaces and Completely Monotone Functions", *The Annals of Mathematics, 39*, 811–841.

SCHOENBERG, I.J. (1938b), "Metric Spaces and Positive Definite Functions", *Transactions of the American Mathematical Society, 44*, 522–536.

SCHÖLKOPF, B. (2000), "The Kernel Trick for Distances", *Advances in Neural Information Processing Systems, 13*, 301–307.

STEIN, M.L. (1999), "Interpolation of Spatial Data: Some Theory for Kriging", Heidelberg: Springer.

TORGESON, W.S. (1958), '*Theory and Methods of Scaling*, New York: Wiley.

VAPKNIK, V.N. (1995), *The Nature of Statistical Learning Theory*, Heidelberg: Springer.

VERBOVEN, S., and HUBERT, M. (2005), "LIBRA: a MATLAB library for Robust Analysis", *Chemometrics and Intelligent Laboratory Systems, 75*, 127–136.

YOUNG, G., and HOUSEHOLDER, A.S. (1938), "Discussion of a Set of Points in Terms of Their Mutual Distances", *Psychometrika, 3*, 19–22.

WILLIAMS, C.K.I. (2002), "On a Connection Between Kernel PCA and Metric Multidimensional Scaling", *Machine Learning, 46*, 11–19.