Journal of Machine Learning Research 5 (2004) 801-818 Submitted 12/03; Published 7/04

Feature Discovery in Non-Metric Pairwise Data

Julian Laub JULIAN.LAUB@FIRST.FHG.DE
Fraunhofer FIRST.IDA
Kekulestrasse 7
12489 Berlin, Germany

Klaus-Robert Müller KLAUS@FIRST.FHG.DE
Fraunhofer FIRST.IDA
Kekulestrasse 7
12489 Berlin, Germany, and
University of Potsdam, Department of Computer Science
August-Bebel-Strasse 89
14482 Potsdam, Germany

Editor: Isabelle Guyon

Abstract

Pairwise proximity data, given as similarity or dissimilarity matrix, can violate metricity. This occurs either due to noise, fallible estimates, or due to intrinsic non-metric features such as they arise from human judgments. So far the problem of non-metric pairwise data has been tackled by essentially omitting the negative eigenvalues or shifting the spectrum of the associated (pseudo-)covariance matrix for a subsequent embedding. However, little attention has been paid to the negative part of the spectrum itself. In particular no answer was given to whether the directions associated to the negative eigenvalues would at all code variance other than noise related. We show by a simple, *exploratory* analysis that the negative eigenvalues *can* code for relevant structure in the data, thus leading to the discovery of new features, which were lost by conventional data analysis techniques. The information hidden in the negative eigenvalue part of the spectrum is illustrated and discussed for three data sets, namely USPS handwritten digits, text-mining and data from cognitive psychology.

Keywords: Feature discovery, exploratory data analysis, embedding, non-metric, pairwise data, unsupervised learning

1. Introduction

A large class of data analysis algorithms is based on a vectorial representation of the data. However, for major fields such as bioinformatics (e.g. Altschul et al., 1997), image analysis (Hofmann et al., 1998; Jacobs et al., 2000) or cognitive psychology (Gati and Tversky, 1982; Goldstone et al., 1991), the data is not available as points lying in some vector space but solely arises as scores of pairwise comparisons, typically measuring similarity or dissimilarity between the data points. A global overview of pairwise proximity data can be found in Everitt and Rabe-Hesketh (1997). Table 1 gives a simple instance of pairwise data obtained from human similarity judgments of Morse code (Everitt and Rabe-Hesketh, 1997).

	1	2	3	4	5
1	84	63	13	8	10
2	62	89	54	20	5
3	18	64	86	31	23
4	5	26	44	89	42
5	14	10	30	69	90

Table 1: Test subjects are asked to judge the pairwise similarity of auditory Morse code (long and short tones). Here we show the similarity matrix for the first five digits. The entries correspond to the percentage of a large number of observers who responded 'same' to the row signal followed by the column signal. (Excerpt from Table 1.3 given in Everitt and Rabe-Hesketh (1997)). Note that this proximity matrix is *asymmetric*.

These pairwise proximities, or "data points", are in no natural way related to the common viewpoint of objects lying in some "well behaved" space like a vector space which always is—albeit possibly high dimensional—a very restrictive structure.

There is a class of algorithms specifically designed for pairwise data, e.g. KNN, pairwise k-means (Duda et al., 2001). Otherwise the pairwise data is first "embedded" into a vector space in order to make it available for the numerous algorithms based on vectorial input (Cox and Cox, 2001). Pairwise data satisfying restrictive conditions can be embedded distortionless with respect to metricity into a Euclidean space (Roth et al., 2003b).

Often pairwise data is non-metric and the dissimilarity matrix does not satisfy the mathematical requirements of a metric function. Metric violations preclude the use of well established machine learning methods, which typically have been formulated for metric data only, and more precisely, for vectorial data, thus limiting the interest of raw pairwise data. Non-metric pairwise data can not be embedded distortionless into a (Euclidean) vector space. So, in general, embedding into a Euclidean space (and often subsequent dimension reduction) amounts to distorting pairwise data to enforce Euclidean metricity. This procedure is exemplified by Multi Dimensional Scaling (Cox and Cox, 2001). Other popular methods are e.g. *Locally Linear Embedding* (Roweis and Saul, 2000) and *Isomap* (Tenenbaum et al., 2000).

However, little is known about the real information loss incurred by enforcing metricity, when non-metric data is forcefully embedded into a vector space on the assumption that non-metricity be a mere artifact of noise. This assumption certainly holds for many cases, especially when the pairwise comparison is the output of some algorithm tuned to be metric but relying on some random initialization. It does not hold for pairwise data which is inherently non-metric, e.g. for human similarity judgments, summarizing geometrical (metric) and categorial thinking (possibly non-metric).

Technically, non-metricity translates into indefinite similarity matrices, also called "pseudo-covariance" matrices (Torgerson, 1958), a fact, which imposes severe constraints on the data analysis procedures. Typical approaches to tackle these problems involve omitting altogether the negative eigenvalues like in classical scaling (Young and Householder, 1938) or shifting the spectrum (Roth et al., 2003a) for subsequent (kernel-)PCA (Schölkopf et al., 1998).

The central and so far unanswered question is therefore: *Does the negative part of the spectrum of a similarity matrix code anything useful other than noise?*

This work will contribute by showing that the negative eigenvalues *can* code for relevant structure in the data. The *exploratory* technique outlined below can lead to the discovery of *systematically new* features, which were so far lost or have gone unnoticed by existing algorithms. This is discussed for three illustrative examples after a brief theoretic discussion of the issue of negative eigenvalues and a simple explanation of *how* negative spectra can code specific information.

2. Embedding of Non-Metric Data

In this section we will describe the issue of embedding pairwise proximity data into a vector space. Embedding pairwise data corresponds to finding points in a vector space, such that their mutual distance is as close as possible to the initial dissimilarity matrix with respect to some cost function. Embedding yields points in a vector space, thus making the data available to the numerous machine learning techniques which require vectorial input. Embedding also allows visualization after dimension reduction.

Let $D = (D_{ij})$ be an $n \times n$ dissimilarity matrix. We want to find n vectors x_i in a p-dimensional vector space such that the distance between x_i and x_j is close to the dissimilarity D_{ij} with respect to some cost function measuring the distortion incurred by the embedding procedure.

Let X be the matrix whose column are given by the vectors x_i. The matrix defined by $\frac{1}{n}XX^\top$ is called the *covariance* matrix and is positive semi-definite, i.e., all its eigenvalues λ_i are positive or zero ($\lambda_i \geq 0$). The covariance matrix plays an important role in spectral methods like (kernel-)PCA. The directions corresponding to the leading eigenvalues describe the directions which capture large variance in the data. Thus we expect to find interesting features there.

2.1 Mathematical Statement of the Embedding Problem

We will briefly state the mathematical formulation of the embedding problem and give the necessary and sufficient condition for a Euclidean embedding.

A dissimilarity matrix D will be called *metric* if there exists a metric function d such that $D_{ij} = d(\cdot,\cdot)$. In other words, D is positive and symmetric, its elements are 0 if and only if they are on the diagonal,[1] and they satisfy the triangle inequality. $D = (D_{ij})$ will be called squared-Euclidean if the metric function d derives from the Euclidean norm l_2.

Let $C = -\frac{1}{2}QDQ$ where $Q = I - \frac{1}{n}ee'$. Q is the projection matrix onto the orthogonal complement of $e = (1, 1, \ldots 1)^\top$. The operation $D \to QDQ$ corresponds to the *centralizing* operation. The meaning of C will become clear subsequently.

We have the following important theorem (Torgerson, 1958; Young and Householder, 1938):

Theorem 1 *D is squared-Euclidean iff C is positive semi-definite.*

In other words, the pairwise dissimilarities given by D can be embedded into a Euclidean vector space if and only if the associated matrix C is positive semi-definite.

2.2 Embedding when D is Squared-Euclidean

When D is squared-Euclidean, C is semi-definite positive and can readily be interpreted as covariance matrix (via simple algebra). The embedded vectors can be recovered by usual kernel-PCA (Schölkopf et al., 1998; Cox and Cox, 2001):

1. We reasonably suppose that there are no two identical data points with different labels in the data set.

$$
\begin{aligned}
&D \xrightarrow{C=-1/2QDQ} C \text{ with } n \text{ positive eigenvalues}\\
&C \xrightarrow{\text{spectral decomposition}} V\Lambda V^\top\\
&X_K = |\Lambda_K|^{1/2} V_K^\top,
\end{aligned}
$$

where $V = (v_1, \ldots, v_n)$ with eigenvectors v_i's and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. K is the subspace of chosen directions v_i. The columns of X_K contain the vectors x_i in p-dimensional subspace K, where V_K is the column-matrix of the selected eigenvectors and Λ_K the diagonal matrix of the corresponding eigenvectors.

If $K = \{v_1 \ldots v_p\}$ the distances between these vectors differ the least from the distances D with respect to the quadratic approximation error. For $p = n-1$ the mutual distances coincide with D, i.e. $D_{ij} = ||x_i - x_j||^2$. In other words: *there is a direct algebraic transformation between D and the set of x_i's.*

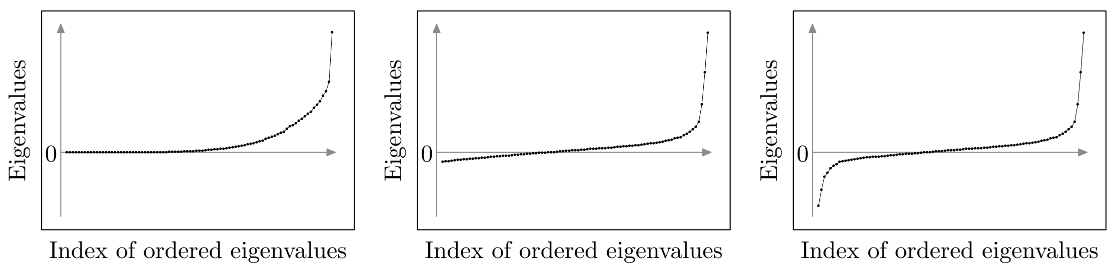

Figure 1: Examples of spectra of C for a squared-Euclidean dissimilarity (left) and non squared-Euclidean dissimilarity matrices. The eigenvalues are plotted against rank order.

2.3 Embedding for General D's

For non squared-Euclidean dissimilarity matrices D the associated C is not positive semi-definite and is not a covariance matrix. In this case we will call it *pseudo-covariance* matrix. Figure 1 shows an instance of a spectrum associated to a positive semi-definite C (left) and two instances of negative spectra: in the middle, a spectrum associated to a pairwise dissimilarity which essentially is metric but corrupted by noise which translates into a spectrum with only a few negative eigenvalues of small magnitude. On the right, a spectrum with negative eigenvalues large in magnitude associated to intrinsic non-metricity.

In order to study the possible loss incurred by omitting the negative part of the spectrum we propose the following simple algorithm, which allows to specifically visualize the information coded by the negative eigenvalues.

Algorithm. Start with some symmetric dissimilarity D or similarity S. If non-symmetric cases the pairwise proximity matrix must first be symmetrized. Furthermore, when the proximity data are similarities, a problem specific dissimilarity is computed. D typically is related to S via, e.g., $D_{ij} = 1 - S_{ij}$ or $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$. Recall that $Q = I - \frac{1}{n}ee'$ with $e = (1, 1, \ldots 1)^\top$. C denotes the pseudo-covariance matrix.

$$
\begin{aligned}
&D \xrightarrow{C=-1/2QDQ} C \text{ with } p \text{ positive and } q \text{ negative eigenvalues}\\
&C \xrightarrow{\text{spectral decomposition}} V\Lambda V^\top = V|\Lambda|^{\frac{1}{2}}M|\Lambda|^{\frac{1}{2}}V^\top\\
&X_K = |\Lambda_K|^{1/2}V_K^\top,
\end{aligned}
$$

where M is the block matrix consisting of the blocks $I_{p\times p}$, $-I_{q\times q}$ and $0_{k\times k}$ (with $k = n-p-q$).

The columns of X_K contain the vectors x_i in the p-dimensional subspace K. At this point K can be very general. However, as for PCA, we will find it sensible to choose a few leading eigendirections *which can also include eigendirections associated to the negative part of the spectrum*.

Visualization. Retaining only the first two coordinates ($K = \{v_1, v_2\}$) of the obtained vectors corresponds to a projection onto the first two leading eigendirections. Retaining the last two ($K = \{v_{n-1}, v_n\}$) is a projection onto the last two eigendirections: *This corresponds to a projection onto directions related to the metric violations of C*

This simple algorithm[2] is very akin to the well known PCA algorithm except that it does not require the spectrum to be positive.

To finish this short overview about embedding pairwise data, it is important to stress the fact that in general metricity is not enough for pairwise data to be loss-free embeddable into a Euclidean vector space. Pairwise data may be metric, yet have an associated spectrum with negative eigenvalues. The interesting case, however, is given for the Euclidean metric since Theorem 1 establishes a very simple relationship between the requirements on the dissimilarity matrix and its loss-free embeddability into a Euclidean vector space.

2.4 Issue of Information Loss

Classical approaches to the embedding into a Euclidean vector space usually involve techniques like multi-dimensional scaling (Cox and Cox, 2001). In its simplest version, classical scaling, MDS proceeds as the algorithm in Section 2.1. However $\Lambda_K^{1/2}$ is only defined for $K \subset \{v_1 \dots v_p\}$ with $p \leq t$ where t is the number of positives eigenvalues. The requirement $p \leq t$ leads to a cut-off of the negative eigenvalues. Another variant of MDS is called *non-metric* MDS and treats ordinal-scale data, where the projections only try to preserve the rank order between the distances, not their absolute value (Kruskal, 1964; Shepard, 1962). It is important to notice here that in our work non-metricity refers to the violations of metric requirements and the subsequent impossibility of a loss-free embedding into a Euclidean vector space. Non-metric MDS does not discover the information coded specifically by metric violations.

Recently *Constant Shift Embedding* was introduced which guarantees distortionless embedding of non-metric pairwise data w.r.t. cluster assignment in the case of a shift invariant cluster cost function (Roth et al., 2003b,a). However, in practical applications, the need for dimension reduction to speed up optimization and robustify solutions, effectively results in retaining only the leading eigendirections and cutting off large parts of the spectrum. For other cases than noise corrupted non-metric pairwise data (Figure 1, middle) it is an open question whether the removal of negative eigenvalues leads to an information loss.

2. A Matlab implementation can be found under `http://ida.first.fraunhofer.de/~jlaub/`.

The above methods, unlike ours, do not permit to specifically study the information coded by non-metricity.

3. Interpretation and Modeling of Negative Spectra

In this section we will discuss the issue of information loss raised by the preceding considerations. We will first show by simple from-scratch constructions how negative spectra can occur. Further understanding will be gained by interpretation of the negative eigenspaces. In particular, a model is presented that can explain negative spectra in the case of human similarity judgments in cognitive psychology. A simple toy illustration will conclude this section.

3.1 Constructing Negative Spectra

Let us first introduce two simple constructions of non-metric pairwise data sets whose non-metric part codes for specific information. These constructions typically come about in situations involving penalization and/or competition of dissimilarity measures by subtraction or division:

$$D_{ij} = (D_1 - D_2)_{ij} \quad \text{or} \quad D_{ij} = \frac{(D_1)_{ij}}{(D_2)_{ij}}, \tag{1}$$

with the assumption that $(D_2)_{ij} \neq 0 \; \forall i, j = 1, 2 \ldots n$ in the latter case. See Figure 2 for a schematic illustration. The structure of the penalized cells is reflected in non-metricity. Such similarity scores occur in various image matching algorithms or in text mining via e.g. the *min-max* dis/similarity (Banerjee and Ghosh, 2002; Dagan et al., 1995), but also in alignment algorithms from e.g. bioinformatics.

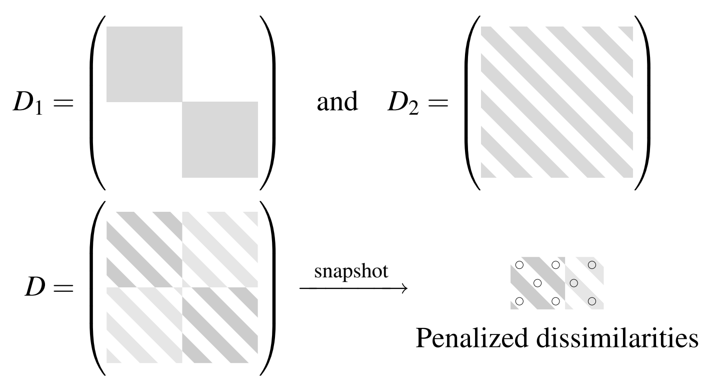

Figure 2: Principle of penalization: the penalized cells can form a structure on their own, which is reflected in non-metricity. The snapshot shows the alternate structure of the penalized entries (small circles).

3.2 Interpreting Negative Eigenspaces

For a positive semi-definite C the projections along the leading eigendirections can readily be interpreted as projections along the axis of high variances of the data. For pseudo-covariance matrices this still holds up to a scaling factor when shifting the spectrum so as to assure positive semi-definiteness.

For projections onto the negative eigendirections the interpretation is not so straightforward since there is no clear intuition on what "negative variance" represents. However the above presented algorithm relies on a pseudo-Euclidean-style decomposition of the embedding space. The pseudo-Euclidean space effectively amounts to two Euclidean spaces one of which has a positive semi-definite inner product and the other a negative semi-definite inner product. An interesting interpretation of the distances in a squared-Euclidean space is that they can be looked at as a difference of squared-Euclidean distances from the "positive" and the "negative" space, by the decomposition $\mathbb{R}^{(p,q)} = \mathbb{R}^p + i\mathbb{R}^q$, so that $D_{ij} = D_{ij}^{(\mathbb{R}^p)} - D_{ij}^{(\mathbb{R}^q)}$, where the D_{ij} are squared-Euclidean. This is the rationale behind the first construction of a non-metric D via $D_{ij} = (D_1 - D_2)_{ij}$.

The power of this decomposition resides in the fact that the negative eigenvalues now admit the natural interpretation of variances of the data projected onto directions in $\mathbb{R}^q$. Thus the variance along v_n is $\sqrt{|\lambda_n|}$, the variance along v_{n-1} is $\sqrt{|\lambda_{n-1}|}$, etc. (c.f. Pękalska et al., 2001).

3.3 Modeling Negative Spectra

While the dissimilarity matrix constructions obtained from Equation 1 can account for a class of non-metric pairwise data from domains such as image matching, text-mining or bioinformatics, where penalization models underlie the computed similarities, they cannot necessarily appropriately model the generic case where a pairwise dissimilarity matrix is given from an experiment, say in cognitive psychology. The simple model for non-metric pairwise data introduced in the following is inspired by approaches in cognitive psychology to explain human similarity judgments (Tenenbaum, 1996). Let $\{f_1, f_2, \ldots f_n\}$ be a basis. A given data point x_i can be decomposed in this basis as $x_i = \sum_{k=1}^n \alpha_k^{(i)} f_k$. The squared l_2 distance between x_i and x_j therefore reads: $d_{ij} = ||x_i - x_j||^2 = \left|\left| \sum_{k=1}^n \left(\alpha_k^{(i)} - \alpha_k^{(j)}\right) f_k \right|\right|^2$. However this assumes constant feature-perception, i.e. a constant mental image with respect to different tasks. In the realm of human perception this is often not the case, as illustrated by the following well known visual "traps" (Figure 3). Our perception has several ways to interpret the figures which can give rise to asymmetry (Thomas and Mareschal, 1997) by a different weighting of the perceived dissimilarities. It is important to notice here that for human similarity judgments, one can hardly speak of artifact or erroneous judgments with respect to a Euclidean norm. The latter seems rather exceptional in these cases.

A possible way to model different interpretations of a given geometric object is to introduce states $\{\omega^{(1)}, \omega^{(2)} \ldots \omega^{(d)}\}$, $\omega^{(l)} \in \mathbb{R}^n$ for $l = 1, 2, \ldots d$, affecting the features. The similarity judgment between objects then depends on the perceptual state (weight) the observer is in. Assuming that the person is in state $\omega^{(l)}$ the distance becomes:

$$d_{ij} = ||x_i - x_j||^2 = \left|\left| \sum_{k=1}^n \left(\alpha_k^{(i)} - \alpha_k^{(j)}\right) \omega_k^{(l)} f_k \right|\right|^2. \tag{2}$$

With no further restriction this model yields non metric distance matrices.

In the worst case l is random, but usually perception-switches can be modeled and l becomes some function of (i, j). For random l, non-metricity is an artifact of sample size, since when averaging the d's over p observers the mean dissimilarity is asymptotically metric in p ($\langle d \rangle \to$ metric as $p \to \infty$): the mean ω becomes constant for all i, j equal to the expectation of its distribution.

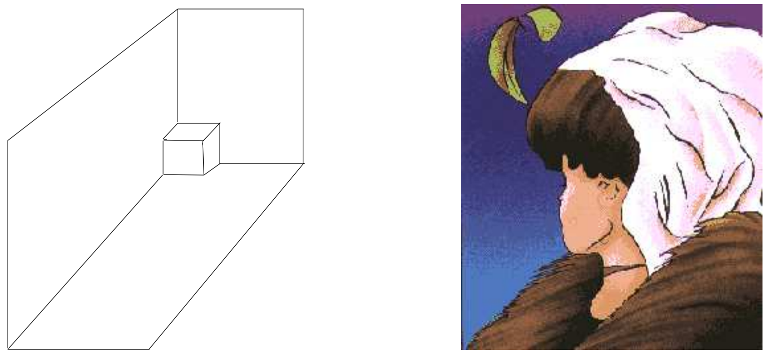

Figure 3: Left: What do you see? A small cube in the corner of a room or a large cube with a cubic hole or a small cube sticking with one corner on a large one? Right: What do you see? A young lady or an old woman? If you were to compare this picture to a large set of images of young ladies or old women, the (unwilling) perception switch could induce large individual weights on the similarity.

On the other hand, if we suppose that the function l of (i,j) does not vary much between observers, then the averaging does not flatten out the non-metric structure induced by the systematic perception-switch.

3.4 Illustration (Proof of Principle)

The importance of the information coded by the negative eigenvalues is exemplified by the following simple example: consider n objects, labeled $1,2,\ldots n$, presenting two salient features. Suppose that they cluster into $\{1,\ldots \frac{n}{2}\}$ and $\{\frac{n}{2}+1,\ldots n\}$ according to the first feature, and into $\{1,3,5,\ldots n-1\}$ and $\{2,4,6,\ldots n\}$ according to the second. Let D_1 and D_2 be the dissimilarity matrices corresponding to feature 1 and 2 respectively. Save very pathological cases the spectra associated to the D obtained by subtraction or division of D_1 and D_2 contain steeply falling negative eigenvalues. Furthermore the projection onto the first two eigenvalues exhibits a clear distinction w.r.t. feature 1 whereas the projection onto the last two eigenvalues exhibits a clear distinction w.r.t. feature 2.

Let e.g. $n=8$. Two artificial dissimilarity matrices D_1 and D_2 were constructed to reflect the above surmise about the underlying structure:

$$D_1 = \begin{pmatrix} 0.00 & 2.36 & 2.59 & 1.78 & 4.74 & 4.82 & 4.98 & 4.72 \\ 2.36 & 0.00 & 2.39 & 1.60 & 4.98 & 5.06 & 5.22 & 4.96 \\ 2.59 & 2.39 & 0.00 & 2.09 & 5.29 & 5.37 & 5.53 & 5.27 \\ 1.78 & 1.60 & 2.09 & 0.00 & 5.08 & 5.16 & 5.32 & 5.06 \\ 4.74 & 4.98 & 5.29 & 5.08 & 0.00 & 1.20 & 1.82 & 1.62 \\ 4.82 & 5.06 & 5.37 & 5.16 & 1.20 & 0.00 & 2.98 & 1.78 \\ 4.98 & 5.22 & 5.53 & 5.32 & 1.82 & 2.98 & 0.00 & 2.02 \\ 4.72 & 4.96 & 5.27 & 5.06 & 1.62 & 1.78 & 2.02 & 0.00 \end{pmatrix} \text{ and } D_2 = \begin{pmatrix} 0.00 & 4.15 & 2.03 & 4.14 & 1.26 & 4.33 & 0.690 & 4.85 \\ 4.15 & 0.00 & 4.70 & 0.570 & 4.37 & 1.82 & 4.24 & 2.02 \\ 2.03 & 4.70 & 0.00 & 4.69 & 1.85 & 4.88 & 1.68 & 5.40 \\ 4.14 & 0.570 & 4.69 & 0.00 & 4.36 & 1.83 & 4.23 & 2.67 \\ 1.26 & 4.37 & 1.85 & 4.36 & 0.00 & 4.55 & 0.730 & 5.07 \\ 4.33 & 1.82 & 4.88 & 1.83 & 4.55 & 0.00 & 4.42 & 2.14 \\ 0.690 & 4.24 & 1.68 & 4.23 & 0.730 & 4.42 & 0.00 & 4.94 \\ 4.85 & 2.02 & 5.40 & 2.67 & 5.07 & 2.14 & 4.94 & 0.00 \end{pmatrix}.$$

Figure 4 shows the result obtained by Algorithm 2.3. The spectrum associated to $D = D_1 - D_2$ is non positive. The information contained in the positive and the negative part is recovered: We see that the information represented in the first two eigendirections is related to the variance due to the cluster structure $\{1,\ldots 4\}$ and $\{5,\ldots 8\}$ whereas the information represented in the last two eigendirection relates to the cluster structure $\{1,3,\ldots 7\}$ and $\{2,4,\ldots 8\}$. *This last information*

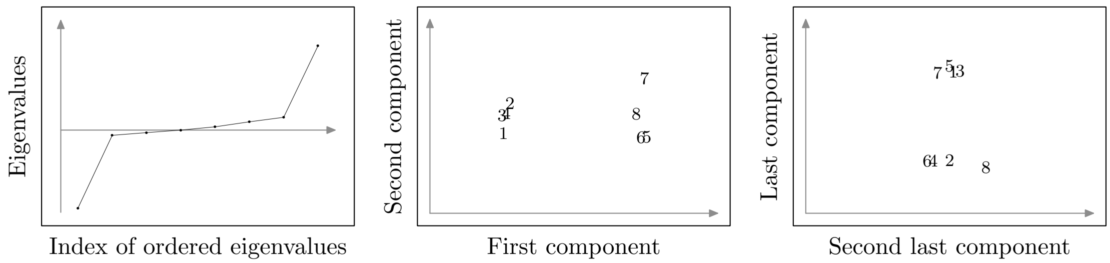

Figure 4: Sorted spectrum associated to the non-metric E (right), Projection onto the two leading positive eigendirection and projection onto the two leading negative eigendirections (right).

would have been lost by usual methods relying on high variance and thus neglecting the negative eigenvalues.

4. Summary

We summarize the procedure and the rationale behind it (see *schematic* diagram Figure 5).

Consider the following illustrative setting: we have apples of different sizes and two colors. There are two salient features: size (geometric) and color (categorial). These apples are pairwise compared, either by a computer algorithm, a human test subject or any other mechanism. This comparison yields a dissimilarity matrix D or a similarity matrix S. In the latter case a problem specific dissimilarity matrix D is obtained from S.

From D we compute the centralized (pseudo-)covariance matrix C and its spectrum. C is positive semi-definite if and only if D is squared-Euclidean. For generic D this is not the case and the usual techniques fail to take this into account.

We project the data onto the first two leading eigenvectors explaining the variance associated to the first feature (size). Second, we project the data onto the last two eigenvectors accounting for the variance of the second feature (color). This last step is done by an embedding into the pseudo-Euclidean space.

The second feature is lost by any method relying exclusively on high variance, that is, the majority of machine learning techniques. We propose the exploration of the negative eigenspectrum for *feature discovery*.

5. Further Illustrative Applications

To go beyond toy examples showing that non-metricity can code for interesting features, we will now illustrate our feature discovery technique by three applications from real-world domains, namely image matching, text mining and cognitive psychology.

5.1 USPS Handwritten Digits

A similarity matrix is computed from binary image matching on the digits 0 and 7 of the USPS data set. Digits 0 and 7 have been chosen since they exhibit clear geometric differences. All images have

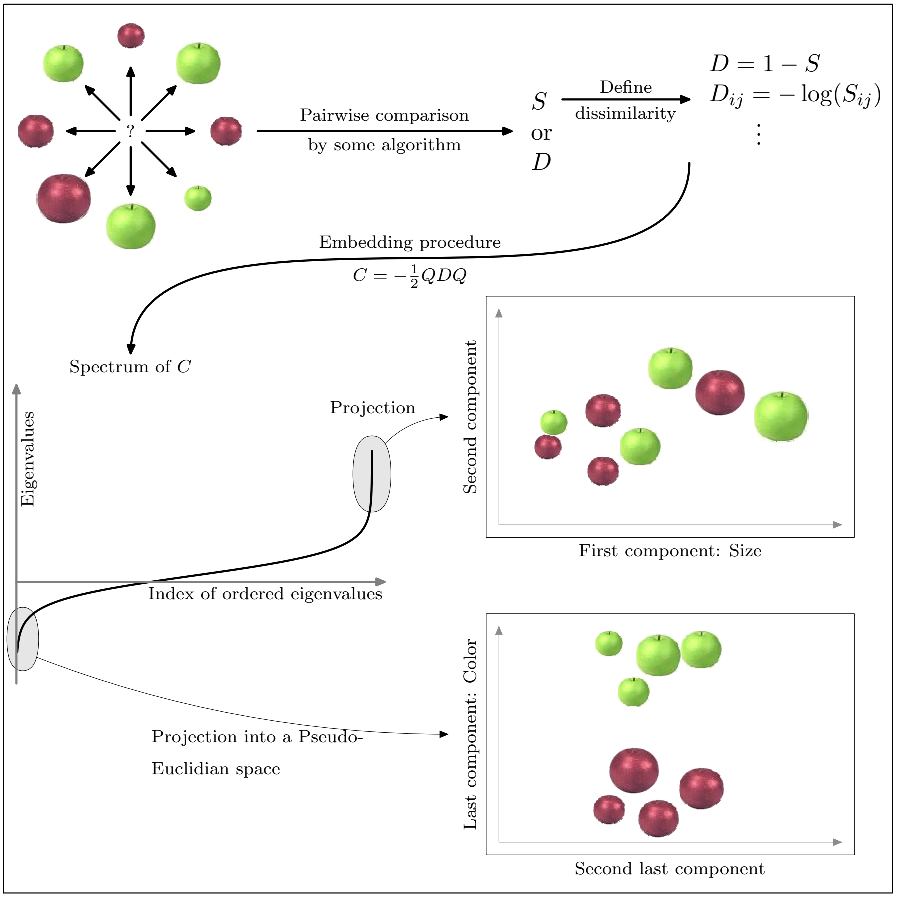

Figure 5: Summarizing diagram.

been sorted according to decreasing sum of pixel value (1 to 256) thus separating the bold digits from the light ones. A total of 1844 samples have been retained. The images have been normalized and discretized to have binary pixel values 0 and 1.

Binary image matching. Let r and s denote the labels of two images and S_{rs} the score rating mutual similarity. In the case of binary images, S_{rs} is a function of a, b, c and d, where a counts the number of variables where both objects s and r score 1, b the number where r scores 1 and s scores 0, c the number of variables where r scores 0 and s scores 1 and d the number of where both objects score 0. The counting variables a, b, c and d allow to define a variety of similarity scores S_{rs} (see Cox and Cox, 2001). We will be interested in the *Simpson* score, defined by

$$S_{rs} = \frac{a}{\min(a+b, a+c)}.$$

It exhibits a strongly falling negative spectrum, corresponding to highly non-metric data. Projection onto the eigenvectors associated to the first leading eigenvalues and projection onto the eigenvectors associated to the last eigenvalues yield results different in nature.

In each case there is a clear interpretation of the variance according to salient features: (i) Figure 6 shows that the variance in the "positive" eigenvectors corresponds to the geometrical dis-

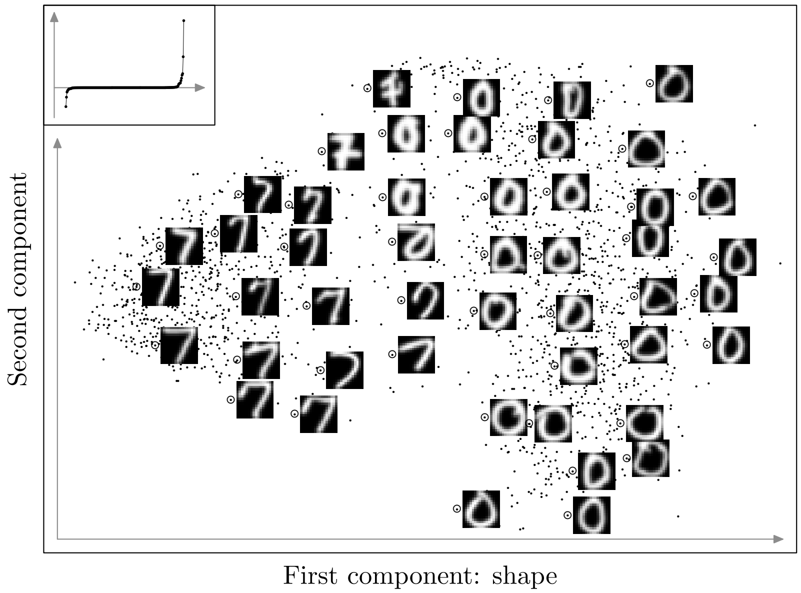

Figure 6: Projection onto the first two positive eigendirections. The explained variance is associated to the geometric shape.

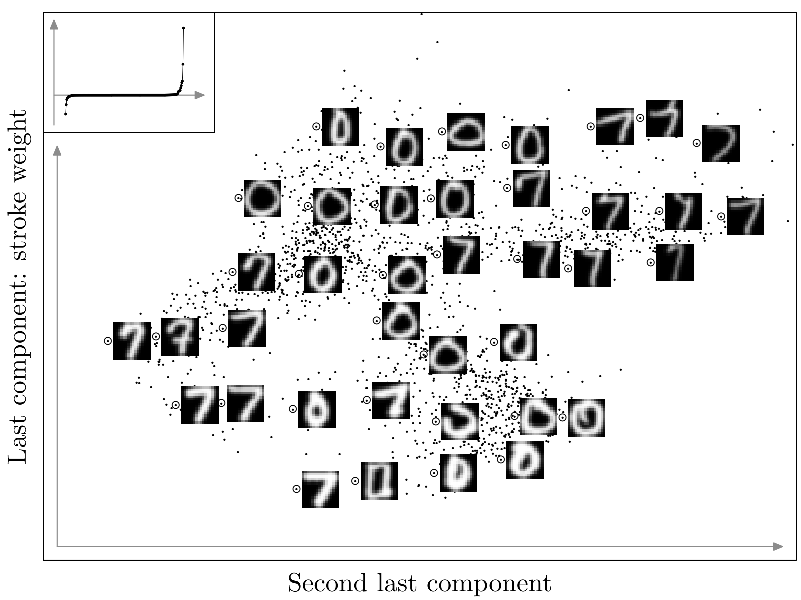

Figure 7: Projection onto the last two negative eigendirections. The explained variance is associated to the stroke weight.

tinction between the shapes of the 0's and the 7's; (ii) Figure 7 on the other hand shows that in the "negative" eigenvectors the variance is associated to the feature of stroke weight. *This inter-*

esting feature would have been lost if we had embedded the data by conventional methods thereby discarding the negative part of the spectrum.

5.2 Text-Mining

We are interested in the semantic structure of nouns and adjectives from two topically unrelated sources, namely Grimm's Fairy Tales (Gutenberg) and popular science articles about space exploration (NASA). Both sources contributed 60 documents containing roughly between 500 and 1500 words each. A subset of 120 nouns and adjectives has been selected, containing both very specific and very general terms out of both data sources.

Similarity measure for words. From a set of p documents and a choice of n keywords we can construct a contingency table, by simply indicating whether word i ($1 \leq i \leq n$) appears in document k ($1 \leq k \leq p$) or not. This yields a $p \times n$ boolean matrix.

We will take the *Keyword Semantic Proximity* as similarity measure (Rocha, 2001; Rocha and Bollen, 2001), which expresses that two words are similar if they often appear together in a document. This similarity is penalized if they individually spread over a large number of documents:

$$s_{ij} = \frac{\#\{\text{documents where word } i \textit{ and } \text{word } j \text{ appear}\}}{\#\{\text{documents where word } i \textit{ or } \text{word } j \text{ appear}\}}.$$

From this similarity measure, we obtain a dissimilarity matrix via, e.g. $d_{ij} = -\log(s_{ij})$. In Rocha (2001) the author uses $d_{ij} = 1/s_{ij} - 1$ which is another possible choice. In either case, the resulting dissimilarity matrix d is not squared-Euclidean such that the associated (pseudo-)covariance matrix exhibits strong negative eigenvalues (see inset in Figure 8).

The data is projected on the first two leading eigenvectors (Figure 8). On the far left we find the words stemming from the popular science articles (e.g. "nuclear", "computer", "physics" etc.) whereas on the far right we have those from Grimm's Fairy Tales (e.g. "castle", "queen", "ravens" etc.). The captured variance can be interpreted as the semantic context of the words.

Projection onto the last two eigendirections yields a distribution over a new interesting feature (Figure 9). We notice that in the upper half we find words of high specificity of either of the sources (e.g. "astronauts", "wolf", "witch" etc.). In the lower half we see an accumulation of words with general, unspecific, meaning, expected to be found in a large variety of documents (e.g. "day", "world", "thing" etc.). Thus to our understanding the variance associated to the last eigendirection again corresponds to the *specificity* of the words (relative to the data source). *This feature would have gone unnoticed by algorithms not specifically taking into account the negative eigenvalues.*

5.3 Cognitive Psychology

We finally present an example from human similarity judgments in cognitive psychology. This will also allow us to illustrate the model presented in Section 3.3.

The pairwise dissimilarity data is obtained from Gati and Tversky (1982). The stimuli tested consist of 16 images of flowers having leaves of varying elongation and stems of increasing size (Figure 10). These two stimuli were presented to a group of thirty undergraduate students from the Hebrew University who, individually, evaluated the mutual dissimilarity of the flowers on a 20-point scale (see Gati and Tversky, 1982).

We have processed the data according to Algorithm 2.3. In the positive eigendirections we obtain a very good reconstruction of the two geometric features, namely the elongation of the leaves

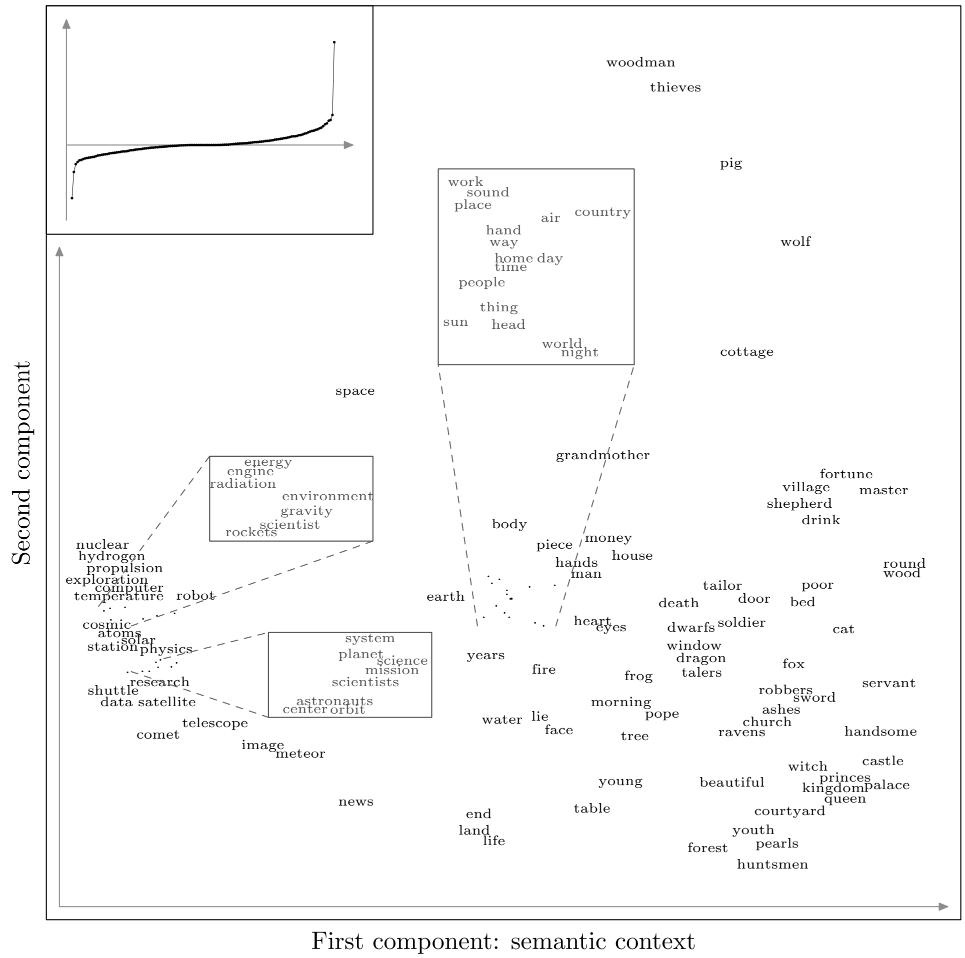

Figure 8: Projection onto the first two eigendirections.

and the size of the stem: see Figure 11, middle. There seems to be no tendency to favor one over the other. The first component explains the variance in leaf elongation (horizontal axis), the second the variance of the stem size (vertical axis).

Interestingly, the projection onto the last two negative eigendirections exhibits further structure, as shown by Figure 11, right. The interpretation, however, is not so straightforward. Two clusters loosely form, separated by the last eigendirection (vertical axis). They are $\{1,2,5,6,11,12,15,16\}$ and $\{3,4,7,8,9,10,13,14\}$. A possible feature could be the oddness of a plant, such that the first cluster contains the odd plants, and the second the "normal" ones, since one could expect plants with small leaves to be of small size and plants with large leaves to be of greater size. The odds here are the small plants with large leaves and the large plants with small leaves. This would correspond to categorial perception while judging similarity.

Features related to the concept of normality, or expectation, are not uncommon in cognitive psychology. In Navarro and Lee (2002) features like the normality or usuality of faces are discussed in the context of the Modified Contrast Model, along with certainly not easily graspable features like relationships in parenthood. While the authors focus on common and distinctive features and

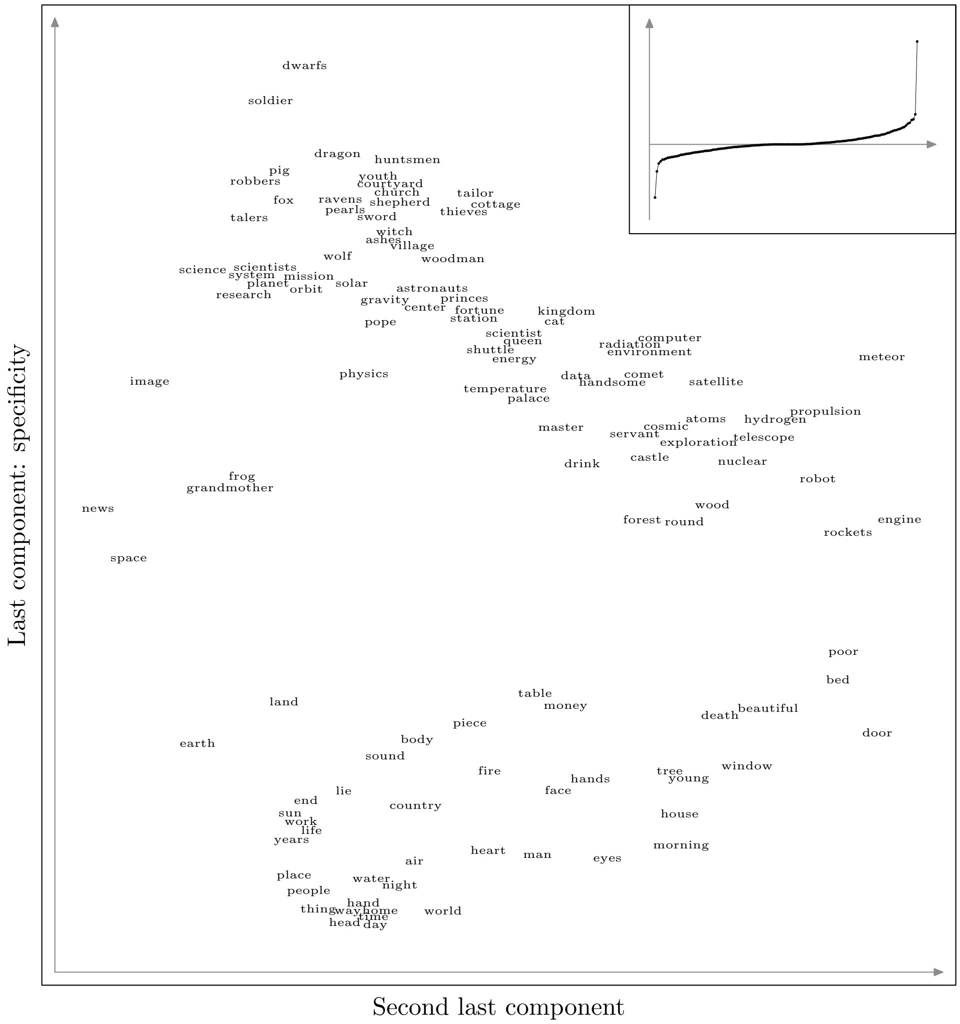

Figure 9: Projection onto the last two negative eigendirections.

distinguish between conceptual and perceptual features, the interpretation of the discovered features remains as a second independent step in data analysis.

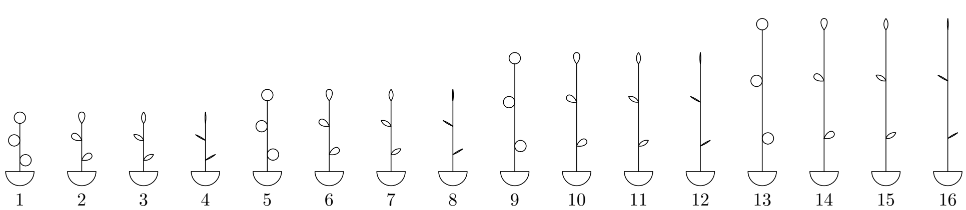

Figure 10: Images of the flowerpots as presented to the test person.

We explain the flowerpot experiment according to the model presented in Section 3.3, starting from a uniform distribution of 16 points in three dimensions and choosing the feature vectors f_k, $k = 1, 2, 3$ to be the unit vectors $e_1 = (1, 0, 0)$ etc.

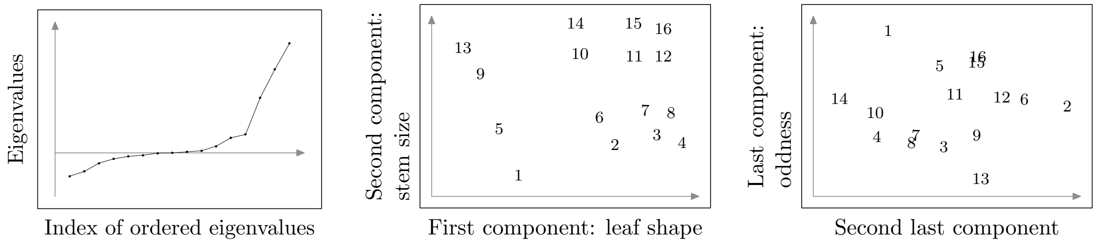

Figure 11: Left: Sorted eigenvalues. Middle: Projection onto the leading two positive eigenvalues. Right: projection onto the last two eigendirections.

The states are obtained by fitting the d defined in Equation 2 heuristically to the experimental dissimilarity by minimization of the difference of the means over all matrix elements.

We obtain a good model fit for six states $\{(8.3,0,0), (0,3.5,0), (4.7,4.7,4.7), (6.4,6.4,0), (0,3.4,3.4), (3.1,0,3.1)\}$. See Figure 12.

In other words, following the semantics of the model presented, one can explain the results of the obtained dissimilarities by six perceptual states of the observer; i.e. the weight vectors model the bias in perception. These seem to outnumber the actually observed features (in the two-dimensional representations) which are three in number (the two geometric features in the positives and the categorial one in the negatives). However, we must keep in mind that one may reduce the number of weights required to approximate d by a deeper knowledge of the initial feature presentation, including its dimensionality. We have taken a uniform distribution in three dimensions for lack of more precise knowledge.

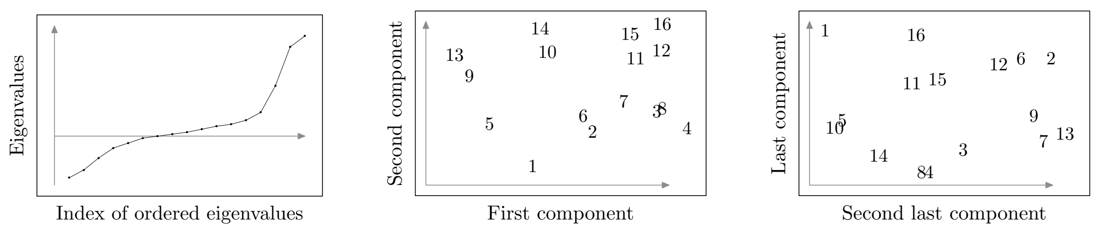

Figure 12: Prediction of flowerpot experiment.

6. Conclusion and Outlook

This work studies the potential of relevant information being coded specifically by the non-metric part of the spectrum of a pseudo-covariance matrix. It has been shown that non-metricity *can* indeed code for features relevant to a better understanding of the data set. The proposed algorithm effectively overcomes the drawback of most variance based algorithms which take only into account the variance of the leading eigendirections. Model.illustrations provide a simple intuition and explanation of the phenomena. Note, however, that spectra like Figure 1 (right) are only potentially—not necessarily—containing interesting information in their negative part, some fancy noise process might also be the cause of such a structure.

Concluding, it is an important step in unsupervised data analysis to find out whether the negative part of the spectrum codes for interesting variance. The present technique can be employed as a general exploratory feature discovery tool. Application fields range from cognitive psychology, marketing, biology, engineering and bioinformatics, where in principle new structure awaits its discovery.

A further interesting direction is to go beyond visualization toward automated structure learning. Investigations to overcome low-dimensional feature discovery based on visualization will focus on, e.g., stability analysis (Roth et al., 2002; Meinecke et al., 2002) of various projections onto the possibly negative eigenspace in order to assess quantitatively relevant structure and to rule out noise related, erroneous, feature interpretations.

A major focus will concern the automated distinction of structure induced by intrinsic non-metricity from mere artifacts of some fancy noise process with the overall goal to provide automated learning and procedures that can optimally make use of the information coded by intrinsic non-metricity.

Acknowledgments

Special thanks go to Volker Roth for valuable discussions, as well as to Motoaki Kawanabe, Christin Schäfer, Sebastian Mika, Mikio Braun and Andreas Ziehe for thorough reading of the manuscript. This work is partially supported by DFG grants # MU 987/1-1 and # JA 379/13-2 as well as PASCAL Network of Excellence (EU # 506778).

A particular acknowledgment goes to the reviewers and the action editor who helped with many stringent and interesting remarks and suggestions improve this work.

References

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

A. Banerjee and J. Ghosh. Clickstream clustering using weighted longest common subsequences. *In Proceedings of the Web Mining Workshop at the 1 st SIAM Conference on Data Mining, Chicago, April 2001.*, 2002.

T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 2001.

I. Dagan, S. Marcus, and S. Markovitch. Contextual word similarity and extimation from sparse data. *Computer Speech and Language*, (9(2)):123–152, 1995.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, second edition, 2001.

B. S. Everitt and S. Rabe-Hesketh. *The Analysis of Proximity Data*. Arnold, London, 1997.

I. Gati and A. Tversky. Representation of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 8(2):325–340, 1982.

R. L. Goldstone, D. L. Medin, and D. Gentner. Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, pages 222–262, 1991.

Project Gutenberg. `http://promo.net/pg/`.

T. Hofmann, J. Puzicha, and J. M. Buhmann. Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8): 803–818, 1998.

D. W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1964.

F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. *IEEE Transactions on Biomedical Engineering*, 49:1514–1525, 2002.

NASA. `http://science.nasa.gov/headlines/news_archive.htm`.

D. J. Navarro and M. D. Lee. Common and distinctive features in stimulus similarity: A modified version of the contrast model. *submitted*, 2002.

E. Pękalska, P. Paclík, and R. P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, (2):175–211, 2001.

L. M. Rocha. Talkmine: A soft computing approach to adaptive knowledge recommendation. *Soft Computing Agents: New Trends for Designing Autonomous Systems*, pages 89–116, 2001.

L. M. Rocha and J. Bollen. Biologically motivated distributed designs for adaptive knowledge management? *Design Principles for the Immune System and other Distributed Autonomous Systems*, pages 305–334, 2001.

V. Roth, T. Lange, M. Braun, and J. M. Buhmann. A resampling approach to cluster validation. *Statistics–COMPSTAT*, pages 123 – 128, 2002.

V. Roth, J. Laub, J. M. Buhmann, and K.-R. Müller. Going metric: Denoising pairwise data. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 817–824. MIT Press: Cambridge, MA, 2003a.

V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann. Optimal cluster preserving embedding of non-metric proximity data. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25 (12):1540–1551, 2003b.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2000.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

R. N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27, 1962.

J. B. Tenenbaum. Learning the structure of similarity. *Advances in Neural Information Processing Systems*, 1996.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2000.

M. S. C. Thomas and D. Mareschal. Connectionism and psychological notions of similarity. *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, 1997.

W. S. Torgerson. *Theory and Methods of Scaling*. John Wiley and Sons, New York, 1958.

G. Young and A. S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22, 1938.