Seventh Framework Programme

Information and Communication Technologies

FET Open - Collaborative Project

Technical Report n. 2009_9

November 2009

Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Project reference number	213250
Authors	Robert P.W. Duin and Elzbieta Pekalska
Title	Datasets and tools for dissimilarity analysis in pattern recognition
Month	November 2009
web site	http://simbad-fp7.eu/techreports.php

Abstract

This report presents a summary on a set of proximity datasets collected within the SIMBAD project. This is done in order to enable a further study of general symmetric similarity and dissimilarity data matrices. The sets are either artificially generated or derived as proximity measurements from real world experiments. As these datasets are based on unconstrained proximity measures, the resulting similarity matrices represent indefinite measures while the dissimilarity matrices have a non-Euclidean behavior, i.e. they cannot be isometrically embedded in Euclidean spaces. Various indices are derived and computed to characterize the datasets. Also, a number of ways are used to convert the proximity matrices into Euclidean distance matrices so that they can be represented in Euclidean vector spaces.

Initial classification experiments are performed always in the same, consistent way for all datasets and for a number of dissimilarity-based embedded spaces. Various scatterplots are presented by which the usefulness of the various Euclidean conversions can be evaluated.

The report is just an introduction to the collected datasets. It presents possible ways that can be used for their characterization and comparison. It does not provide specific or definite scientific conclusions. Its target is to show that the collected datasets show a large variability of characteristics. Thereby, they constitute a good set for further study of the analysis and application of dissimilarity data.

Contents

1 Introduction

A collection of (dis)similarity datasets is described and analyzed in the coming sections. Our own Matlab packages PRTools and DisTools are used for this purpose. PRTools is a general package for statistical pattern recognition, while the DisTools package resulted from previous projects studying dissimilarity data. The toolbox is however significantly upgraded for the analysis presented here.

The results for all datasets are presented always in the same way, starting with a short description and references and followed by data characteristics, and classification results. Below, we will explain the notation and briefly summerize how the sections are built.

Some datasets are public domain obtained from the Internet data sources. Other datasets are either especially generated by us for illustration purposes or received via private communication that resulted from our previous research. Data descriptions are copied, summarized or generated by us and possible references are listed. Whenever applicable, the original web location is mentioned. In addition, we always report the Matlab version we used in the PRTools format.

Proximity datasets come either as similarity data $k(i,j)$ or dissimilarity data $d(i,j)$ and we always report the original type. Since we focus on dissimilarity data, similarities $k(i,j)$ are converted to the corresponding dissimilarities $d(i,j)$ by the following equation

$$d(i,j) = \sqrt{k(i,i) + k(j,j) - k(i,j) - k(j,i)} \tag{1}$$

All the dissimilarity matrices are then scaled such that the average dissimilarity is one, i.e.:

$$\frac{d(i,j)}{\frac{1}{m(m-1)} \sum_{i,j} d(i,j)} = 1 \tag{2}$$

This is done to assure that the results are comparable over the datasets as we deal with dissimilarity data in various ranges and scales. Such scaled dissimilarities are denoted as $\tilde{d}$. In addition, we assume here that the dissimilarities are symmetric. So, every dissimilarity $\tilde{d}(i,j)$ has been transformed by

$$\tilde{d}(i,j) := \frac{\tilde{d}(i,j) + \tilde{d}(j,i)}{2} \tag{3}$$

In order to characterize dissimilarity sets in the same way, we compute a number of indices that focus on various aspects of the data. These are:

asymmetry coefficient

The asymmetry γ is defined over similarities as well as dissimilarities. In the latter case it becomes $\tilde{d}(i,j)$ as

$$\gamma = \sum_{i \neq j} \frac{|\tilde{d}(i,j) - \tilde{d}(j,i)|}{|\tilde{d}(i,j) + \tilde{d}(j,i)|} \tag{4}$$

number of objects

It is the number of rows m of a dissimilarity matrix (must be equal to the number of columns).

number of significant eigenvectors

The dissimilarity matrix D is embedded in a pseudo-Euclidean (PE) space of $m-1$ dimensions. Let L be a vector consisting of the $m-1$ eigenvalues found by the embedding. We list the minimum number of components L_i of L for which the following holds:

$$\frac{\sum_{i<m} |L_i|}{\sum_i |L_i|} > 0.95 \tag{5}$$

number of triangle inequality violations

There are $m(m-1)(m-2)$ triangle inequalities. We report this quantity as well as the total number of violations.

numbers of positive and negative eigenvalues

There are p positive eigenvalues and q negative eigenvalues from the total $m-1$ eigenvalues L found in the PE embedding. The so-called signature $[p, q]$ is listed.

negative eigenfraction, NEF

This is a measure for the non-Euclidean behavior. It is the ratio of absolute sum of all negative eigenvalues and the sum of all absolute eigenvalues:

$$NEF = \frac{\sum_{i:L_i<0} |L_i|}{\sum_i |L_i|} \tag{6}$$

negative eigenratio NER

This is another measure of the non-Euclidean behavior. It is the ratio of the largest negative to the largest positive eigenvalue.

$$NER = \frac{\max_i |L_i < 0|}{\max_i (L_i > 0)} \tag{7}$$

number of classes, class sizes

The number of classes and the class sizes follow from the given labelling of the objects.

average within-class and between-class dissimilarities

The average within-class dissimilarity is the mean of all dissimilarities between different objects of the same class. The average between-class dissimilarity is the mean of all dissimilarities of objects of different classes. Due to the normalization (2), the former number is usually somewhat smaller than one, while the latter number is usually larger than one.

LOO nearest neighbor errors in the PE Space, Pos Space and Dis Space

These errors are computed in a leave-one-out (LOO) fashion. The nearest neighbor error in the PE space is computed on the original dissimilarity matrix, as the distances in the PE space are identical to the original dissimilarities as we use a full embedding. The Pos Space is a Euclidean space determined by the positive eigenvectors of the PE embedding only. The Dis Space is the dissimilarity space, which is a Euclidean space whose dimensions are defined by the columns of the dissimilarity matrix; see **?**.

In addition to these characteristic indices we show classification results for four classifiers in ten spaces. This is done for majority of the data. In order to make these results comparable over datasets of different sizes, ten random subsets of 50 objects per class were selected for each data. These experiments are only executed for problems in which the smallest class is larger than 60 objects. Hence, there are no classification results for some data, e.g. Protein data.

Ten vector spaces are defined for every subset of 50 objects per class by **all** objects in the subset. (This means that objects from both training and test sets are used to define the given space. Note that this approach is different than the one taken in many of our earlier studies on proximity-based learning.) Then, we perform 2-fold crossvalidation, averaged over the 10 random selected subsets. The final results are presented as average classification errors in all the spaces. The following spaces are considered (remember that m is the number of all objects):

PE Space

This is an $(m-1)$-dimensional pseudo-Euclidean (PE) space defined by $m-1$ eigenvectors, properly scaled by the corresponding eigenvalues, both determined in the embedding of

the dissimilarity data into the PE space. Classifiers in this space make use of the specific (non-degenerate and indefinite) inner product definition of this space, which also leads to a PE distance.

Ass Space
The $(m-1)$-dimensional associated space is defined by the same vector space as the PE Space, but is equipped with the traditional Euclidean inner product and the Euclidean distance measure. The 'negative' subspace of the PE space is thereby used as a 'positive' one.

Pos Space
This is the p-dimensional positive subspace of the complete PE space. It is defined by the p eigenvectors scaled appropriately by the corresponding p positive eigenvalues. If the eigenvectors corresponding to the q negative eigenvalues are the result of noise or class unrelated disturbances, it is expected that classifiers in the Pos Space perform better than in the PE Space.

Neg Space
This is the q-dimensional negative subspace of the complete PE space defined by the q eigenvectors suitably scaled by the corresponding magnitudes of the negative eigenvalues. Here, we neglect its 'negative' nature and treat this subspace as a Euclidean space. This means that distances are computed in the traditional Euclidean way. In order to study whether the negative part is possibly informative in itself, we also perform classification experiments in the negative space only.

Cor Space
There are several ways to transform a pseudo-Euclidean space into a Euclidean space. This is important as it makes a large toolbox of classifiers available. However, this might be paid by a decrease in performance. We use a simple and well defined correction from a pseudo-Euclidean space to a Euclidean one. It relies on transforming the given (scaled) dissimilarity matrix into a Euclidean distance matrix and then by determining the corresponding embedding (by the use of classical scaling). This is realized by adding twice the magnitude of the smallest eigenvalue $L_{\min} = \min_{i<m} L_i$ (form the PE embedding) to all off-diagonal elements of the squared dissimilarity data:

$$\tilde{d}_{\mathrm{cor}}(i,j) = \sqrt{d^2(i,j) + 2|L_{\min}|}, \quad i \neq j. \tag{8}$$

This corresponds to a Euclidean space defined by the eigenvectors from the PE embedding, but scaled by the square roots of the transformed eigenvalues $L_i + |L_{\min}|$. Consequently, the dimensions of the PE space scaled by the smallest (i.e. largest in magnitude) negative eigenvalues contribute the least in terms of inner products and distances in the corrected Euclidean space. Note also that the 1-nearest neighbor and Parzen classifiers perform identically to the PE space.

All the above spaces are variants of the original PE space. In addition, we also study various dissimilarity spaces. Such spaces postulate Euclidean spaces whose dimensions (features) are defined by dissimilarities to individual objects, $d(\cdot, x_i)$. Several distances or dissimilarities can be chosen for this purpose. Among others, distances computed between the objects in each of the above five spaces can be used to build a dissimilarity space. The dissimilarity space based on the PE Space uses the original dissimilarities. It is what we called the "dissimilarity space" in our earlier publications (and above just Dis Space).

In the bottom tables, the five dissimilarity spaces are denoted as **PE Dis Space**, **Ass Dis Space**, **Pos Dis Space**, **Neg Dis Space** and **Cor Dis Space**. For example, the Ass Dis Space means

that first, the Associate Space is derived from the PE embedding as described above. Then, the Euclidean distances are recomputed between all the objects in that space. The resulting Euclidean distance matrix is used to define a dissimilarity space, named as Ass Dis Space.

As above, also these experiments rely on a full embedding that combines both training and test sets. Note also that **no feature reduction** has been applied. This is done so to assure a fair comparison between spaces and between datasets. Better classification results may be (and often are) obtained for some classifiers by a proper regularization or prototype/feature selection. We, however, decided to skip it in the presented summary as such a regularization may not be optimal and moreover will differ from space to space and vary over different datasets.

The following eight graphs are presented for each dataset.

(a) The dissimilarity matrix as an intensity image. Zero dissimilarities are encoded as black pixels and the lighter the pixels the higher the dissimilarity value. Objects belonging to the same class are grouped together.

(b) A scatterplot of the first two positive eigenvectors of the PE embedding. Different classes are denoted by different markers.

(c) Learning curves of the 1-nearest neighbor in the five embedded spaces, i.e. in the PE space and its variants. First, the embedding is determined on the complete data subset of the given size. The learning curves are based on a theoretical computation of the expected nearest neighbor classification error for training sets of the given size. The curves are smooth because this result is equivalent (but much faster) to drawing an infinite number of random training sets. In addition, the learning curve for the PE Dis Space is drawn in this graph as well in order to provide a basic comparison to a dissimilarity space result.

(d) Learning curves of the 1-nearest neighbor errors in the dissimilarity spaces. Similarly as above, the spaces are first defined by using all the training and test data in the given subset and then the errors are computed in these spaces. The learning curve for the PE Space is drawn in this graph, as well, for a basic comparison with the embedded space.

(e) The spectrum of the eigenvalues found in the PE embedding of the complete dissimilarity data into the PE Space. They are ranked from the most positive to the most negative eigenvalue. Negative values are marked in red.

(f) The negative eigenfraction (NEF) is a function of the size of the data. We draw a sufficient number of random subsets to have an accuracy of at least 95

(g) Subsets of N points may either be perfectly embedded into a Euclidean space or not. For non-metric dissimilarity datasets such subsets can be found even for $N = 3$. For metric but non-Euclidean datasets such subsets may be found for $N > 3$. Here, in this figure, we report the probability that a random subset of N points from a given dissimilarity set **does not** perfectly fit into a Euclidean space.

(h) This figure shows two histograms of dissimilarities. The red histogram shows the dissimilarities of objects in the same class. It is drawn on top of the blue histogram of all dissimilarities. The visible blue part corresponds thereby to the between-class dissimilarities.

2 Dataset Descriptions

Next subsections present a detailed description of various datasets based on either similarities or dissimilarities. Some of them are artificially generated, some rely on real data created inside the SIMBAD project, while others are obtained thanks to research collaborations or from public domain collections found on the Internet.

Artificial datasets. There are two datasets, **GaussM1** and **GaussM02** defined by non-Euclidean distance measures in a vector space. We also randomly generated sets of artificial objects: polygons, related by the standard and the modified Hausdorff distance (**PolyDisH57** and **PolyDisM57**) and balls of various sizes related by their spatial distances in a hypercube (**Balls3D** and **Balls50D**).

SIMBAD real world datasets. The York and Delft groups study graph matching. They used the same graphs obtained from feature points in a set of object images (obtained from the COIL dataset) and computed various dissimilarity measures: **CoilYork**, **CoilDelftDiff** and **CoilDelftSame**.

Donated real world datasets. The following datasets are obtained from various collaborations:
DelftGestures, based on dynamic time warp distances between gestures,
FlowCyto, four sets of histogram distances between flow-cytometry measurements of possible tumor cells,
ProDom, Similarities between protein sequences,
WoodyPlants50, a subset of a large dataset comparing the shapes of leaves,
Zongker, distances between handwritten digits based on deformable templates.

Public domain real world datasets. The following datasets can be found on the Internet.
Chickenpieces, A family of 44 datasets based on shape distances computed by different weighted edit distances,
NewsGroups, A non-metric correlation measure between newsgroup messages,
CatCortex, the connection strengths between cortical areas of a cat,
Protein, a comparison of protein sequences based on the concept of an evolutionary distance.

References, links and other details are presented in the following subsections.

2.1 Balls3D

Description Dissimilarity Dataset

This dataset has been generated by the DisTools command GENBALLD([100 100], 3, [0.02 0.04]) which generates the given numbers of 3-D balls with sizes [0.02 0.04] in a3-D hypercube. Balls do not overlap. Dissimilarities are computed as the shortest distance between two points on the surface of two balls. The intention is to study strong examples in which non-Euclidean dissimilarities are informative.

Reference(s)

E. Pekalska, A. Harol, R.P.W. Duin, D. Spillman, and H. Bunke, Non-Euclidean or non-metric measures can be informative, in: D.-Y. Yeung et al., Proc. SSSPR2006 Lecture Notes in Comp. Sc., vol. 4109, Springer, Berlin, 2006, 871-880.

R.P.W. Duin, E. Pekalska, A. Harol, W.J. Lee, and H. Bunke, On Euclidean corrections for non-Euclidean dissimilarities, in: N. da Vitoria Lobo et al., Proc. SSSPR2008, Lecture Notes in Comp.Sc., vol. 5342, Springer, Berlin, 2008, 551-561.

J. Laub, V. Roth, J.M. Buhmann, K.R. Mueller, On the information and representation of non-euclidean pairwise data, Pattern Recognition, vol. 39, 2006, 1815-1826.

Web page(s)

The PRTools version of the data: http://prlab.tudelft.nl/data/disdatasets.html

The DisTools package which contains GENBALLD: http://prlab.tudelft.nl/software/DisTools.html

0.000	asymmetry
200	number of objects
2	number of significant eigenvectors
1266	number of triangle inequality violations out of 7880400
9, 190	number of positive and negative eigenvalues
0.001	negative eigenfraction
0.000	negative eigenratio
2	number of classes, class sizes [100 100]
0.999, 1.001	average within-class and between-class dissimilarity
43.5, 43.5, 48.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 1: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	47.4 (2.0)	47.4 (2.0)	47.4 (2.0)	44.2 (1.5)	47.4 (2.0)
Parzen	45.7 (1.7)	45.5 (1.6)	45.6 (1.7)	35.5 (1.7)	45.7 (1.7)
NM	47.5 (2.0)	47.7 (2.0)	47.6 (1.9)	49.6 (0.2)	48.1 (1.8)
SVM-1	50.7 (2.2)	50.0 (2.7)	50.0 (2.5)	62.1 (1.7)	50.1 (2.0)

Table 2: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	49.8 (2.2)	49.8 (2.2)	49.8 (2.2)	5.1 (0.8)	49.7 (2.2)
Parzen	47.9 (2.2)	47.9 (2.2)	47.9 (2.2)	4.6 (0.5)	47.9 (2.2)
NM	49.8 (2.2)	49.8 (2.2)	49.8 (2.2)	5.0 (0.8)	49.9 (2.2)
SVM-1	50.2 (1.6)	50.8 (1.7)	50.7 (1.7)	1.9 (0.5)	49.8 (1.5)

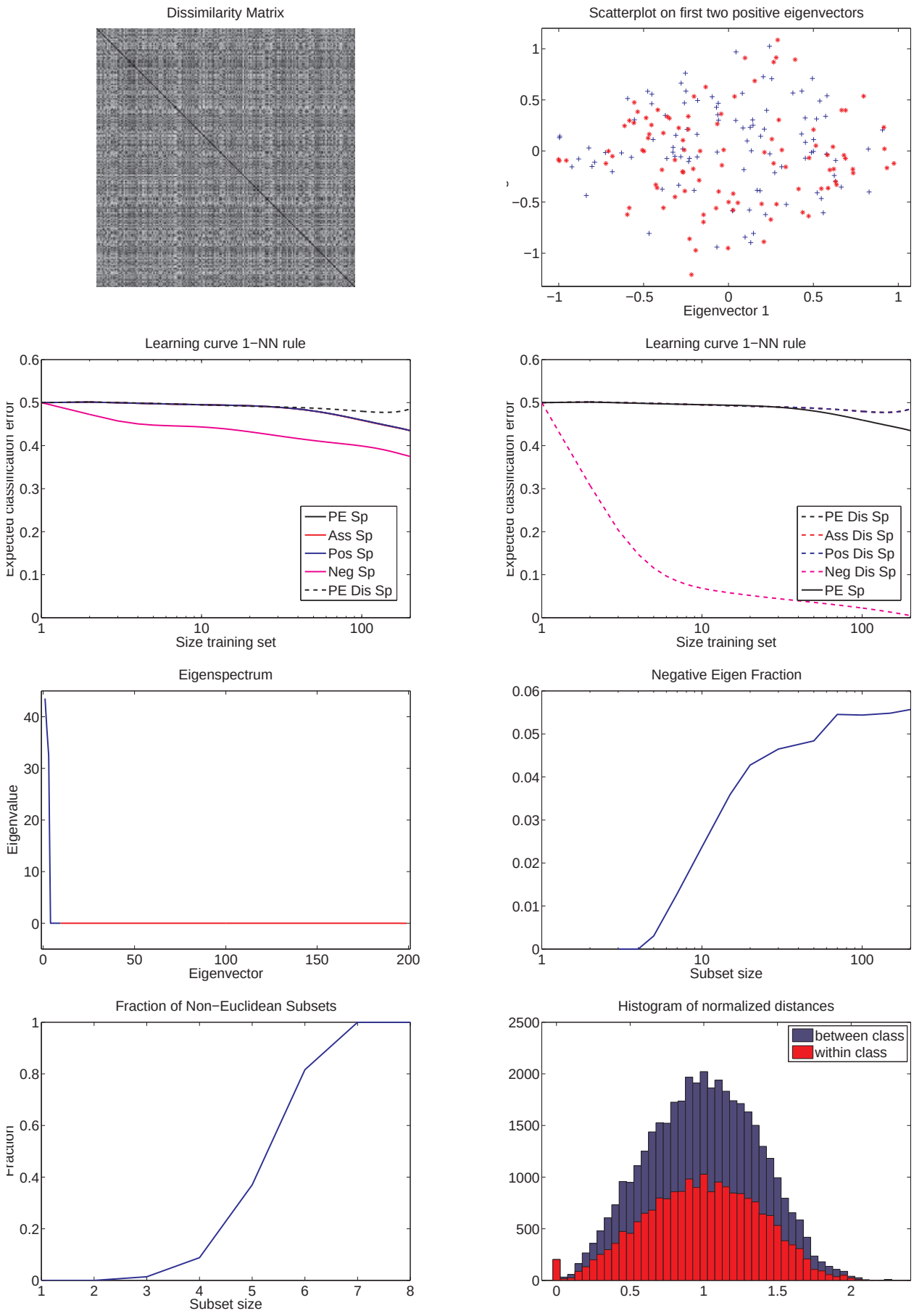


Figure 1: Graphical results for Balls3D.

2.2 Balls50D

Description Dissimilarity Dataset

This dataset has been generated by the DisTools command GENBALLD([500 500 500 500], 50, [0.01 0.02 0.04 0.08]) which generates the given numbers of 50-D balls with sizes [0.01 0.02 0.04 0.08] in a50-D hypercube. Balls do not overlap. Dissimilarities are computed as the shortest distance between two points on the surface of two balls. The intention is to study strong examples in which non-Euclidean dissimilarities are informative.

Reference(s)

E. Pekalska, A. Harol, R.P.W. Duin, D. Spillman, and H. Bunke, Non-Euclidean or non-metric measures can be informative, in: D.-Y. Yeung et al., Proc. SSSPR2006 Lecture Notes in Comp. Sc., vol. 4109, Springer, Berlin, 2006, 871-880.

R.P.W. Duin, E. Pekalska, A. Harol, W.J. Lee, and H. Bunke, On Euclidean corrections for non-Euclidean dissimilarities, in: N. da Vitoria Lobo et al., Proc. SSSPR2008, Lecture Notes in Comp.Sc., vol. 5342, Springer, Berlin, 2008, 551-561.

J. Laub, V. Roth, J.M. Buhmann, K.R. Mueller, On the information and representation of non-euclidean pairwise data, Pattern Recognition, vol. 39, 2006, 1815-1826.

Web page(s)

The PRTools version of the data: http://prlab.tudelft.nl/data/disdatasets.html

The DisTools package which contains GENBALLD: http://prlab.tudelft.nl/software/DisTools.html

0.000	asymmetry
2000	number of objects
46	number of significant eigenvectors
0	number of triangle inequality violations out of 7988004000
51, 1948	number of positive and negative eigenvalues
0.000	negative eigenfraction
0.000	negative eigenratio
4	number of classes, class sizes [500 500 500 500]
1.000, 1.000	average within-class and between-class dissimilarity
74.2, 74.2, 74.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 3: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	73.4 (1.3)	73.4 (1.3)	73.4 (1.3)	75.0 (0.0)	73.4 (1.3)
Parzen	74.7 (1.2)	74.7 (1.2)	74.7 (1.2)	64.3 (1.4)	74.7 (1.2)
NM	74.3 (1.1)	74.3 (1.1)	74.3 (1.1)	75.0 (0.0)	74.3 (1.1)
SVM-1	73.9 (1.0)	73.9 (1.1)	74.0 (1.1)	75.0 (0.0)	73.9 (1.0)

Table 4: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	74.7 (1.1)	74.8 (1.0)	74.7 (1.1)	1.0 (0.2)	74.8 (1.0)
Parzen	74.3 (1.1)	74.2 (1.1)	74.2 (1.1)	1.5 (0.2)	74.3 (1.1)
NM	74.6 (1.1)	74.7 (1.0)	74.6 (1.1)	0.5 (0.1)	74.6 (1.1)
SVM-1	72.8 (1.0)	72.8 (1.0)	72.8 (1.0)	0.5 (0.1)	72.8 (1.0)

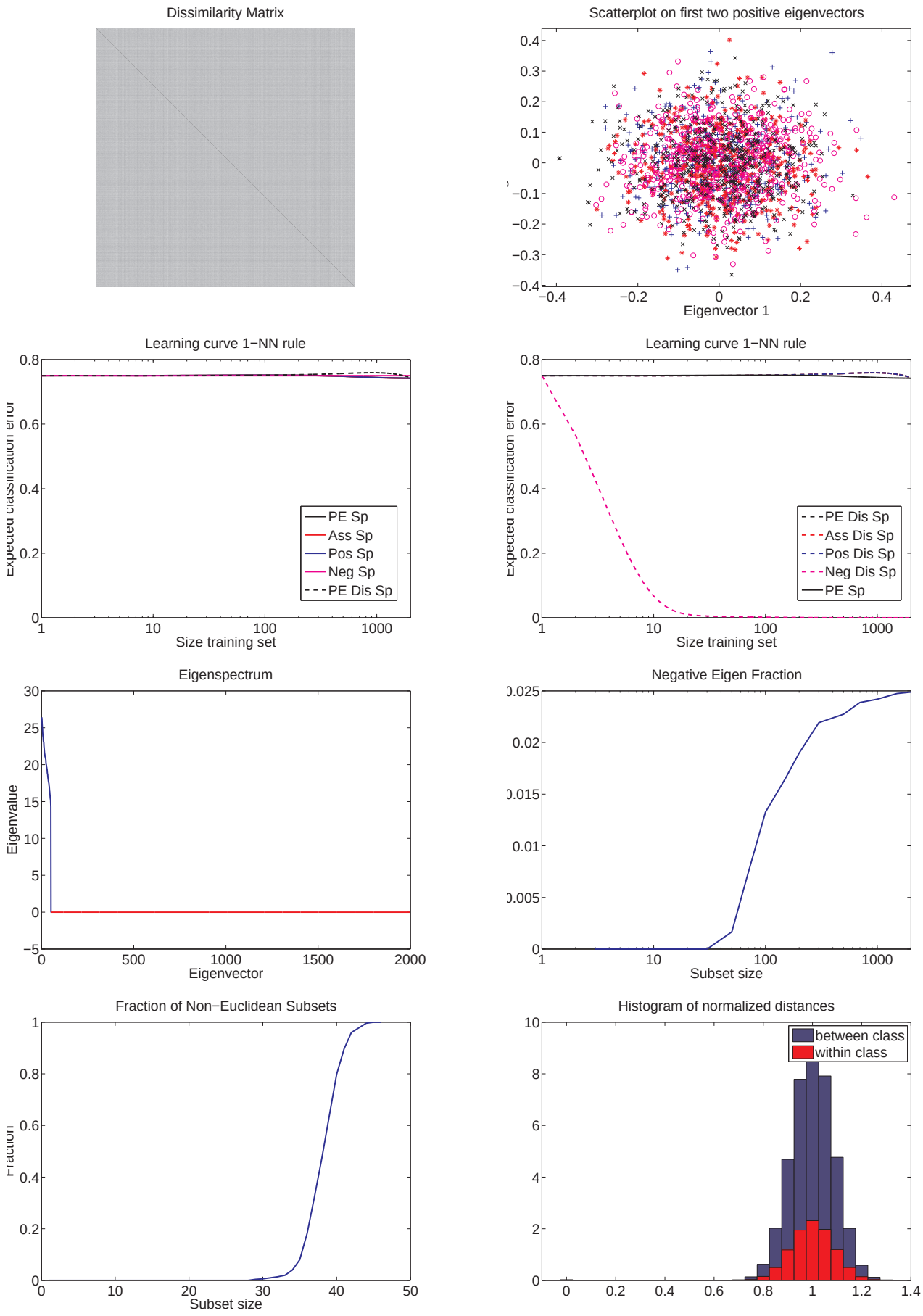


Figure 2: Graphical results for Balls50D.

2.3 CatCortex

Description Dissimilarity Dataset

The cat-cortex data set is provided as a 65x65 dissimilarity matrix describing the connection strengths between 65 cortical areas of a cat from four regions (classes): auditory (A), frontolimbic (F), somatosensory (S) and visual (V). The data was collected by [Scannell] and used for classication [Graepel] and clustering [Denoeux and Masson]. The dissimilarity values are measured on an ordinal scale.

Reference(s)

J. Scannell, C. Blakemore, M. Young, Analysis of connectivity in the cat cerebral cortex. Journal of Neuroscience, vol. 15, 1463-1483, 1995.

T. Graepel, R. Herbrich, P. Bollmann-Sdorra, K. Obermayer, Classification on pairwise proximity data. In Advances in Neural Information System Processing vol. 11, 438-444, 1999.

T. Denoeux, T. and M.-H. Masson, EVCLUS: Evidential clustering of proximity data. IEEE Transations on Systems, Man and Cybernetics, vol. 34, 95-109, 2004.

Web page(s)

The original data http://www.hds.utc.fr/ tdenoeux/software.htm

The PRTools version http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
65	number of objects
44	number of significant eigenvectors
286	number of triangle inequality violations out of 262080
41, 23	number of positive and negative eigenvalues
0.208	negative eigenfraction
0.272	negative eigenratio
4	number of classes, class sizes [10 19 18 18]
0.802, 1.066	average within-class and between-class dissimilarity
12.3, 6.2, 7.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

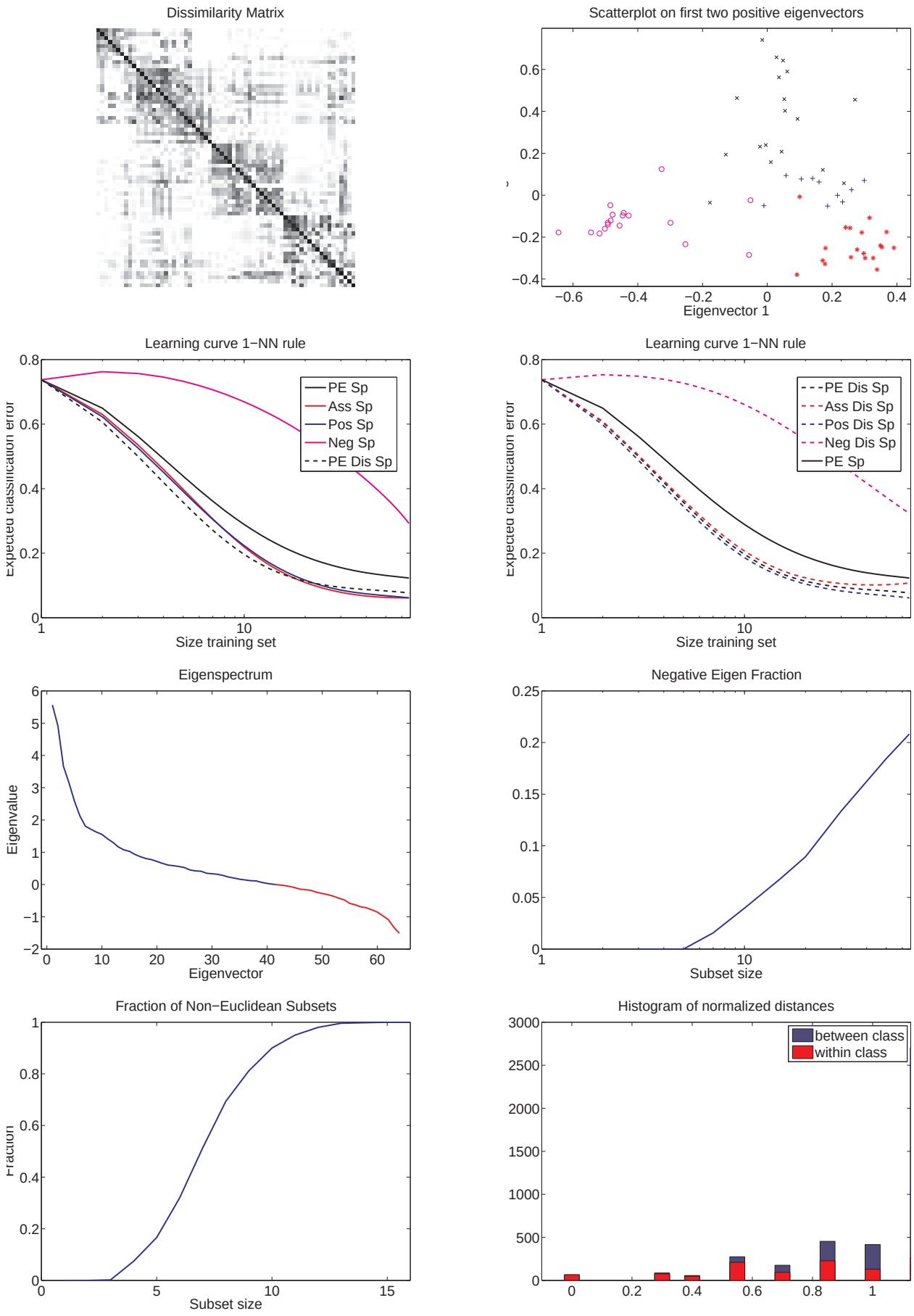


Figure 3: Graphical results for CatCortex.

2.4 CoilDelftDiff

Description Dissimilarity Dataset

This is a dissimilarity matrix between set of graphs derived from four objects of the COIL database computed by Richard Wilson. The graphs are the Delaunay triangulations derived from corner points found in these images, see [Xia and Hancock]. Graphs are compared in the eigenspace with a dimensionality determined by the smallest graph in every pairwise comparison by the JoEig approach. see [Lee and Duin].

Reference(s)

B. Xiao and E. R. Hancock. Geometric characterisation of graphs. ICIAP 2005, LNCS 3617:471-478, 2005

W.J. Lee and R.P.W. Duin, An Inexact Graph Comparison Approach in Joint Eigenspace, in: N. da Vitoria Lobo et al. (eds.), Proc. SSSPR2008, Lecture Notes in Computer Science, vol. 5342, Springer, Berlin, 2008, 35-44.

Web page(s)

The PRTools version of the data: http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
288	number of objects
168	number of significant eigenvectors
1	number of triangle inequality violations out of 23639616
163, 124	number of positive and negative eigenvalues
0.128	negative eigenfraction
0.051	negative eigenratio
4	number of classes, class sizes [72 72 72 72]
0.967, 1.011	average within-class and between-class dissimilarity
47.2, 47.6, 42.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 5: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	50.8 (1.0)	52.4 (0.8)	51.2 (0.6)	78.5 (0.9)	50.8 (1.0)
Parzen	48.7 (1.1)	48.7 (1.0)	48.8 (1.0)	80.9 (0.9)	48.7 (1.1)
NM	51.0 (0.7)	50.2 (0.7)	50.1 (0.7)	76.5 (0.5)	50.9 (0.8)
SVM-1	50.3 (0.9)	44.4 (1.0)	45.1 (0.9)	68.0 (1.4)	45.5 (0.6)

Table 6: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	47.1 (0.7)	47.6 (1.1)	46.9 (0.8)	69.8 (1.2)	47.0 (0.7)
Parzen	50.0 (1.2)	52.3 (0.8)	50.7 (1.0)	72.5 (1.0)	48.9 (1.3)
NM	46.5 (0.8)	47.5 (1.2)	46.7 (0.9)	70.0 (1.1)	45.9 (0.8)
SVM-1	41.4 (1.2)	41.7 (0.9)	41.3 (1.1)	71.7 (1.6)	44.6 (0.8)

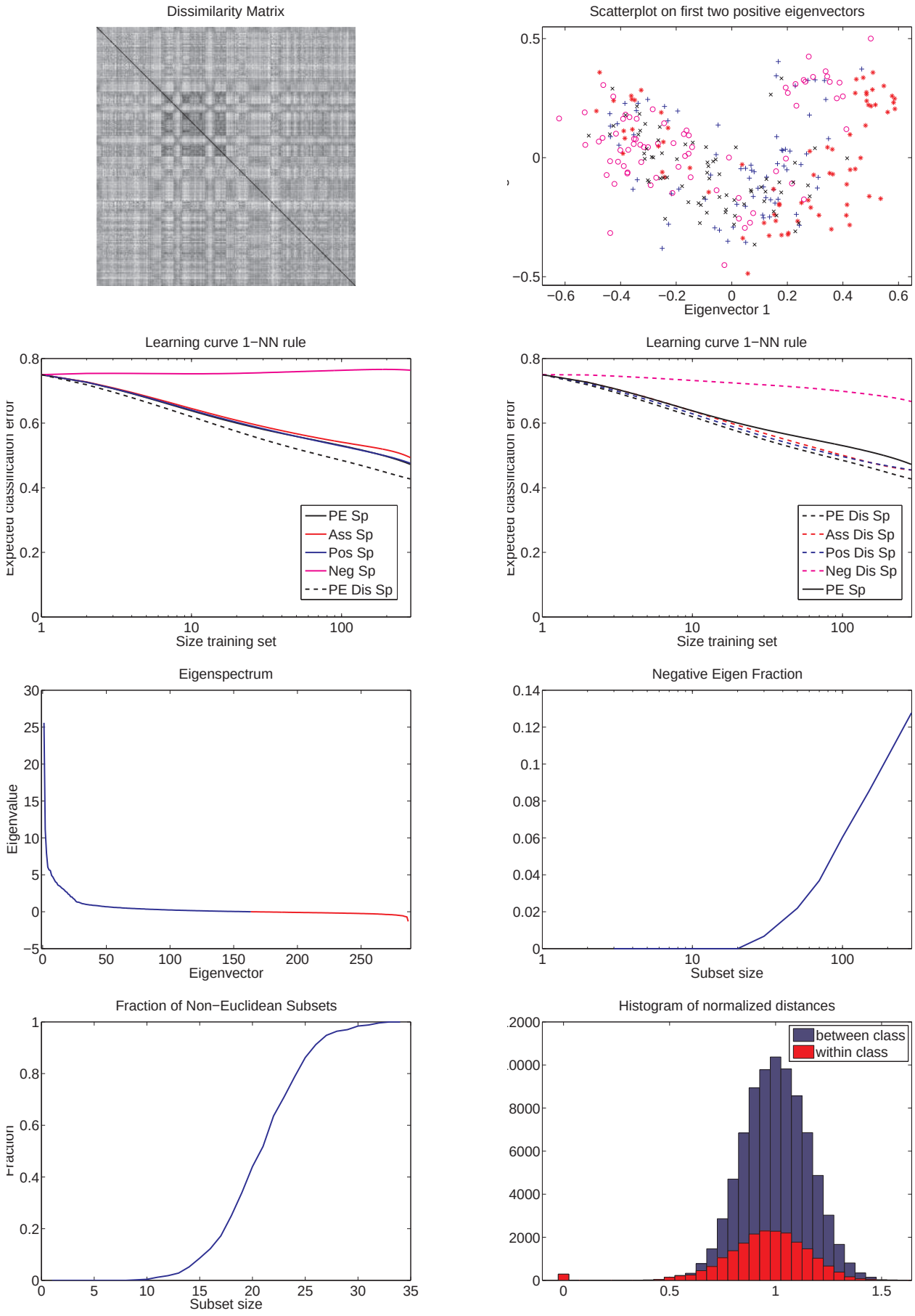


Figure 4: Graphical results for CoilDelftDiff.

2.5 CoilDelftSame

Description Dissimilarity Dataset

This is a dissimilarity matrix between set of graphs derived from four objects of the COIL database computed by Richard Wilson. The graphs are the Delaunay triangulations derived from corner points found in these images, see [Xia and Hancock]. Distances are obtained in a pairwise fashion in a 5D space of eigenvectors derived from the two graphs by the JoEig approach, see [Lee and Duin].

Reference(s)

B. Xiao and E. R. Hancock. Geometric characterisation of graphs. ICIAP 2005, LNCS 3617:471-478, 2005

W.J. Lee and R.P.W. Duin, An Inexact Graph Comparison Approach in Joint Eigenspace, in: N. da Vitoria Lobo et al. (eds.), Proc. SSSPR2008, Lecture Notes in Computer Science, vol. 5342, Springer, Berlin, 2008, 35-44.

Web page(s)

The PRTools version of the data: http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
288	number of objects
158	number of significant eigenvectors
0	number of triangle inequality violations out of 23639616
249, 38	number of positive and negative eigenvalues
0.027	negative eigenfraction
0.181	negative eigenratio
4	number of classes, class sizes [72 72 72 72]
0.978, 1.007	average within-class and between-class dissimilarity
64.6, 60.8, 38.9	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 7: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	65.8 (0.7)	63.6 (1.1)	64.1 (0.8)	68.0 (1.3)	65.8 (0.7)
Parzen	54.0 (1.3)	48.3 (1.1)	49.7 (1.2)	61.2 (0.8)	54.0 (1.3)
NM	64.6 (1.4)	65.0 (1.0)	64.5 (1.6)	65.8 (2.0)	64.7 (1.9)
SVM-1	71.6 (1.1)	51.6 (1.4)	62.1 (1.2)	63.2 (1.4)	59.8 (1.2)

Table 8: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	41.3 (0.6)	41.9 (0.5)	41.3 (0.6)	65.8 (1.4)	45.0 (1.0)
Parzen	51.5 (1.3)	51.5 (1.2)	51.8 (1.2)	60.1 (0.9)	48.5 (0.7)
NM	41.5 (0.6)	41.4 (0.6)	40.9 (0.6)	63.3 (1.6)	44.3 (0.7)
SVM-1	44.3 (1.3)	42.9 (1.1)	43.8 (1.3)	65.5 (1.1)	44.9 (1.1)

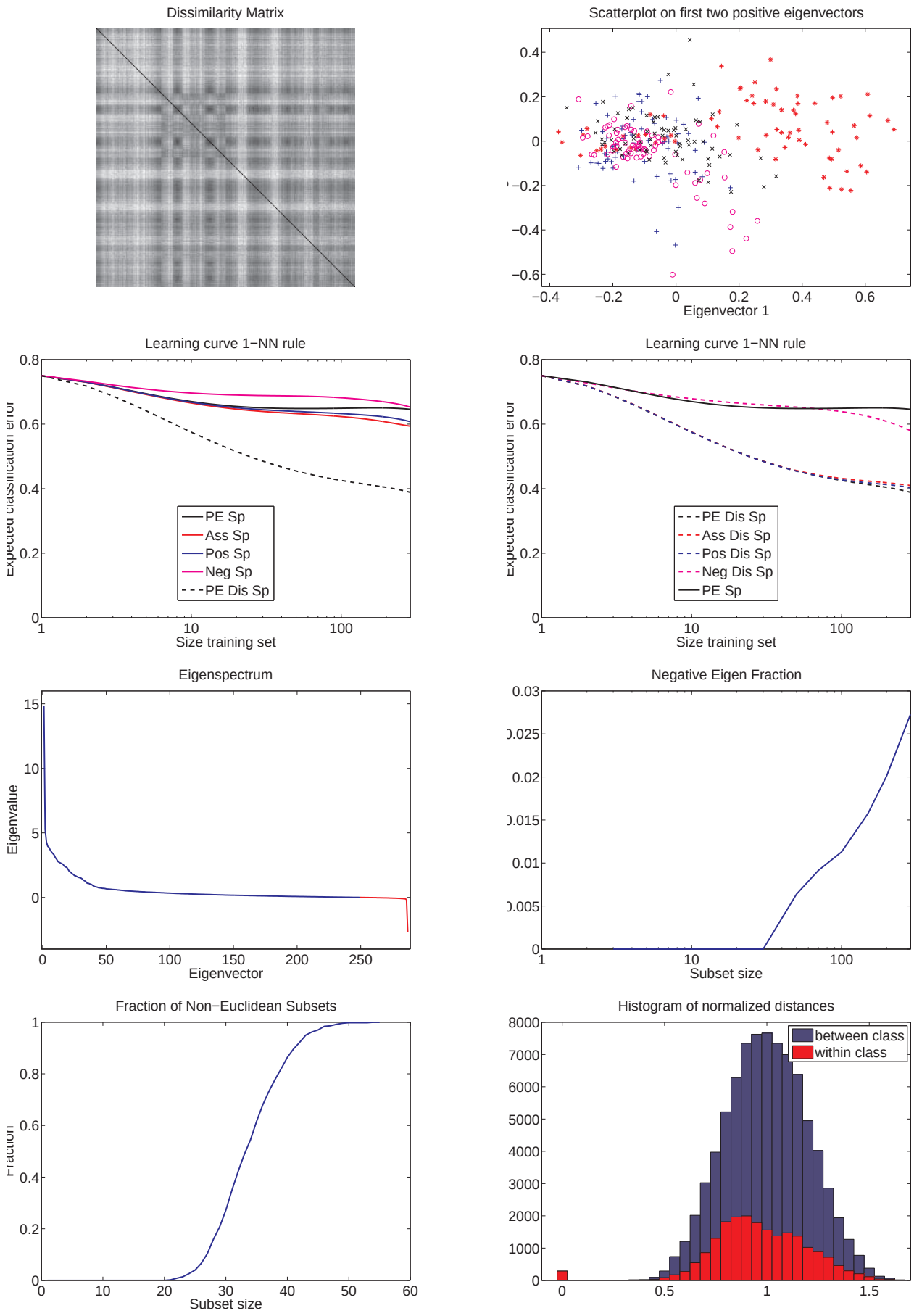


Figure 5: Graphical results for CoilDelftSame.

2.6 CoilYork

Description Dissimilarity Dataset

This is a dissimilarity matrix between set of graphs derived from four objects of the COIL database computed by Richard Wilson. The graphs are the Delaunay triangulations derived from corner points found in these images, see [Xia and Hancock]. The distance matrix is constructed by graph matching, using the algorithm of [Gold and Ranguranjan]

Reference(s)

B. Xiao and E. R. Hancock. Geometric characterisation of graphs. ICIAP 2005, LNCS 3617:471-478, 2005

S. Gold and A. Rangarajan. A graduated assignment algorithm for graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18:377-388, 1996.

Web page(s)

The PRTools version of the data: http://prlab.tudelft.nl/data/disdatasets.html

0.009	asymmetry
288	number of objects
204	number of significant eigenvectors
1	number of triangle inequality violations out of 23639616
169, 118	number of positive and negative eigenvalues
0.258	negative eigenfraction
0.046	negative eigenratio
4	number of classes, class sizes [72 72 72 72]
0.973, 1.009	average within-class and between-class dissimilarity
23.3, 33.7, 30.9	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 9: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	35.9 (1.2)	48.9 (1.1)	39.1 (0.9)	75.0 (1.4)	35.6 (0.9)
Parzen	53.4 (1.3)	53.0 (1.5)	53.2 (1.5)	83.9 (0.7)	53.4 (1.3)
NM	33.6 (1.2)	47.3 (0.9)	37.5 (1.3)	82.6 (1.1)	33.6 (1.3)
SVM-1	52.9 (1.3)	42.3 (1.5)	42.0 (1.7)	66.4 (0.8)	44.5 (1.4)

Table 10: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	39.7 (1.1)	47.3 (1.0)	41.7 (1.0)	74.2 (1.2)	36.5 (1.1)
Parzen	62.5 (1.1)	64.1 (1.1)	63.1 (1.1)	75.7 (1.3)	60.9 (1.0)
NM	37.2 (1.3)	44.0 (1.0)	39.1 (1.2)	74.0 (1.2)	34.5 (1.4)
SVM-1	36.5 (1.8)	38.0 (1.3)	36.8 (1.8)	67.2 (0.8)	40.6 (1.8)

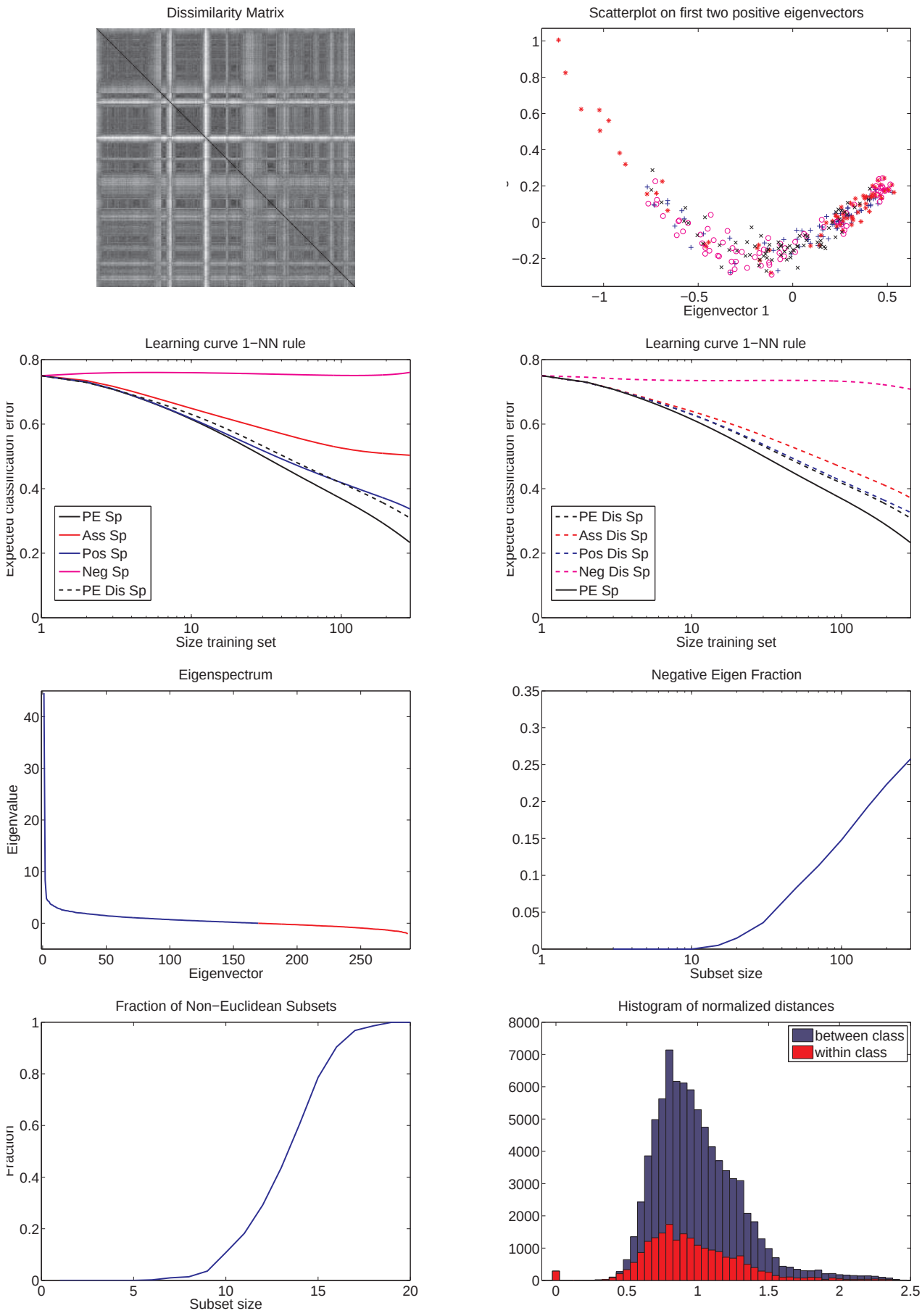


Figure 6: Graphical results for CoilYork.

2.7 DelftGestures

Description Dissimilarity Dataset

This dataset consists of the dissimilarities computed from a set of gestures in a sign-language study. They are measured by two video cameras observing the positions the two hands in 75 repititions of creating 20 different signs. The dissimilarities result from a dynamic time warping procedure. The experiments are performed by Gineke ten Holt, Jeroen Arendsen Robbert Eggermont and Jeroen Lichtenauer who also prepared the dataset.

Reference(s)

Jeroen Lichtenauer, Emile A. Hendriks, Marcel J. T. Reinders: Sign Language Recognition by Combining Statistical DTW and Independent Classification. IEEE Trans. Pattern Anal. Mach. Intell. vol. 30, 2040-2046, 2008.

Web page(s)

The PRTools version http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
1500	number of objects
943	number of significant eigenvectors
14798	number of triangle inequality violations out of 3368253000
765, 734	number of positive and negative eigenvalues
0.308	negative eigenfraction
0.035	negative eigenratio
20	number of classes, class sizes [75 75 75 75 75 75 75 75 75 75 75 75 75 75 75 75 75 75 75 75]
0.463, 1.028	average within-class and between-class dissimilarity
4.1, 10.4, 7.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 11: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	5.8 (0.2)	19.8 (0.4)	12.8 (0.3)	73.7 (0.3)	5.8 (0.2)
Parzen	4.8 (0.1)	6.8 (0.2)	5.4 (0.2)	68.3 (0.5)	4.8 (0.1)
NM	5.5 (0.2)	19.3 (0.4)	12.3 (0.4)	84.0 (1.2)	9.2 (0.3)
SVM-1	4.6 (0.2)	3.7 (0.1)	2.9 (0.1)	63.8 (0.4)	5.2 (0.3)

Table 12: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	10.7 (0.3)	19.1 (0.5)	15.7 (0.3)	57.6 (0.6)	8.3 (0.3)
Parzen	95.0 (0.0)	95.0 (0.0)	95.0 (0.0)	95.0 (0.0)	95.0 (0.0)
NM	95.0 (0.0)	95.0 (0.0)	95.0 (0.0)	95.0 (0.0)	95.0 (0.0)
SVM-1	4.4 (0.2)	5.6 (0.2)	4.8 (0.2)	51.7 (0.4)	10.1 (0.3)

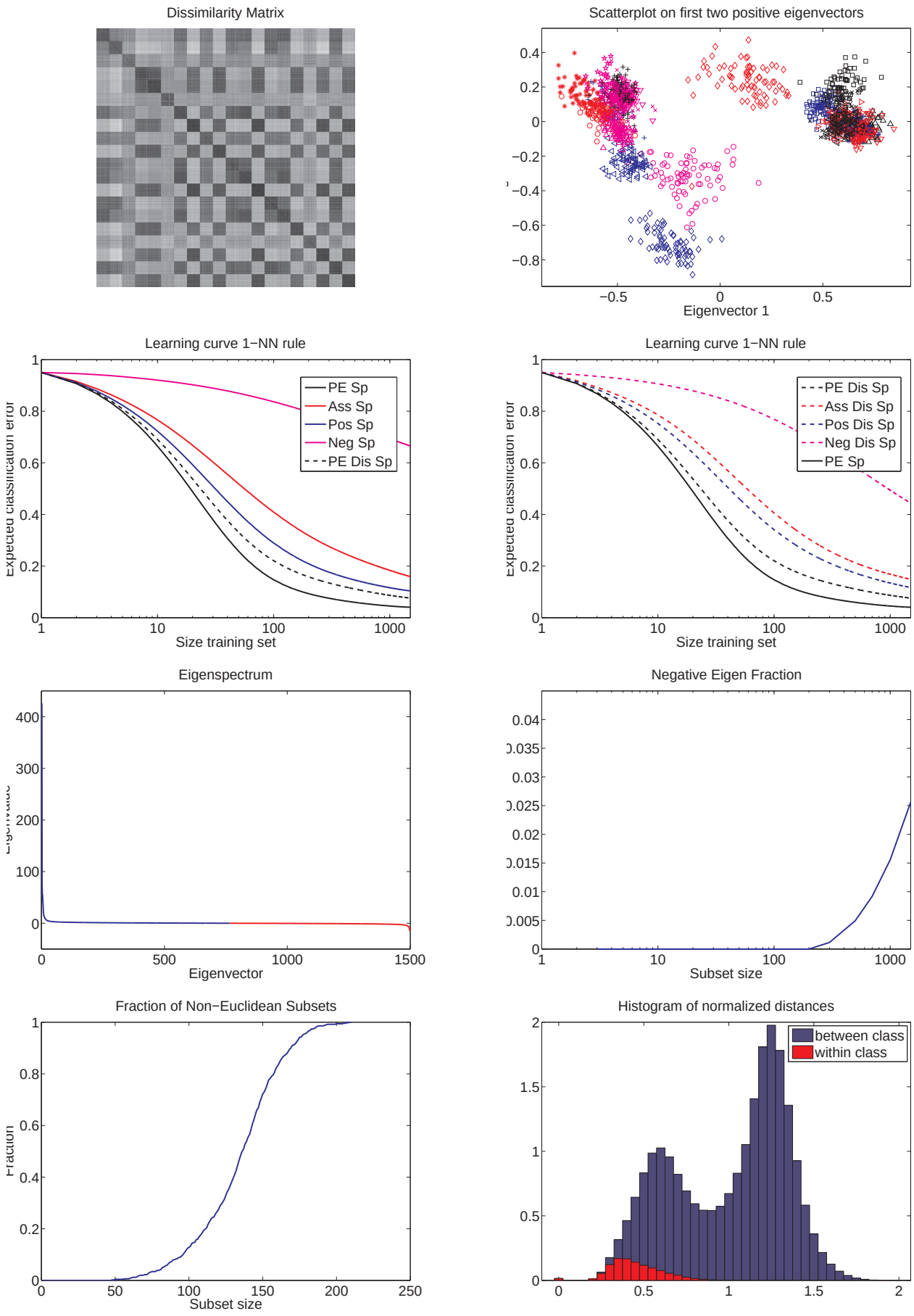


Figure 7: Graphical results for DelftGestures.

2.8 FlowCyto-1

Description Dissimilarity Dataset

This dissimilarity dataset is based on 612 FL3-A DNA flowcytometer histograms from breast cancer tissues in 256 resolution. The initial data were acquired by M. Nap and N. van Rodijnen of the Atrium Medical Center in Heerlen, The Netherlands, during 2000-2004, using tube 3 of a DACO Galaxy flowcytometer. Histograms are labeled in 3 classes: aneuploid (335 patients), diploid (131) and tetraploid (146). Dissimilarities between normalized histograms are computed using the L1 norm, correcting for possible different calibration factors.

Reference(s)

Web page(s)

The original histograms http://prlab.tudelft.nl/data/flowcytohist.html

The PRTools version of the dissimilarities http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
612	number of objects
228	number of significant eigenvectors
272052	number of triangle inequality violations out of 228098520
297, 314	number of positive and negative eigenvalues
0.275	negative eigenfraction
0.201	negative eigenratio
3	number of classes, class sizes [335 131 146]
1.004, 0.997	average within-class and between-class dissimilarity
37.6, 40.5, 37.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 13: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	44.3 (0.8)	46.6 (1.1)	47.1 (1.0)	61.5 (1.8)	44.3 (0.8)
Parzen	39.8 (1.1)	39.8 (1.1)	40.0 (1.1)	66.1 (1.5)	39.8 (1.1)
NM	42.8 (1.1)	42.0 (1.0)	41.5 (1.4)	48.0 (1.1)	44.5 (0.8)
SVM-1	43.9 (1.9)	38.6 (1.4)	39.3 (1.2)	65.9 (1.0)	37.4 (0.9)

Table 14: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	40.3 (1.4)	42.9 (1.7)	42.3 (2.0)	53.5 (1.6)	41.5 (1.2)
Parzen	43.8 (0.7)	46.0 (0.8)	45.3 (1.0)	57.8 (2.1)	43.7 (0.8)
NM	39.9 (1.1)	39.6 (1.7)	40.9 (1.7)	49.0 (1.4)	36.9 (0.8)
SVM-1	38.8 (0.7)	38.3 (1.0)	37.9 (0.7)	59.7 (1.3)	37.4 (0.9)

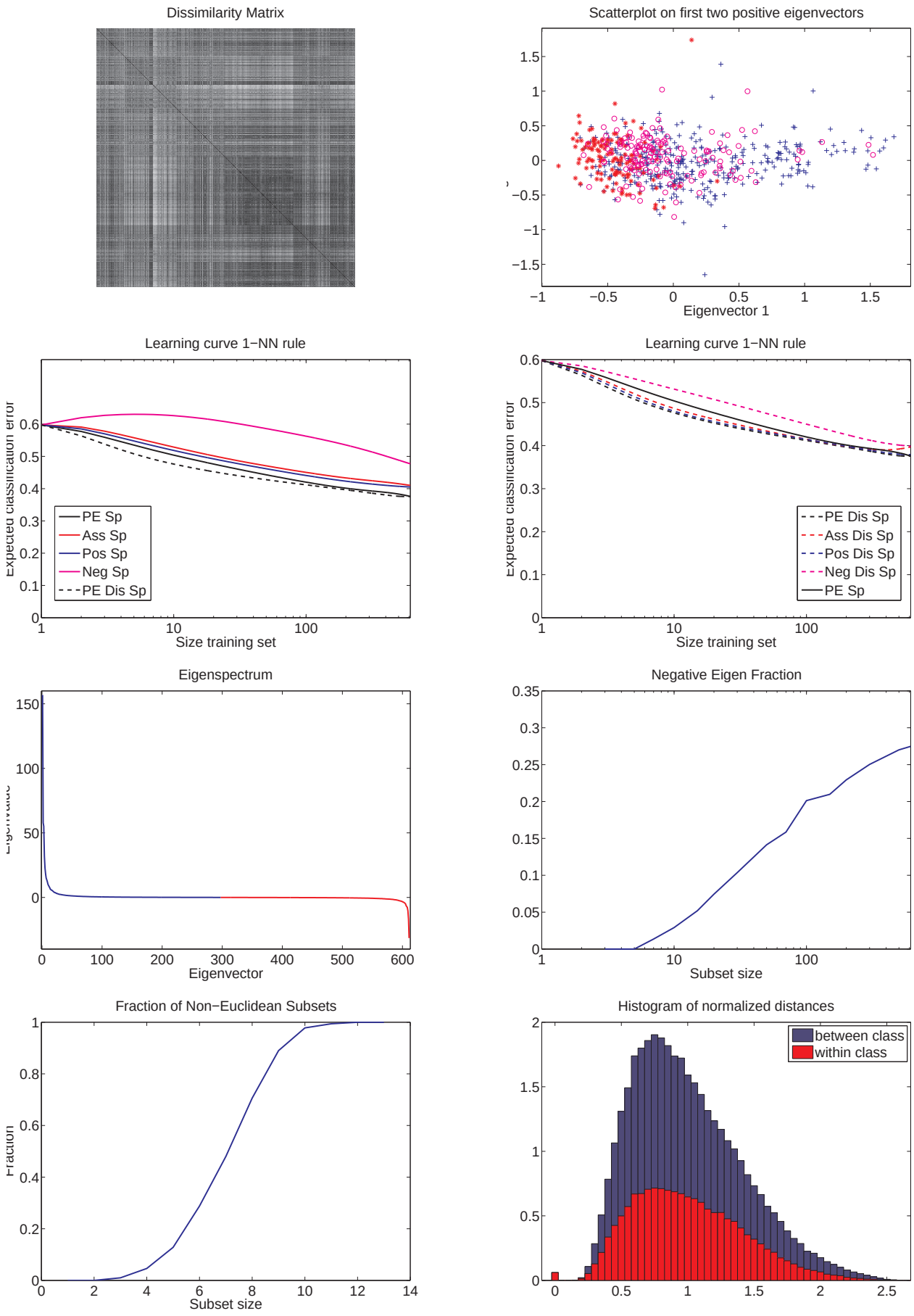


Figure 8: Graphical results for FlowCyto-1.

2.9 FlowCyto-2

Description Dissimilarity Dataset

This dissimilarity dataset is based on 612 FL3-A DNA flowcytometer histograms from breast cancer tissues in 256 resolution. The initial data were acquired by M. Nap and N. van Rodijnen of the Atrium Medical Center in Heerlen, The Netherlands, during 2000-2004, using tube 4 of a DACO Galaxy flowcytometer. Histograms are labeled in 3 classes: aneuploid (335 patients), diploid (131) and tetraploid (146). Dissimilarities between normalized histograms are computed using the L1 norm, correcting for possible different calibration factors.

Reference(s)

Web page(s)

The original histograms http://prlab.tudelft.nl/data/flowcytohist.html

The PRTools version of the dissimilarities http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
612	number of objects
231	number of significant eigenvectors
161517	number of triangle inequality violations out of 228098520
296, 315	number of positive and negative eigenvalues
0.268	negative eigenfraction
0.146	negative eigenratio
3	number of classes, class sizes [335 131 146]
1.007, 0.996	average within-class and between-class dissimilarity
35.8, 37.3, 37.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 15: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	45.5 (0.8)	51.0 (1.0)	49.0 (1.1)	65.4 (1.2)	45.5 (0.8)
Parzen	39.9 (0.8)	39.5 (0.8)	39.8 (0.8)	66.1 (1.5)	39.9 (0.8)
NM	44.9 (0.8)	43.6 (1.2)	42.5 (0.9)	49.4 (1.4)	45.3 (0.0)
SVM-1	50.3 (1.3)	39.3 (1.0)	40.3 (0.7)	64.0 (1.6)	39.1 (1.0)

Table 16: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	44.5 (1.2)	46.0 (1.7)	45.8 (1.4)	59.9 (0.8)	45.5 (1.3)
Parzen	43.3 (1.1)	44.0 (1.3)	43.8 (1.6)	55.7 (2.1)	44.2 (0.9)
NM	43.6 (1.2)	44.3 (1.9)	44.4 (1.7)	57.1 (1.7)	41.1 (1.9)
SVM-1	40.1 (0.8)	39.4 (0.9)	38.9 (0.8)	62.8 (2.1)	39.1 (1.4)

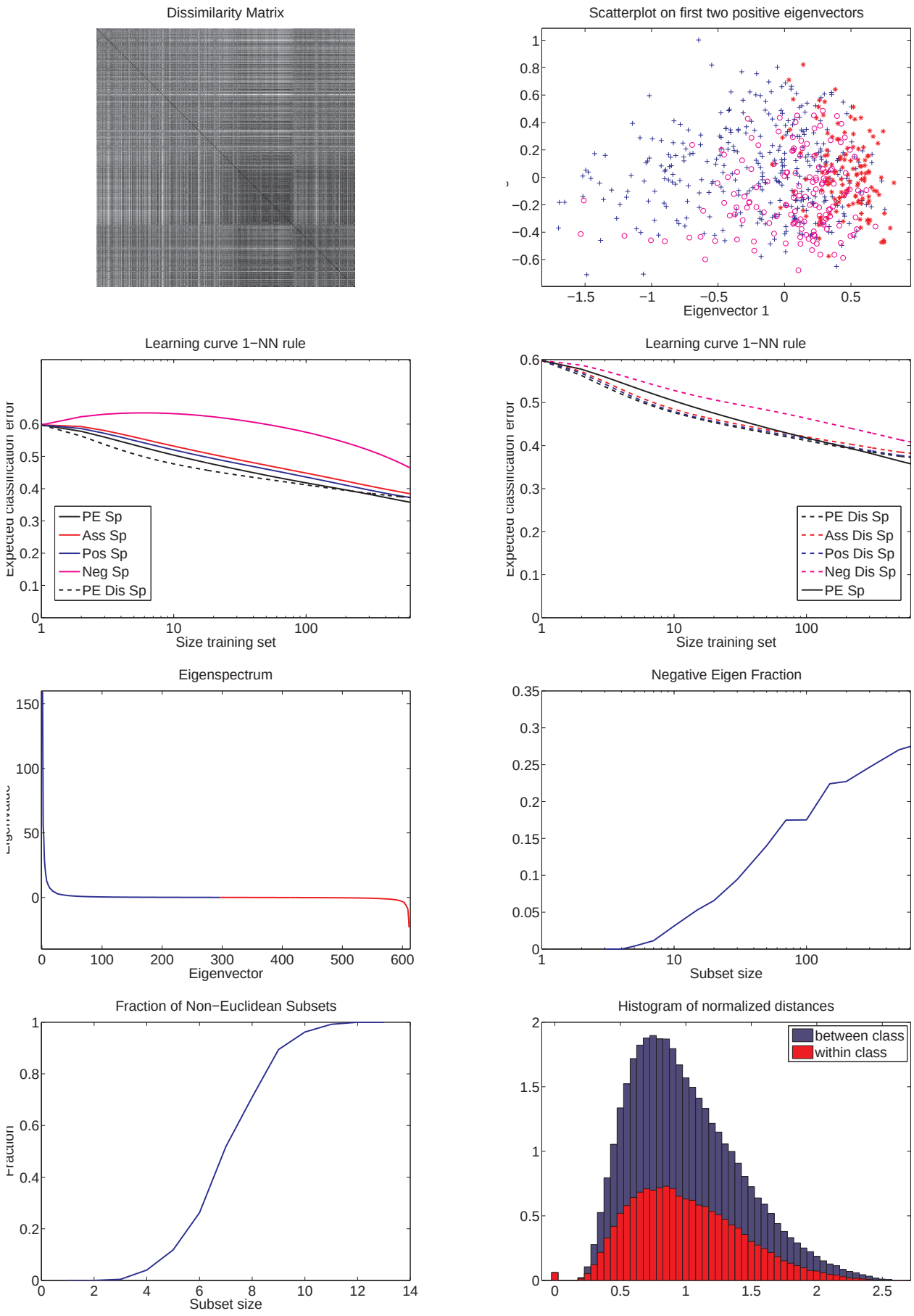


Figure 9: Graphical results for FlowCyto-2.

2.10 FlowCyto-3

Description Dissimilarity Dataset

This dissimilarity dataset is based on 612 FL3-A DNA flowcytometer histograms from breast cancer tissues in 256 resolution. The initial data were acquired by M. Nap and N. van Rodijnen of the Atrium Medical Center in Heerlen, The Netherlands, during 2000-2004, using tube 5 of a DACO Galaxy flowcytometer. Histograms are labeled in 3 classes: aneuploid (335 patients), diploid (131) and tetraploid (146). Dissimilarities between normalized histograms are computed using the L1 norm, correcting for possible different calibration factors.

Reference(s)

Web page(s)

The original histograms http://prlab.tudelft.nl/data/flowcytohist.html

The PRTools version of the dissimilarities http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
612	number of objects
226	number of significant eigenvectors
272879	number of triangle inequality violations out of 228098520
297, 314	number of positive and negative eigenvalues
0.275	negative eigenfraction
0.197	negative eigenratio
3	number of classes, class sizes [335 131 146]
1.006, 0.996	average within-class and between-class dissimilarity
39.1, 40.8, 40.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 17: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	47.7 (1.8)	51.1 (2.2)	49.3 (2.2)	65.7 (1.5)	47.7 (1.8)
Parzen	41.3 (1.2)	41.3 (1.1)	41.6 (1.2)	69.2 (2.2)	41.3 (1.2)
NM	44.2 (0.8)	44.0 (0.9)	43.9 (1.1)	48.9 (1.3)	45.3 (0.0)
SVM-1	47.5 (1.8)	41.8 (1.8)	40.6 (1.5)	64.0 (1.4)	40.3 (1.5)

Table 18: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	45.7 (1.7)	45.9 (1.5)	47.0 (1.8)	55.1 (1.6)	45.8 (1.7)
Parzen	43.0 (1.1)	45.5 (1.7)	44.9 (1.2)	60.1 (1.6)	42.4 (1.4)
NM	43.5 (1.7)	44.4 (1.5)	44.2 (1.5)	52.4 (1.5)	39.4 (1.5)
SVM-1	40.8 (1.7)	40.1 (1.8)	39.9 (1.9)	62.7 (2.3)	38.6 (1.4)

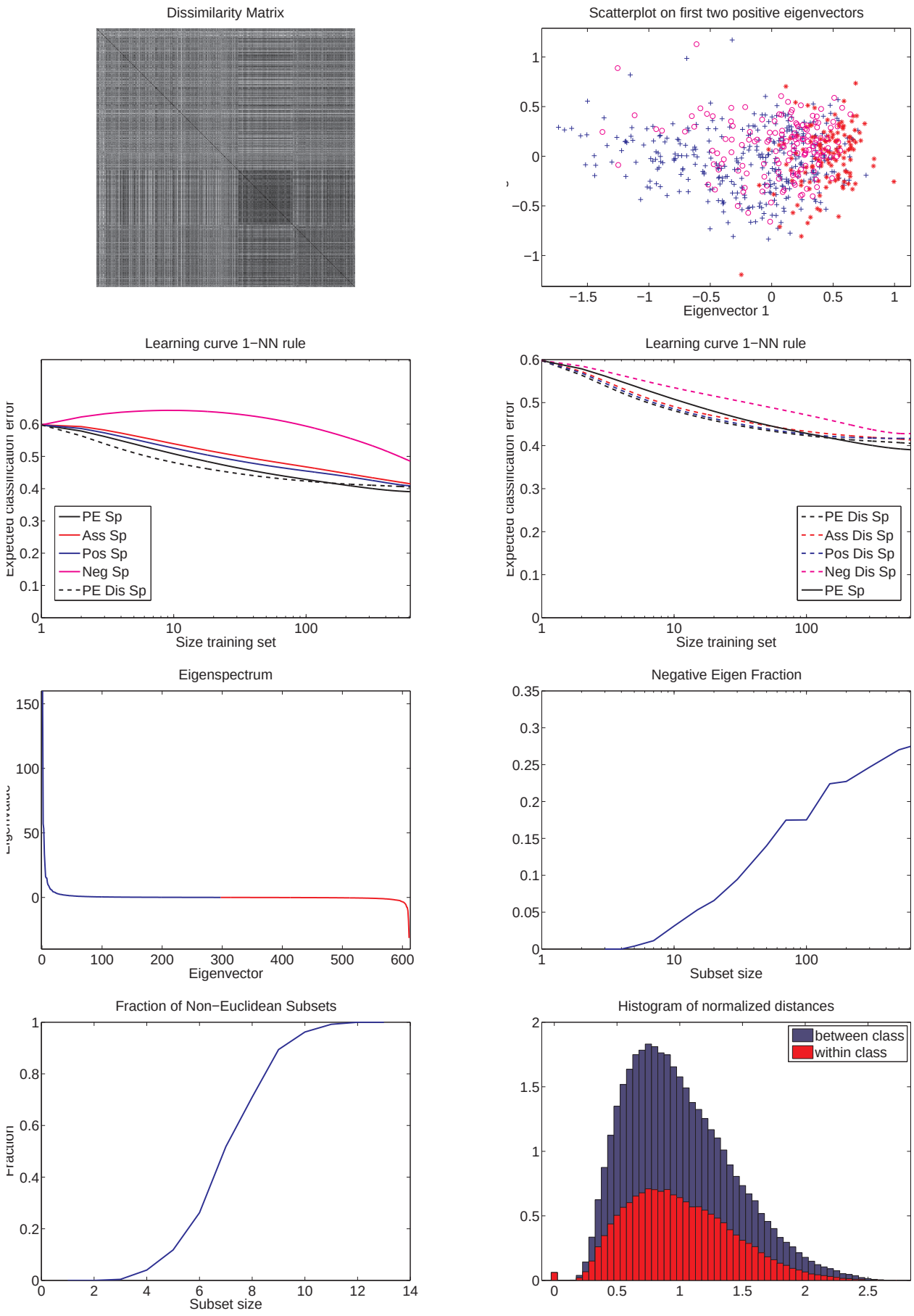


Figure 10: Graphical results for FlowCyto-3.

2.11 FlowCyto-4

Description Dissimilarity Dataset

This dissimilarity dataset is based on 612 FL3-A DNA flowcytometer histograms from breast cancer tissues in 256 resolution. The initial data were acquired by M. Nap and N. van Rodijnen of the Atrium Medical Center in Heerlen, The Netherlands, during 2000-2004, using tube 6 of a DACO Galaxy flowcytometer. Histograms are labeled in 3 classes: aneuploid (335 patients), diploid (131) and tetraploid (146). Dissimilarities between normalized histograms are computed using the L1 norm, correcting for possible different calibration factors.

Reference(s)

Web page(s)

The original histograms http://prlab.tudelft.nl/data/flowcytohist.html

The PRTools version of the dissimilarities http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
612	number of objects
227	number of significant eigenvectors
268991	number of triangle inequality violations out of 228098520
295, 316	number of positive and negative eigenvalues
0.272	negative eigenfraction
0.192	negative eigenratio
3	number of classes, class sizes [335 131 146]
1.008, 0.994	average within-class and between-class dissimilarity
41.8, 42.5, 42.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 19: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	49.2 (1.9)	50.2 (1.3)	49.8 (1.7)	63.4 (1.1)	49.2 (1.9)
Parzen	42.4 (1.1)	42.1 (1.3)	42.4 (1.2)	67.7 (1.7)	42.4 (1.1)
NM	42.7 (0.8)	43.2 (0.8)	42.5 (0.9)	47.4 (0.9)	45.3 (0.0)
SVM-1	47.1 (2.4)	41.4 (1.7)	41.6 (1.7)	67.2 (1.1)	39.7 (1.1)

Table 20: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	47.9 (1.6)	49.8 (2.0)	50.1 (1.8)	55.7 (1.4)	48.1 (1.6)
Parzen	45.2 (1.1)	47.4 (0.9)	46.8 (0.9)	62.9 (1.1)	46.3 (0.9)
NM	46.1 (1.5)	46.6 (1.8)	48.2 (1.4)	50.8 (1.1)	39.9 (1.1)
SVM-1	40.6 (1.4)	40.5 (1.5)	40.7 (1.5)	65.0 (1.3)	40.7 (1.4)

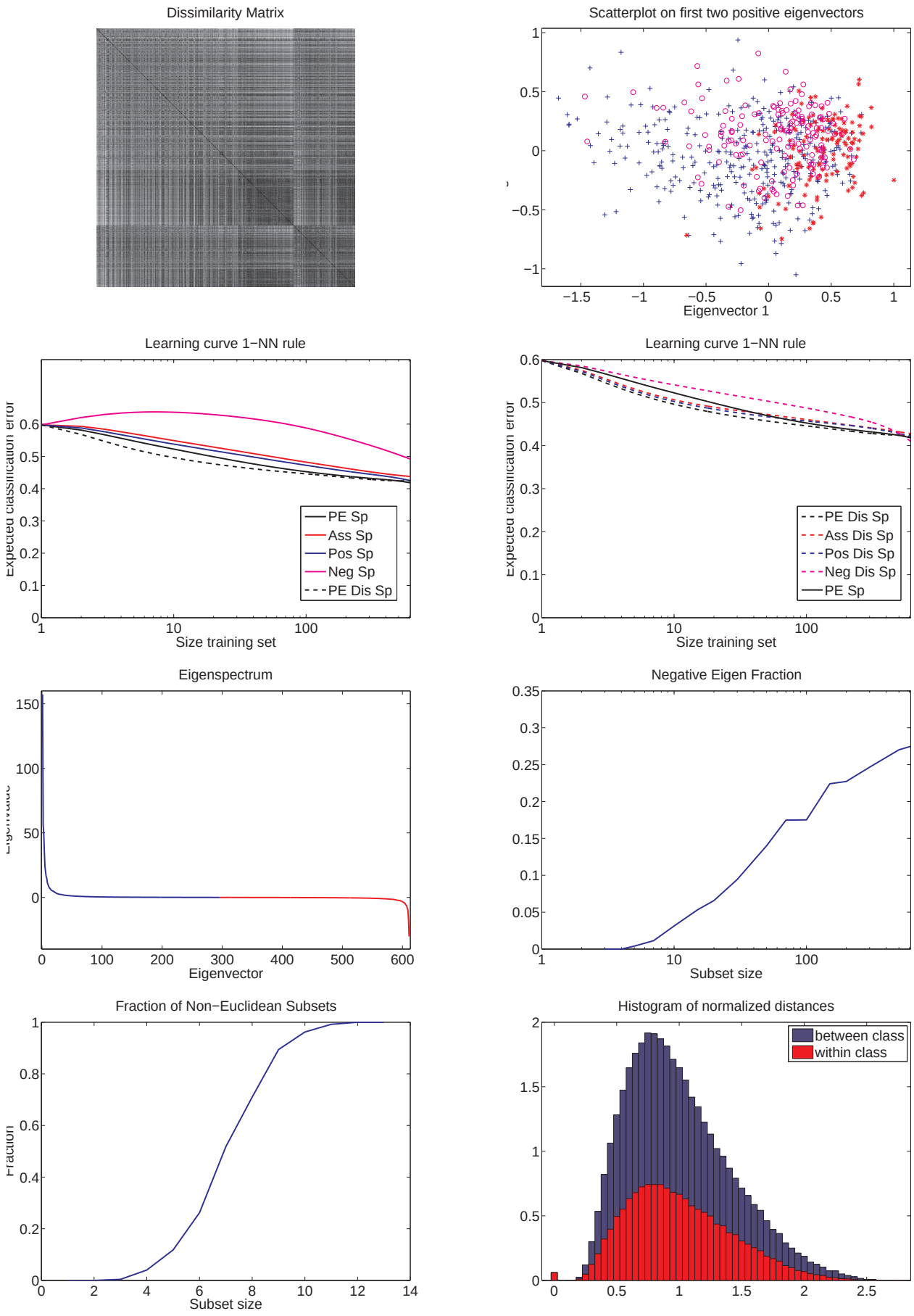


Figure 11: Graphical results for FlowCyto-4.

2.12 GaussM1

Description Dissimilarity Dataset

The dissimilarity dataset consists of the L1 distances between all points of two 20-dimensional Gaussian distributed sets of 1000 points each. Variances in all directions for both sets are 1. The means of the two sets are equal, except for the first dimension, where they have a distance 1.

Reference(s)

Web page(s)

The PRTools version http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
2000	number of objects
1074	number of significant eigenvectors
0	number of triangle inequality violations out of 7988004000
334, 1665	number of positive and negative eigenvalues
0.298	negative eigenfraction
0.036	negative eigenratio
2	number of classes, class sizes [1000 1000]
0.978, 1.022	average within-class and between-class dissimilarity
28.8, 30.4, 24.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 21: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	34.6 (2.0)	35.6 (1.9)	35.3 (1.9)	56.1 (1.5)	34.6 (2.0)
Parzen	20.1 (1.3)	20.1 (1.3)	19.8 (1.4)	63.6 (2.1)	20.1 (1.3)
NM	26.6 (1.9)	28.8 (1.6)	28.0 (1.8)	60.1 (2.0)	27.5 (1.9)
SVM-1	29.3 (1.9)	23.1 (1.6)	25.8 (1.3)	43.9 (2.4)	23.8 (1.8)

Table 22: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	32.9 (2.6)	32.4 (2.3)	32.7 (2.5)	56.1 (1.9)	32.8 (2.2)
Parzen	23.6 (1.6)	24.2 (1.8)	24.2 (1.8)	55.1 (1.3)	22.0 (1.6)
NM	32.2 (2.5)	31.9 (2.4)	31.6 (2.5)	57.0 (1.9)	31.8 (2.1)
SVM-1	26.3 (2.7)	22.8 (1.5)	24.6 (1.8)	40.2 (2.2)	22.6 (1.6)

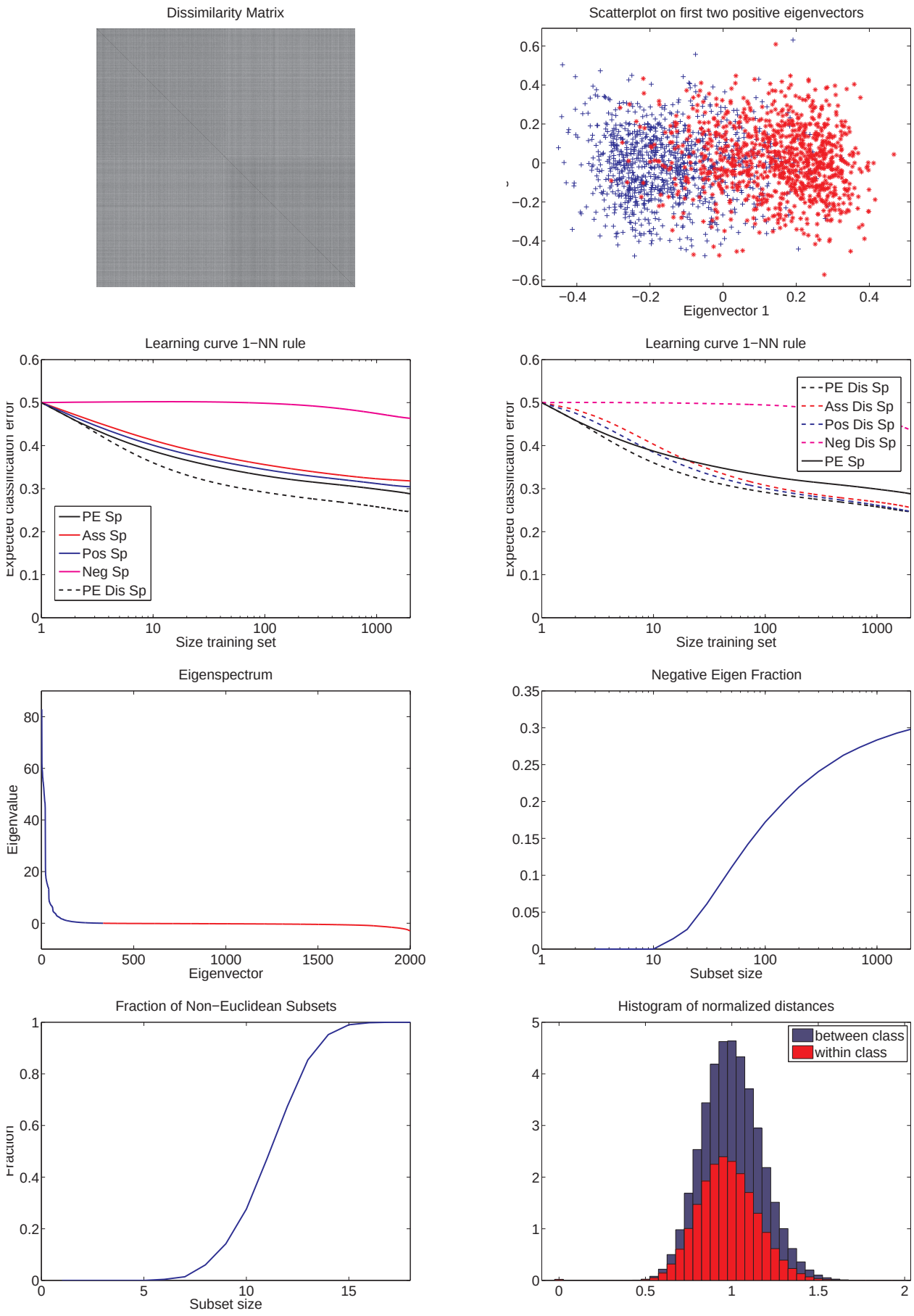


Figure 12: Graphical results for GaussM1.

2.13 GaussM02

Description Dissimilarity Dataset

The dissimilarity dataset consists of the Minkowsky 0.2 distances between all points of two 20-dimensional Gaussian distributed sets of 1000 points each. Variances in all directions for both sets are 1. The means of the two sets are equal, except for the first dimension, where they have a distance 1.

Reference(s)

Web page(s)

The PRTools version http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
2000	number of objects
1306	number of significant eigenvectors
2030263	number of triangle inequality violations out of 7988004000
623, 1376	number of positive and negative eigenvalues
0.426	negative eigenfraction
0.104	negative eigenratio
2	number of classes, class sizes [1000 1000]
0.981, 1.019	average within-class and between-class dissimilarity
34.0, 38.6, 30.1	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 23: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	40.3 (2.3)	41.4 (1.5)	39.9 (1.9)	58.6 (1.4)	40.3 (2.3)
Parzen	27.1 (1.8)	26.4 (1.5)	26.3 (1.8)	64.1 (2.7)	27.1 (1.8)
NM	33.4 (1.5)	38.0 (1.5)	34.5 (1.7)	60.6 (2.0)	34.3 (1.5)
SVM-1	34.9 (2.1)	30.2 (1.4)	32.2 (2.3)	40.7 (1.4)	29.4 (2.4)

Table 24: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	40.2 (2.0)	40.0 (2.2)	40.5 (2.2)	58.7 (1.9)	39.7 (1.9)
Parzen	29.0 (1.5)	34.9 (2.3)	31.7 (1.8)	57.5 (1.7)	27.4 (1.7)
NM	39.9 (1.9)	38.4 (2.1)	40.1 (2.2)	58.6 (2.1)	36.3 (1.9)
SVM-1	29.6 (2.2)	28.3 (1.4)	29.6 (2.4)	41.4 (2.0)	27.6 (1.7)

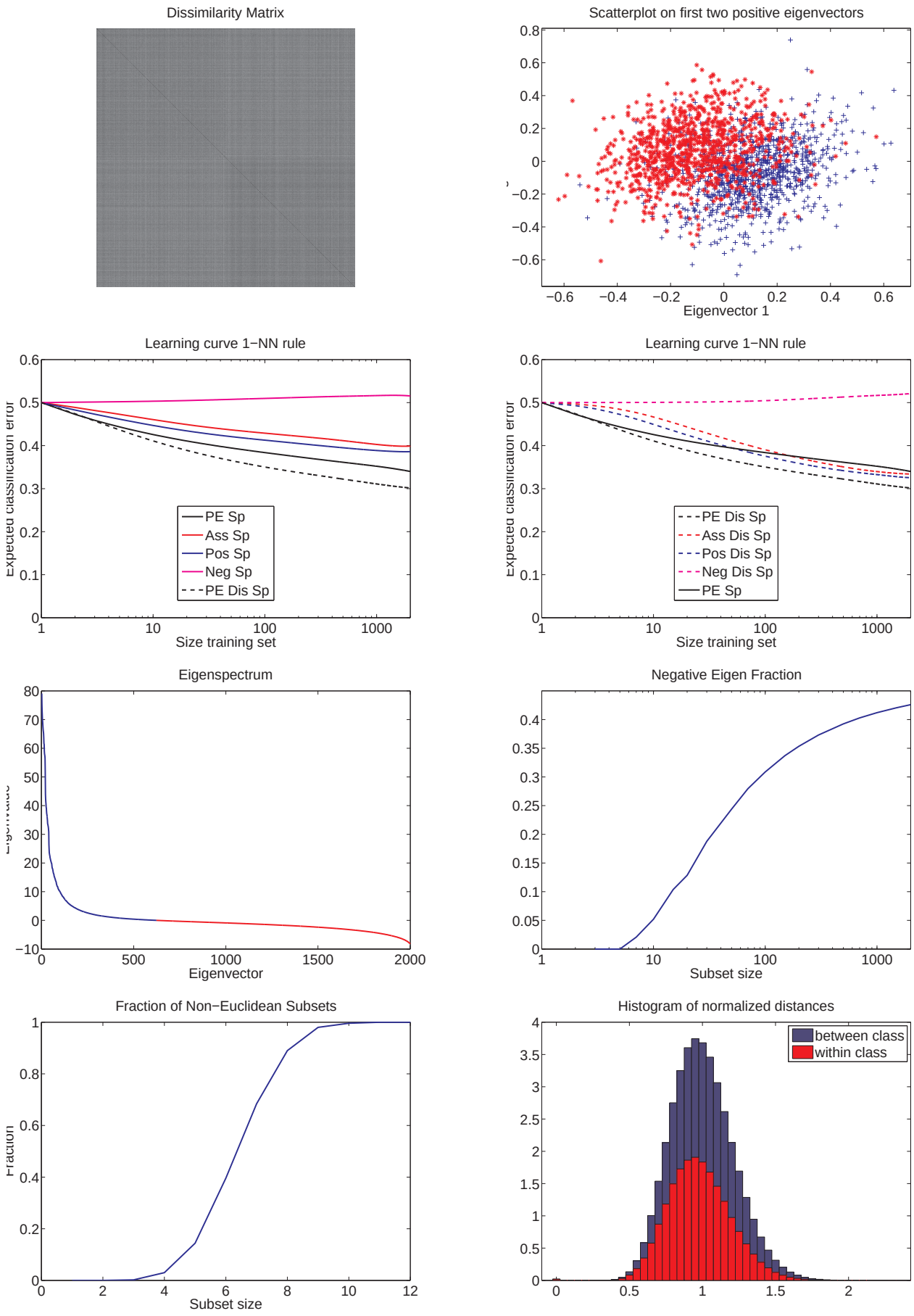


Figure 13: Graphical results for GaussM02.

2.14 NewsGroups

Description Dissimilarity Dataset

This is a small part of the so-called 20Newsgroups data, as considered by Roweis. A non-metric correlation measure for messages from four classes of newsgroups, 'comp.*', 'rec.*', 'sci.*' and 'talk.*' are computed on the occurrence for 100 words across 16242 postings.

Reference(s)

E. Pekalska and R.P.W. Duin, The Dissimilarity Representation for Pattern Recognition, Foundations and Applications, World Scientific, Singapore, 2005.

Web page(s)

The original data: http://www.cs.toronto.edu/?roweis/data.html

The PRTools dissimilarity matrix http://prlab.tudelft.nl/data/disdatasets.

0.000	asymmetry
600	number of objects
343	number of significant eigenvectors
4543	number of triangle inequality violations out of 214921200
153, 387	number of positive and negative eigenvalues
0.202	negative eigenfraction
0.049	negative eigenratio
4	number of classes, class sizes [170 125 102 203]
0.963, 1.013	average within-class and between-class dissimilarity
24.8, 26.7, 25.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 25: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	32.9 (1.2)	37.5 (1.1)	34.5 (1.2)	73.0 (1.2)	32.9 (1.2)
Parzen	28.2 (0.7)	28.5 (1.0)	28.5 (0.9)	83.6 (0.9)	28.2 (0.7)
NM	27.7 (1.1)	33.0 (0.9)	29.2 (0.9)	68.9 (0.9)	28.6 (1.1)
SVM-1	40.0 (0.8)	32.0 (1.3)	34.1 (1.0)	54.8 (0.9)	31.8 (0.8)

Table 26: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	34.7 (1.3)	33.5 (1.2)	33.5 (1.3)	60.8 (1.8)	34.4 (1.1)
Parzen	29.4 (1.0)	31.6 (0.8)	27.8 (1.0)	71.4 (1.1)	29.1 (0.9)
NM	33.8 (1.2)	32.6 (1.3)	32.8 (1.2)	61.5 (1.6)	33.1 (1.2)
SVM-1	32.6 (1.1)	30.4 (1.2)	31.4 (0.9)	63.5 (1.5)	29.6 (1.0)

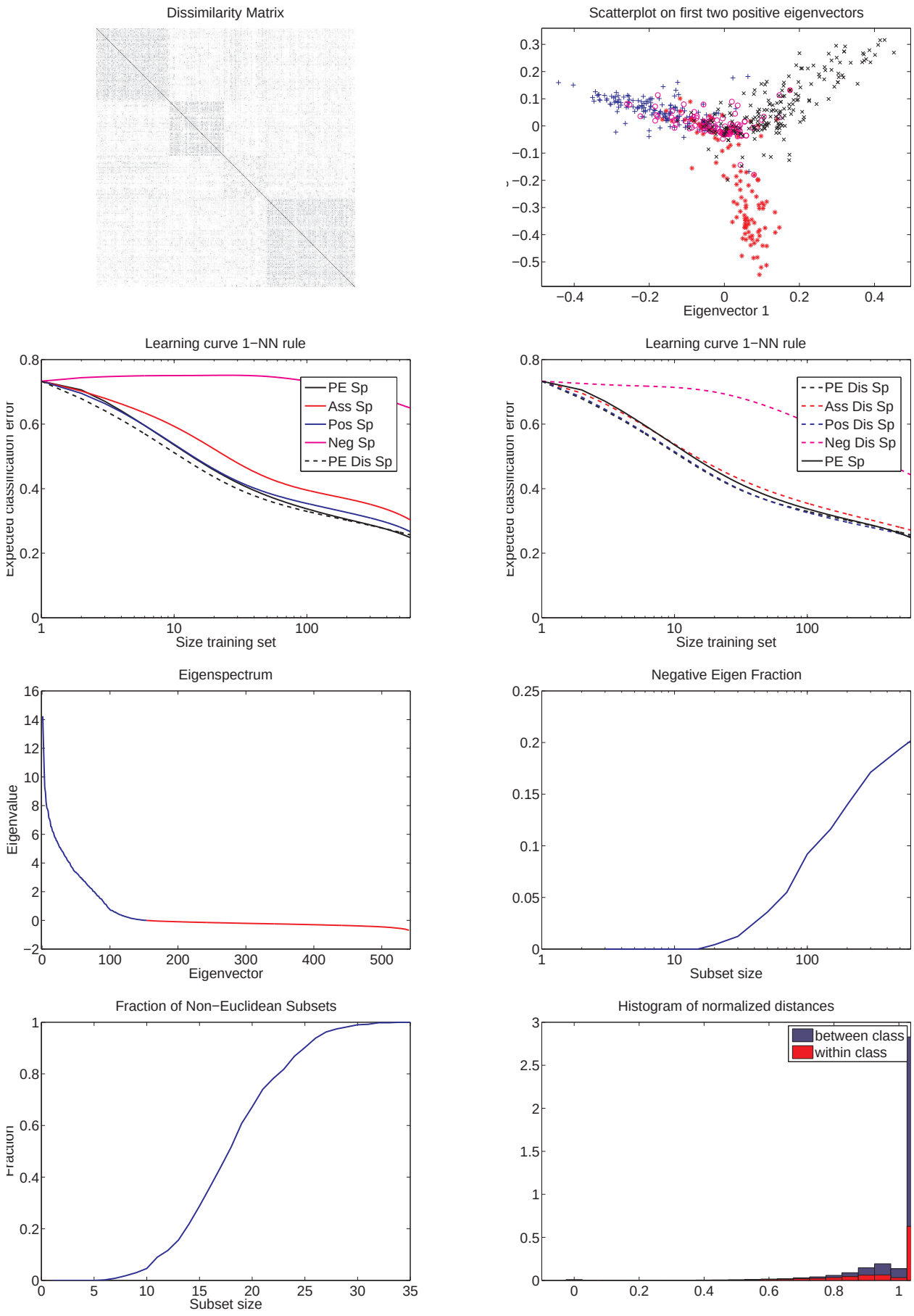


Figure 14: Graphical results for NewsGroups.

2.15 PolyDisH57

Description Dissimilarity Dataset

These are the Hausdorff distances between two randomly generated sets of polygons, pentagons and heptagons, both possibly non-convex. Means are made equal and scales are normalized before the distances are computed, but the polygons are not rotated.

Reference(s)

E. Pekalska and R.P.W. Duin, The Dissimilarity Representation for Pattern Recognition, Foundations and Applications, World Scientific, Singapore, 2005.

Web page(s)

The PRTools version of the data: http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
3000	number of objects
1842	number of significant eigenvectors
0	number of triangle inequality violations out of 26973006000
1546, 1453	number of positive and negative eigenvalues
0.409	negative eigenfraction
0.247	negative eigenratio
2	number of classes, class sizes [1500 1500]
0.968, 1.032	average within-class and between-class dissimilarity
3.4, 6.6, 2.4	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 27: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	16.8 (1.4)	22.8 (1.6)	19.1 (2.1)	48.4 (2.5)	16.8 (1.4)
Parzen	19.4 (1.6)	19.8 (1.4)	19.4 (1.4)	61.7 (1.5)	19.4 (1.6)
NM	14.8 (1.2)	20.5 (1.4)	17.2 (1.7)	60.4 (1.1)	15.1 (1.5)
SVM-1	26.7 (1.9)	12.2 (1.3)	12.6 (1.1)	39.9 (1.8)	16.6 (1.4)

Table 28: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	21.4 (2.2)	25.9 (2.2)	23.7 (2.0)	47.5 (2.2)	18.2 (1.6)
Parzen	28.6 (1.6)	30.9 (1.2)	28.6 (1.4)	58.1 (1.9)	23.2 (1.1)
NM	21.0 (2.2)	25.0 (1.9)	23.1 (2.0)	46.8 (2.2)	15.2 (1.8)
SVM-1	9.5 (1.2)	11.2 (1.3)	8.4 (1.1)	42.1 (1.8)	12.2 (1.2)

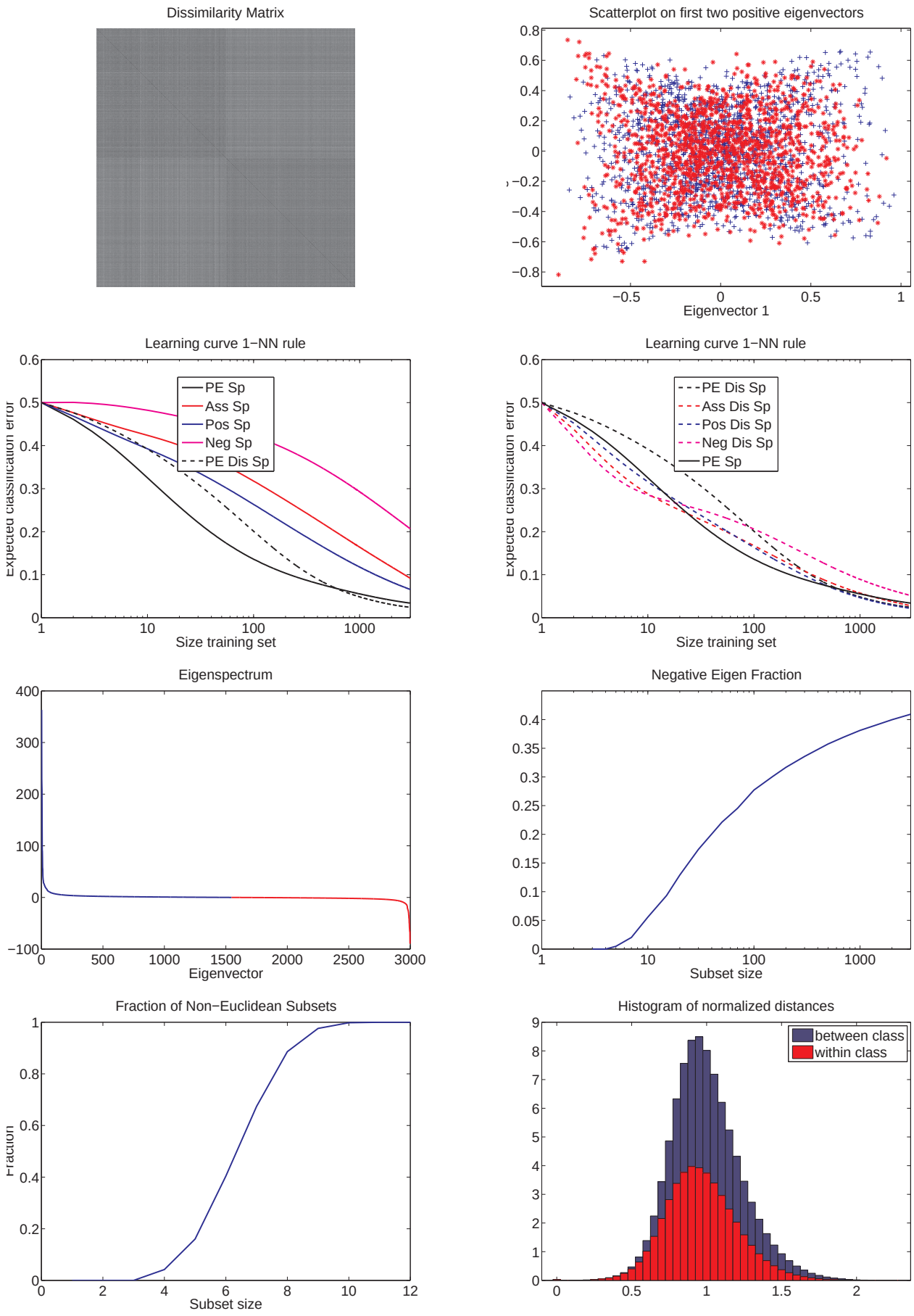


Figure 15: Graphical results for PolyDisH57.

2.16 PolyDisM57

Description Dissimilarity Dataset

These are the modified Hausdorff distances between two randomly generated sets of polygons, pentagons and heptagons, both possibly non-convex. Means are made equal and scales are normalized before the distances are computed, but the polygons are not rotated.

Reference(s)

M.-P. Dubuisson and A.K. Jain, Modified Hausdorff distance for object matching, Proceedings 12th IAPR International Conference on Pattern Recognition (Jerusalem, October 9-13, 1994), vol. 1, IEEE, Piscataway, NJ, USA,94CH3440-5, 1994, 566-568.

E. Pekalska and R.P.W. Duin, The Dissimilarity Representation for Pattern Recognition, Foundations and Applications, World Scientific, Singapore, 2005.

Web page(s)

The PRTools version of the data: http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
3000	number of objects
1673	number of significant eigenvectors
419440	number of triangle inequality violations out of 26973006000
1352, 1647	number of positive and negative eigenvalues
0.348	negative eigenfraction
0.109	negative eigenratio
2	number of classes, class sizes [1500 1500]
0.957, 1.043	average within-class and between-class dissimilarity
1.8, 2.3, 1.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 29: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	14.0 (1.3)	14.0 (1.0)	13.7 (1.4)	45.1 (2.3)	14.0 (1.3)
Parzen	7.3 (1.2)	7.8 (1.1)	7.9 (1.2)	69.2 (1.0)	7.3 (1.2)
NM	11.7 (1.1)	12.2 (1.3)	10.4 (1.4)	50.1 (0.2)	11.3 (1.2)
SVM-1	13.6 (1.1)	3.7 (0.7)	2.8 (0.6)	36.1 (1.6)	3.4 (0.9)

Table 30: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	11.4 (1.6)	12.7 (1.1)	11.5 (1.5)	38.1 (1.5)	12.4 (1.6)
Parzen	10.2 (1.2)	14.3 (1.1)	11.2 (1.2)	29.8 (1.4)	8.2 (1.5)
NM	11.2 (1.6)	12.0 (1.1)	11.1 (1.4)	37.9 (1.4)	10.5 (1.4)
SVM-1	3.1 (0.7)	3.1 (0.5)	2.7 (0.6)	43.8 (1.9)	2.6 (0.9)

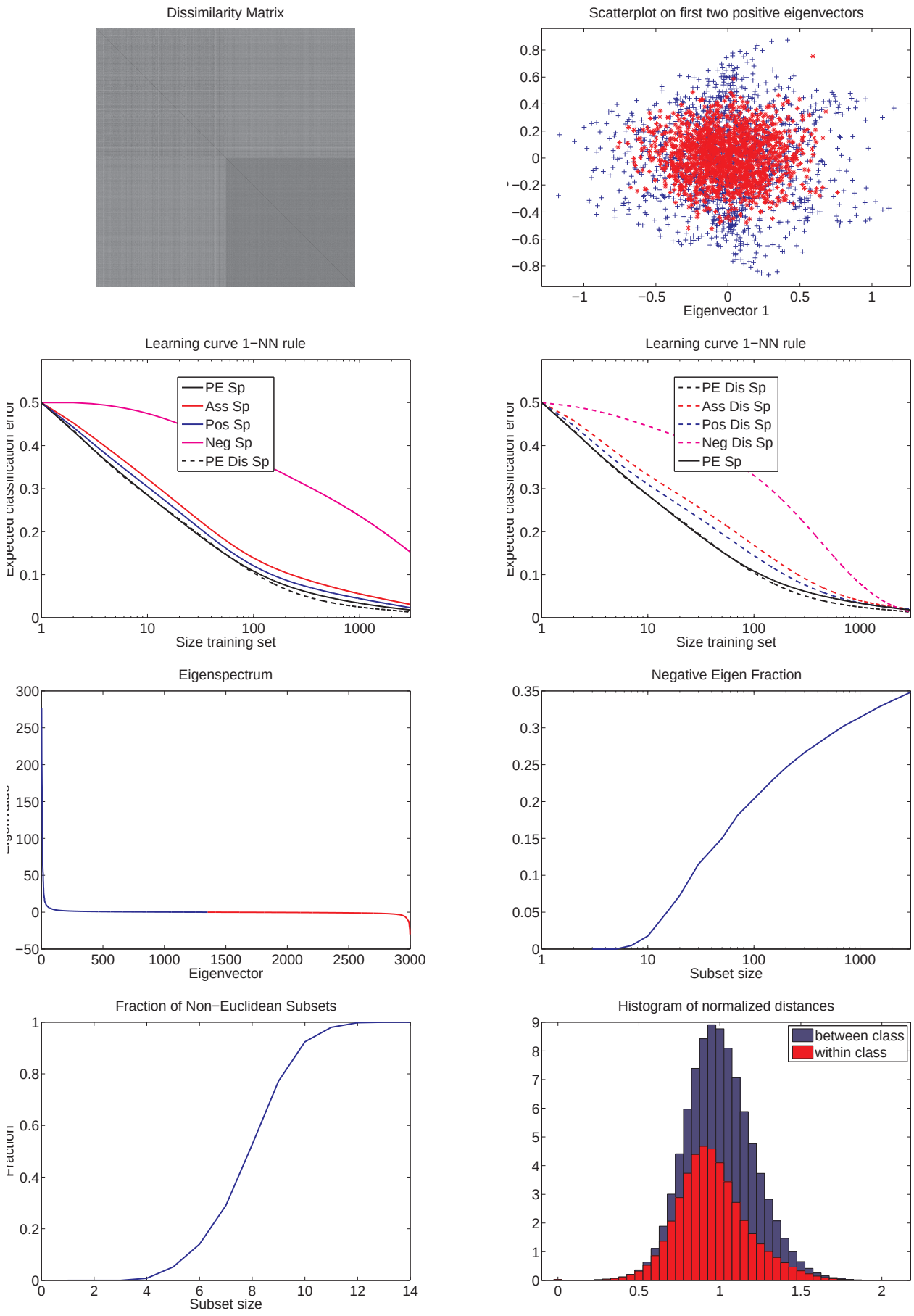


Figure 16: Graphical results for PolyDisM57.

2.17 ProDom

Description Similarity Dataset

ProDom is a comprehensive set of protein domain families [Corpet]. A ProDom subset of 2604 protein domain sequences from the ProDom set was selected by [Roth]. These are chosen based on a high similarity to at least one sequence contained in the first four folds of the SCOP database. The pairwise structural alignments are computed [Roth]. Each SCOP sequence belongs to a group, as labeled by the experts [Murzin]. The same four classes are assigned here.

Reference(s)

V. Roth, J. Laub, J.M. Buhmann, and K.-R. Mueller, Going metric: Denoising pairwise data, Advances in Neural Information Processing Systems, 841-856, MIT Press, 2003.

F. Corpet, F. Servant, J. Gouzy and D. Kahn, ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons, Nucleid Acids Res., vol. 28, 267-269, 2000.

A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, Journal of Molecular Biology, vol. 247, 536-540, 1995.

Web page(s)

The PRTools version http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
2604	number of objects
970	number of significant eigenvectors
136	number of triangle inequality violations out of 17636907624
1502, 680	number of positive and negative eigenvalues
0.043	negative eigenfraction
0.011	negative eigenratio
4	number of classes, class sizes [878 404 271 1051]
0.961, 1.018	average within-class and between-class dissimilarity
0.2, 0.2, 0.8	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 31: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	15.9 (0.8)	15.9 (0.8)	15.9 (0.8)	67.6 (1.2)	15.9 (0.8)
Parzen	15.9 (1.0)	15.9 (1.0)	15.9 (1.0)	74.7 (1.2)	15.9 (1.0)
NM	15.9 (0.7)	15.9 (0.7)	15.9 (0.7)	74.2 (0.6)	15.7 (0.7)
SVM-1	7.0 (0.7)	7.0 (0.7)	7.0 (0.7)	73.7 (0.7)	7.0 (0.7)

Table 32: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	11.5 (0.6)	11.5 (0.6)	11.5 (0.6)	67.3 (1.3)	11.5 (0.6)
Parzen	39.8 (1.7)	39.8 (1.7)	39.8 (1.7)	65.7 (0.9)	40.0 (1.7)
NM	11.3 (0.7)	11.3 (0.7)	11.3 (0.7)	63.7 (1.7)	11.1 (0.7)
SVM-1	8.0 (0.7)	7.9 (0.7)	7.9 (0.7)	69.6 (1.7)	7.9 (0.7)

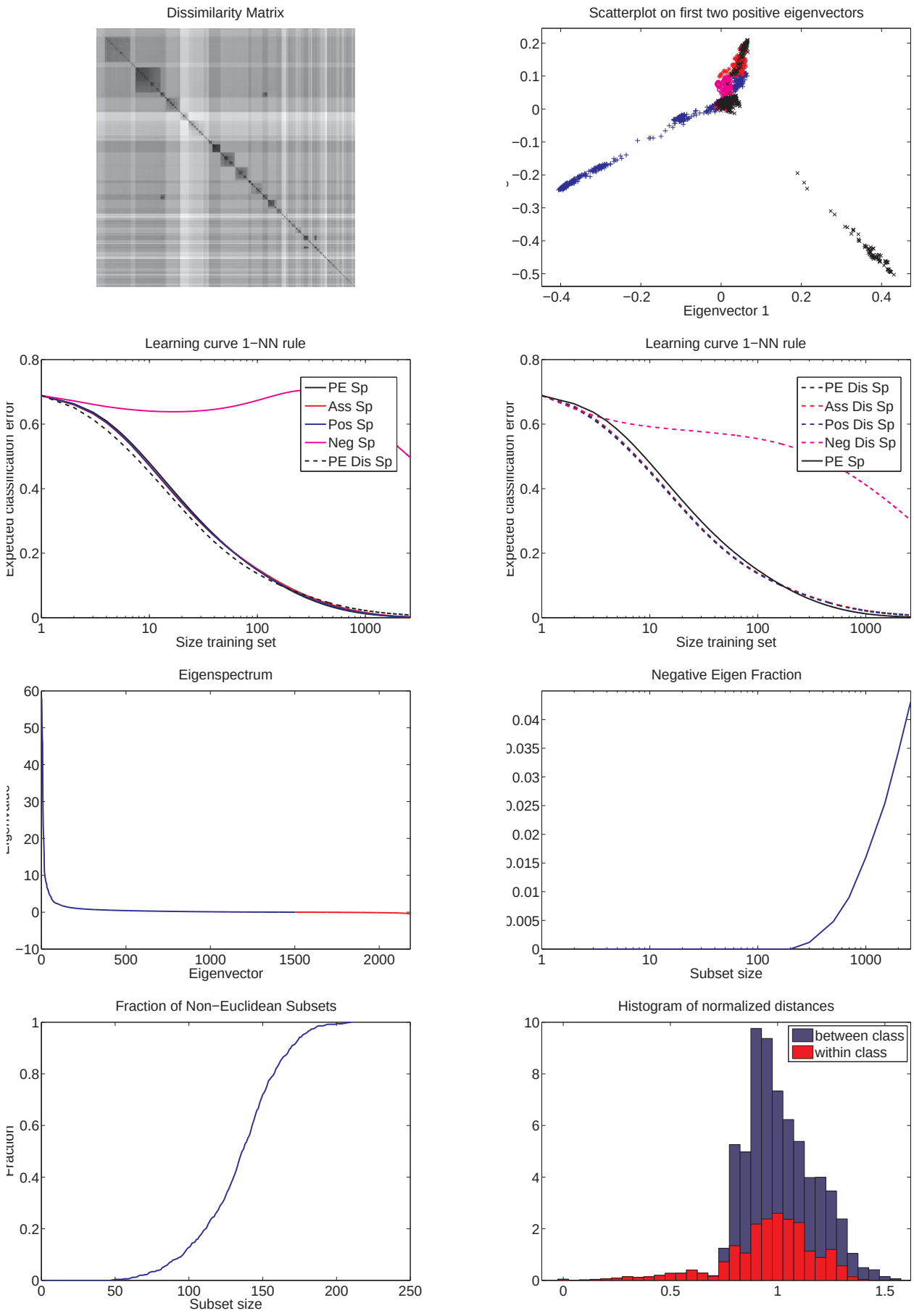


Figure 17: Graphical results for ProDom.

2.18 Protein

Description Dissimilarity Dataset

The protein data are provided as a 213x213 dissimilarity matrix comparing the protein sequences based on the concept of an evolutionary distance. It was used for classication in [Graepel] and for clustering in [Denoeux and Masson]. There are four classes of globins: heterogeneous globin (G), hemoglobin-A (HA), hemoglobin-B (HB) and myoglobin (M).

Reference(s)

T. Graepel, R. Herbrich, P. Bollmann-Sdorra, K. Obermayer, Classification on pairwise proximity data. In Advances in Neural Information System Processing vol. 11, 438-444, 1999.

T. Denoeux, T. and M.-H. Masson, EVCLUS: Evidential clustering of proximity data. IEEE Transations on Systems, Man and Cybernetics, vol. 34, 95-109, 2004.

Web page(s)

The original data http://www.hds.utc.fr/ tdenoeux/software.htm

The PRTools version http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
213	number of objects
86	number of significant eigenvectors
0	number of triangle inequality violations out of 9527916
205, 4	number of positive and negative eigenvalues
0.001	negative eigenfraction
0.002	negative eigenratio
4	number of classes, class sizes [30 72 72 39]
0.715, 1.110	average within-class and between-class dissimilarity
1.9, 1.9, 0.0	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

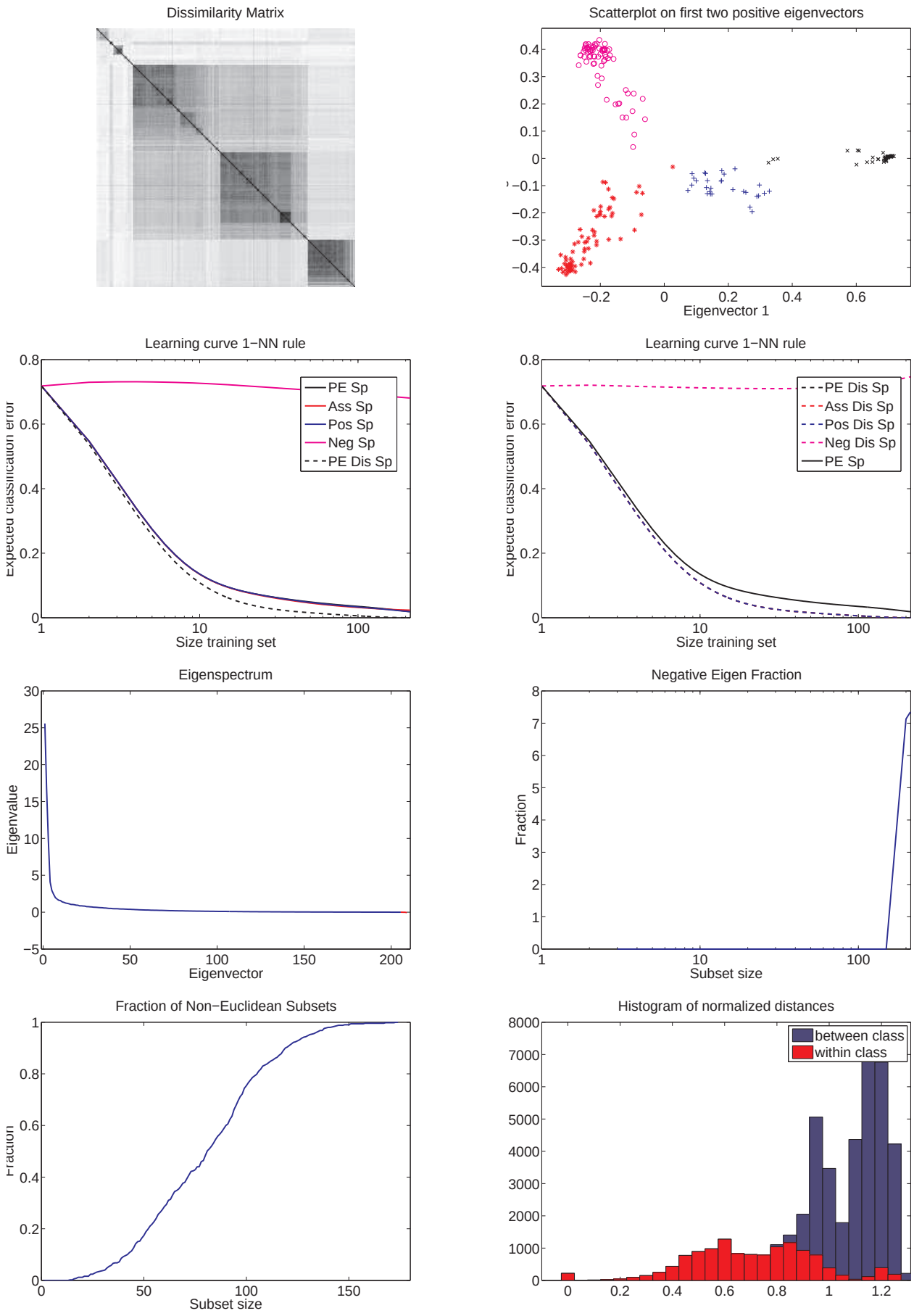


Figure 18: Graphical results for Protein.

2.19 WoodyPlants50

Description Dissimilarity Dataset

This dataset of shape dissimilarities between leaves is a small part of the data that is collected in a study on woody plants. This particular subset has been donated by David Jacobs of University College Maryland and consists out of examples of 14 species for which more than 50 leaves per class are avialable.

Reference(s)

Agarwal, G., Belhumeur, P., Feiner, S., Jacobs, D., Kress, W.J., Ramamoorthi, R., Bourg, N., Dixit, N., Ling, H., Mahajan, D.,Russell, R., Shirdhonkar, S., Sunkavalli, K., and White, S., First Steps Toward an Electronic Field Guide for Plants, Taxon, Journal of the International Association for Plant Taxonomy, vol. 55, August 2006, 597610.

Ling, H., and Jacobs, D.W., Shape Classification Using the Inner-Distance, IEEE Trans on Pattern Anal. and Mach. Intell. (PAMI), vol. 29, no. 2, 286-299, 2007.

Web page(s)

The project website http://herbarium.cs.columbia.edu/About.html

The PRTools version of this data http://prlab.tudelft.nl/data/disdatasets.html

0.000	asymmetry
791	number of objects
441	number of significant eigenvectors
115253	number of triangle inequality violations out of 493038210
395, 395	number of positive and negative eigenvalues
0.229	negative eigenfraction
0.056	negative eigenratio
14	number of classes, class sizes [63 52 61 54 62 53 51 51 55 52 53 66 55 63]
0.589, 1.031	average within-class and between-class dissimilarity
10.0, 10.5, 14.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

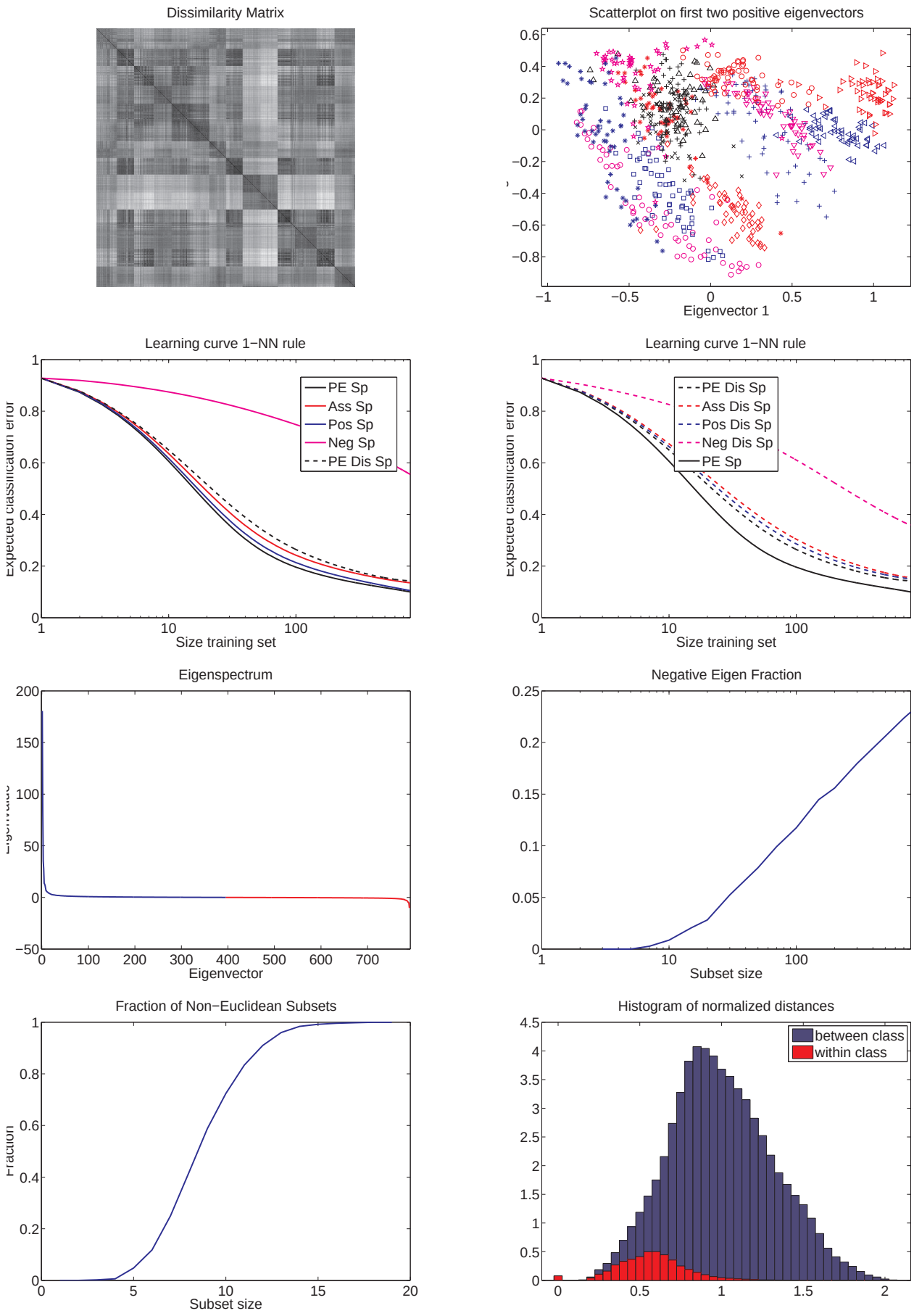


Figure 19: Graphical results for WoodyPlants50.

2.20 Zongker

Description Similarity Dataset

These similarities between 2000 handwritten digits in 10 classes are based on deformable template matching. The dissimilarity measure is the result of an iterative optimization of the non-linear deformation of the grid, see the study by Jain and Zongker. The data has been made available by them to Pekalska who used it in slightly modified version (symmetrized dissimilarities) in several studies.

Reference(s)

A.K. Jain and D. Zongker, Representation and recognition of handwritten digits using deformable templates, PAMI, vol. 19, no. 12, 1997, 1386-1391.

E. Pekalska and R.P.W. Duin, The Dissimilarity Representation for Pattern Recognition, Foundations and Applications, World Scientific, Singapore, 2005.

Web page(s)

The PRTools version http://prlab.tudelft.nl/data/disdatasets.html

0.051	asymmetry
2000	number of objects
1438	number of significant eigenvectors
6583656	number of triangle inequality violations out of 7988004000
1038, 961	number of positive and negative eigenvalues
0.419	negative eigenfraction
0.354	negative eigenratio
10	number of classes, class sizes [200 200 200 200 200 200 200 200 200 200]
0.790, 1.023	average within-class and between-class dissimilarity
44.0, 16.7, 3.9	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 33: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	29.6 (2.0)	31.3 (1.2)	13.8 (0.6)	88.8 (0.4)	29.6 (2.0)
Parzen	9.0 (0.3)	9.0 (0.2)	7.9 (0.3)	87.4 (0.3)	9.0 (0.3)
NM	22.0 (0.9)	32.5 (1.2)	13.5 (0.6)	90.0 (0.1)	17.7 (0.5)
SVM-1	24.1 (0.5)	6.0 (0.3)	5.4 (0.2)	76.1 (0.6)	8.0 (0.4)

Table 34: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	8.8 (0.3)	13.4 (0.6)	9.3 (0.6)	67.7 (1.0)	5.7 (0.4)
Parzen	31.7 (0.4)	46.5 (0.9)	34.9 (0.6)	77.8 (0.7)	11.4 (0.5)
NM	8.7 (0.3)	13.3 (0.5)	9.3 (0.6)	67.4 (1.0)	5.2 (0.4)
SVM-1	6.1 (0.4)	7.5 (0.4)	6.6 (0.3)	72.4 (0.6)	9.0 (0.4)

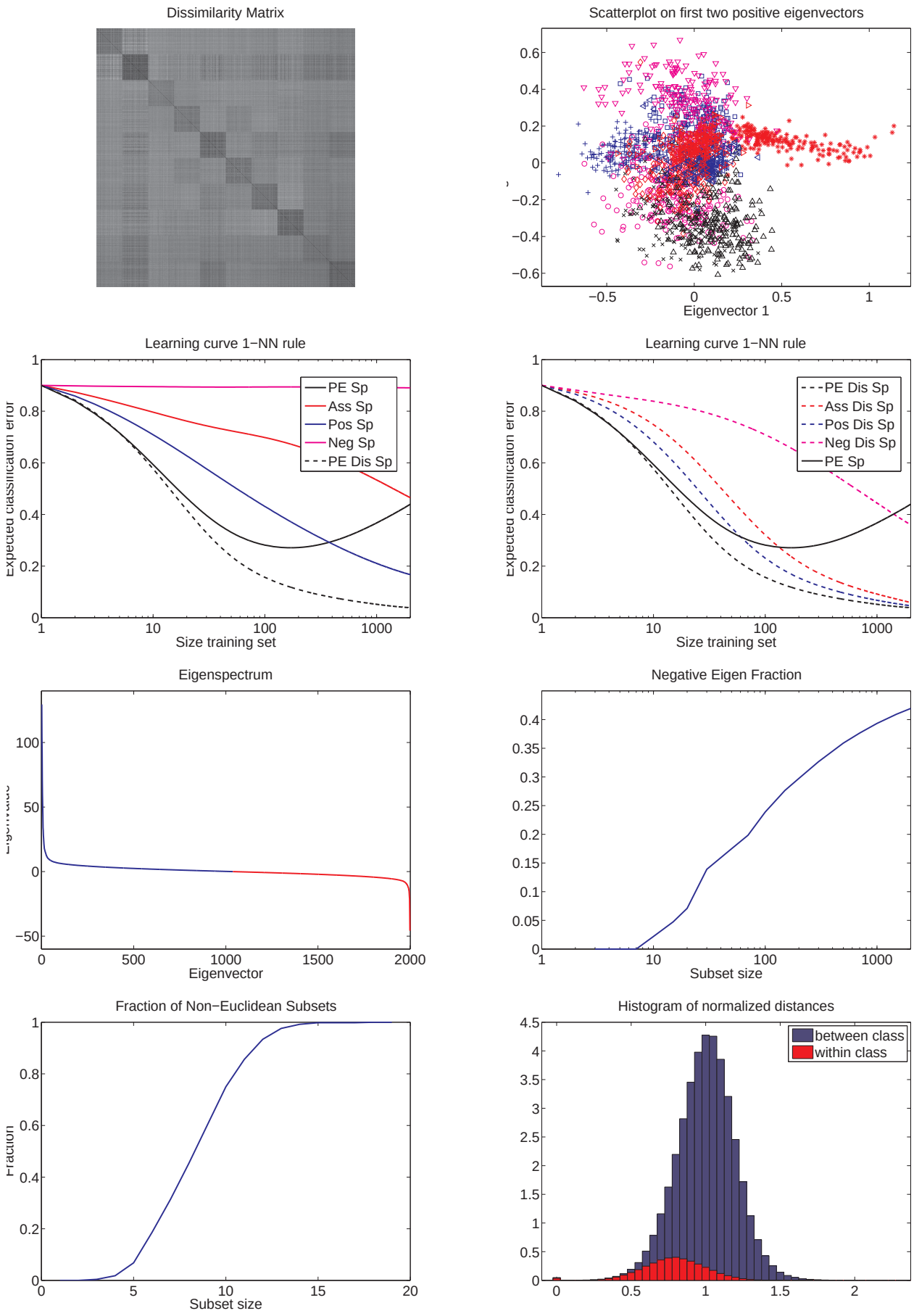


Figure 20: Graphical results for Zongker.

2.21 Chickenpieces-5-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 5. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.044	asymmetry
446	number of objects
300	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
268, 177	number of positive and negative eigenvalues
0.216	negative eigenfraction
0.017	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.814, 1.049	average within-class and between-class dissimilarity
34.5, 46.2, 25.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 35: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	40.1 (1.0)	50.2 (0.8)	46.8 (0.6)	81.4 (0.8)	40.1 (1.0)
Parzen	42.5 (0.6)	42.5 (0.4)	42.2 (0.5)	87.4 (1.0)	42.5 (0.6)
NM	42.1 (0.8)	61.8 (3.7)	45.8 (0.6)	73.8 (0.0)	41.2 (0.8)
SVM-1	35.6 (1.2)	28.8 (0.9)	26.3 (0.9)	70.7 (1.1)	33.5 (0.9)

Table 36: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	29.8 (1.1)	30.9 (1.0)	29.3 (1.0)	75.9 (0.8)	28.1 (0.8)
Parzen	43.7 (0.6)	42.3 (0.7)	42.7 (0.5)	65.7 (1.2)	43.7 (0.6)
NM	27.1 (0.5)	28.5 (0.9)	27.0 (0.7)	75.4 (1.1)	28.8 (0.5)
SVM-1	20.5 (0.6)	24.6 (0.8)	21.7 (0.7)	65.5 (1.3)	26.6 (0.7)

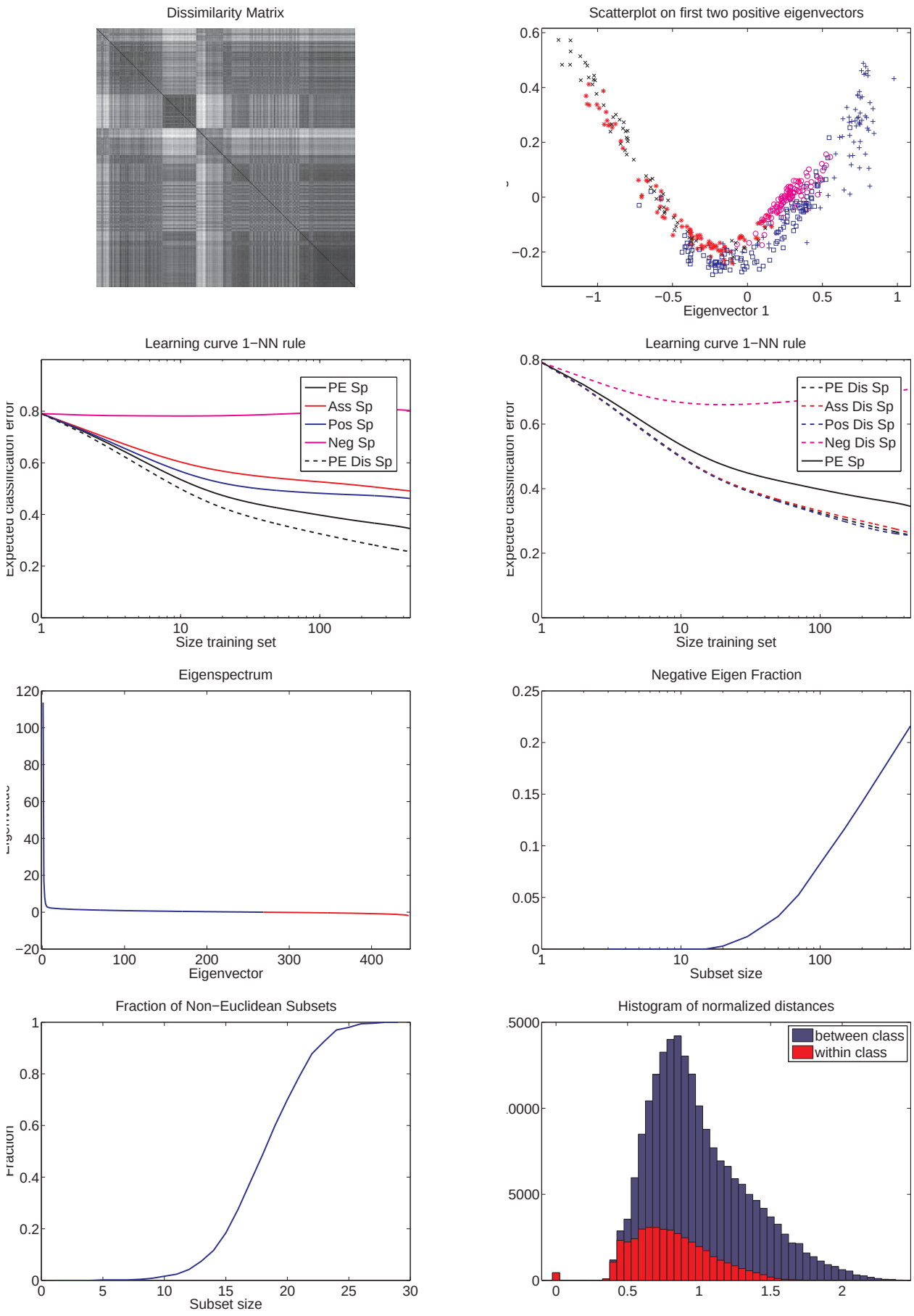


Figure 21: Graphical results for Chickenpieces-5-45.

2.22 Chickenpieces-5-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 5. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.039	asymmetry
446	number of objects
289	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
265, 180	number of positive and negative eigenvalues
0.200	negative eigenfraction
0.012	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.789, 1.055	average within-class and between-class dissimilarity
36.1, 49.3, 28.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 37: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	42.7 (0.8)	51.0 (0.7)	48.3 (0.8)	79.6 (0.8)	42.7 (0.8)
Parzen	43.4 (0.5)	43.1 (0.5)	43.2 (0.5)	86.7 (0.7)	43.4 (0.5)
NM	68.5 (4.2)	49.9 (0.5)	46.4 (0.5)	73.8 (0.0)	41.9 (0.8)
SVM-1	36.0 (1.3)	31.2 (0.8)	29.4 (1.0)	70.5 (0.8)	36.1 (0.7)

Table 38: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	32.0 (0.9)	33.9 (1.0)	32.5 (0.9)	73.9 (0.7)	32.1 (1.3)
Parzen	44.7 (0.5)	44.4 (0.5)	44.8 (0.5)	63.8 (0.9)	45.6 (0.4)
NM	29.6 (0.7)	31.5 (1.0)	30.1 (0.9)	73.0 (0.6)	33.6 (0.6)
SVM-1	21.9 (0.7)	26.9 (0.6)	23.6 (0.5)	66.1 (1.0)	27.4 (0.7)

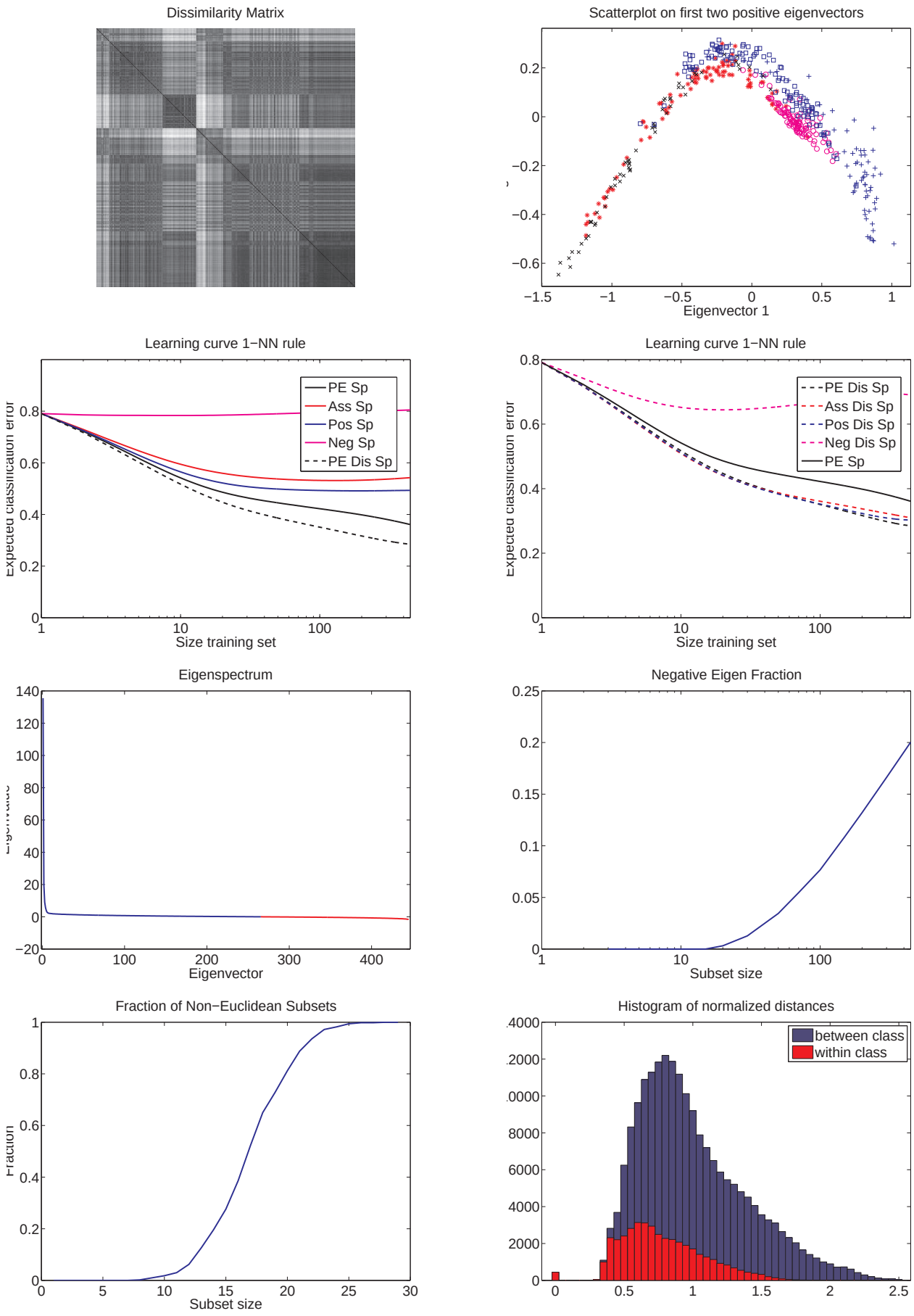


Figure 22: Graphical results for Chickenpieces-5-60.

2.23 Chickenpieces-5-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 5. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.038	asymmetry
446	number of objects
270	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
263, 182	number of positive and negative eigenvalues
0.165	negative eigenfraction
0.008	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.755, 1.064	average within-class and between-class dissimilarity
41.7, 48.2, 32.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 39: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	45.8 (0.9)	50.2 (0.5)	49.3 (0.5)	75.4 (1.0)	45.8 (0.9)
Parzen	44.9 (0.3)	44.8 (0.4)	45.0 (0.3)	82.6 (1.1)	44.9 (0.3)
NM	47.1 (2.8)	52.4 (2.3)	75.7 (3.0)	73.8 (0.0)	44.1 (0.8)
SVM-1	41.5 (1.0)	33.1 (0.5)	30.5 (0.5)	74.0 (0.9)	38.3 (1.0)

Table 40: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	35.0 (1.2)	34.6 (1.0)	35.3 (1.1)	70.6 (0.9)	35.3 (0.9)
Parzen	45.4 (0.4)	45.9 (0.5)	45.9 (0.4)	64.1 (1.1)	46.1 (0.3)
NM	33.2 (0.7)	33.0 (0.8)	33.0 (0.8)	70.1 (1.0)	37.9 (1.0)
SVM-1	24.5 (0.7)	28.2 (0.8)	25.5 (0.8)	65.1 (0.8)	29.6 (0.6)

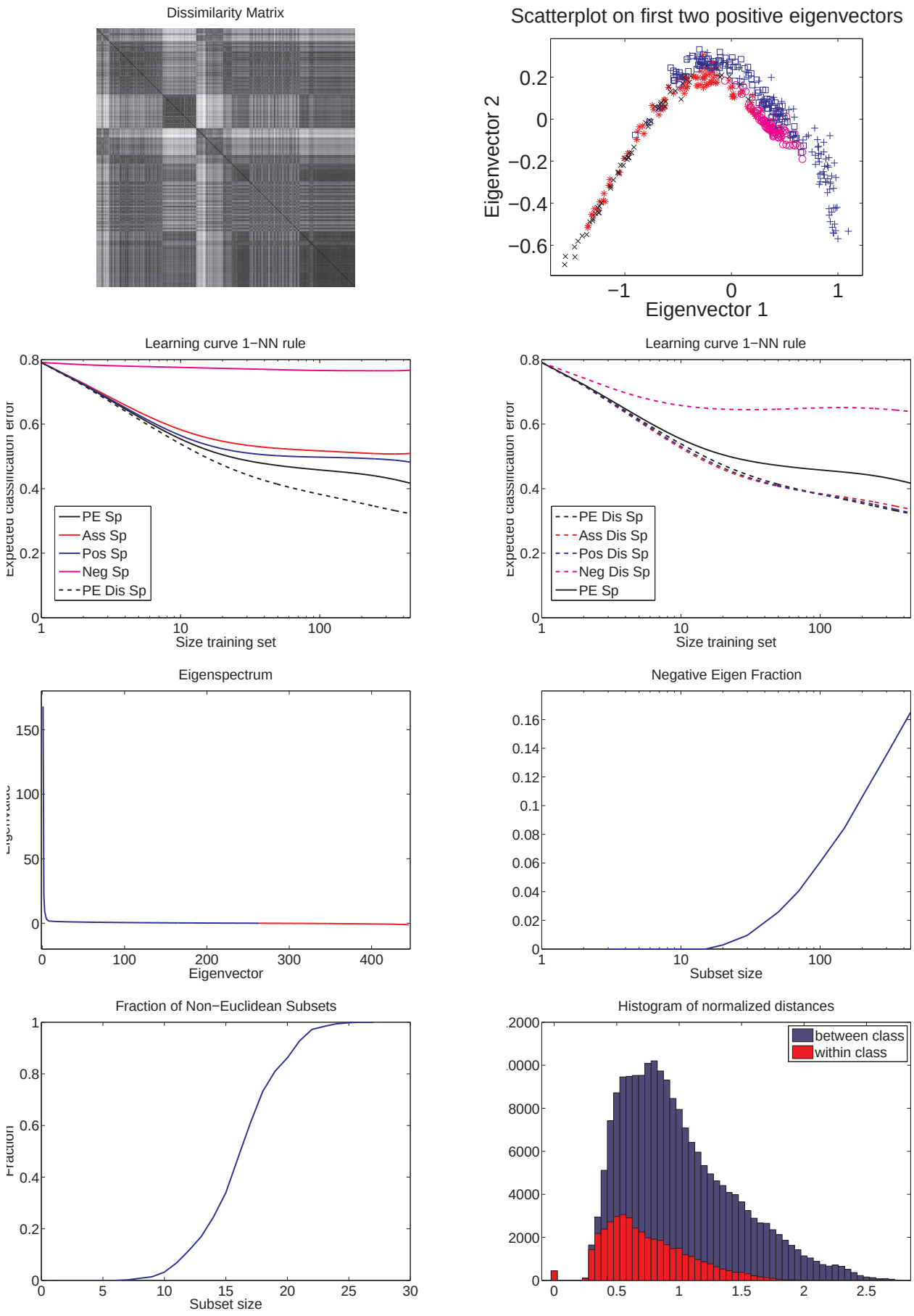


Figure 23: Graphical results for Chickenpieces-5-90.

2.24 Chickenpieces-5-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 5. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.038	asymmetry
446	number of objects
248	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
263, 182	number of positive and negative eigenvalues
0.133	negative eigenfraction
0.006	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.731, 1.070	average within-class and between-class dissimilarity
42.8, 49.3, 32.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 41: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	46.3 (0.6)	50.8 (0.7)	49.7 (0.6)	77.0 (1.0)	46.3 (0.6)
Parzen	45.6 (0.4)	45.5 (0.4)	45.5 (0.4)	79.5 (1.2)	45.6 (0.4)
NM	43.9 (0.8)	71.2 (3.5)	51.7 (3.0)	73.8 (0.0)	45.9 (2.1)
SVM-1	42.4 (1.2)	33.0 (1.2)	31.3 (1.1)	73.0 (1.6)	40.1 (0.9)

Table 42: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	37.0 (1.2)	37.9 (1.1)	36.9 (1.3)	72.2 (1.0)	38.3 (0.8)
Parzen	46.2 (0.4)	46.0 (0.4)	46.2 (0.3)	63.6 (0.8)	46.9 (0.3)
NM	36.1 (0.8)	36.0 (0.9)	36.0 (0.9)	71.1 (1.1)	40.4 (1.1)
SVM-1	26.3 (0.7)	30.1 (0.7)	28.0 (0.7)	66.1 (1.0)	30.8 (0.7)

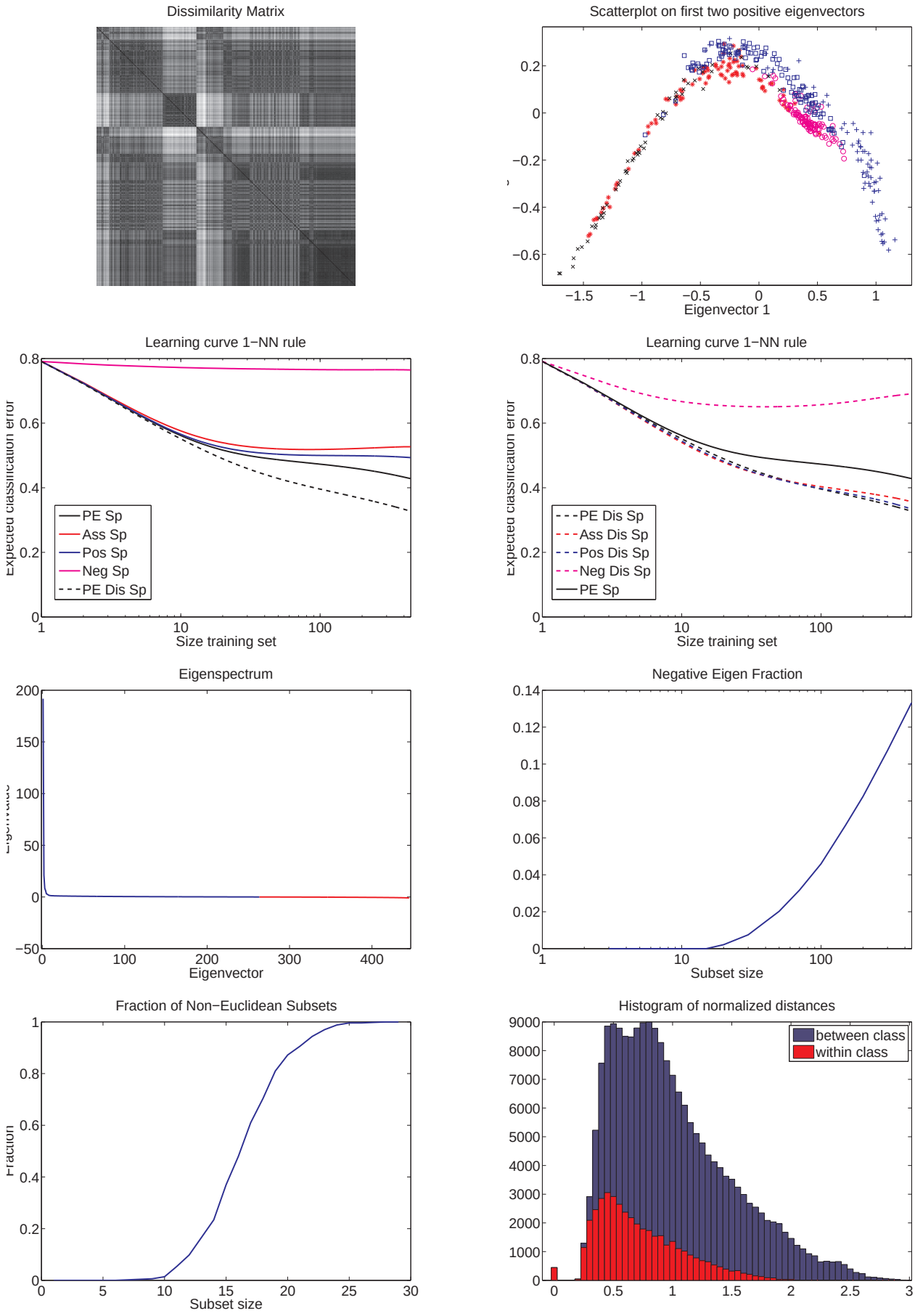


Figure 24: Graphical results for Chickenpieces-5-120.

2.25 Chickenpieces-7-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 7. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.044	asymmetry
446	number of objects
300	number of significant eigenvectors
1	number of triangle inequality violations out of 88120680
263, 182	number of positive and negative eigenvalues
0.235	negative eigenfraction
0.020	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.806, 1.051	average within-class and between-class dissimilarity
24.7, 39.9, 17.9	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 43: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	33.0 (1.0)	47.9 (0.8)	41.1 (0.8)	78.6 (0.7)	33.0 (1.0)
Parzen	40.2 (0.6)	40.4 (0.6)	39.9 (0.7)	87.4 (0.8)	40.2 (0.6)
NM	36.0 (0.9)	64.3 (4.0)	43.4 (0.8)	73.8 (0.0)	35.5 (0.9)
SVM-1	30.6 (0.7)	25.4 (0.9)	20.6 (0.9)	69.5 (0.5)	28.4 (0.6)

Table 44: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	26.8 (0.8)	28.5 (0.7)	27.0 (0.6)	71.2 (1.0)	23.9 (0.9)
Parzen	42.5 (0.6)	39.8 (0.7)	40.6 (0.6)	64.5 (1.1)	41.7 (0.6)
NM	24.6 (0.6)	27.0 (0.5)	25.5 (0.7)	70.5 (1.1)	25.8 (0.8)
SVM-1	16.9 (0.9)	22.1 (1.0)	18.5 (1.0)	61.2 (0.9)	23.9 (0.5)

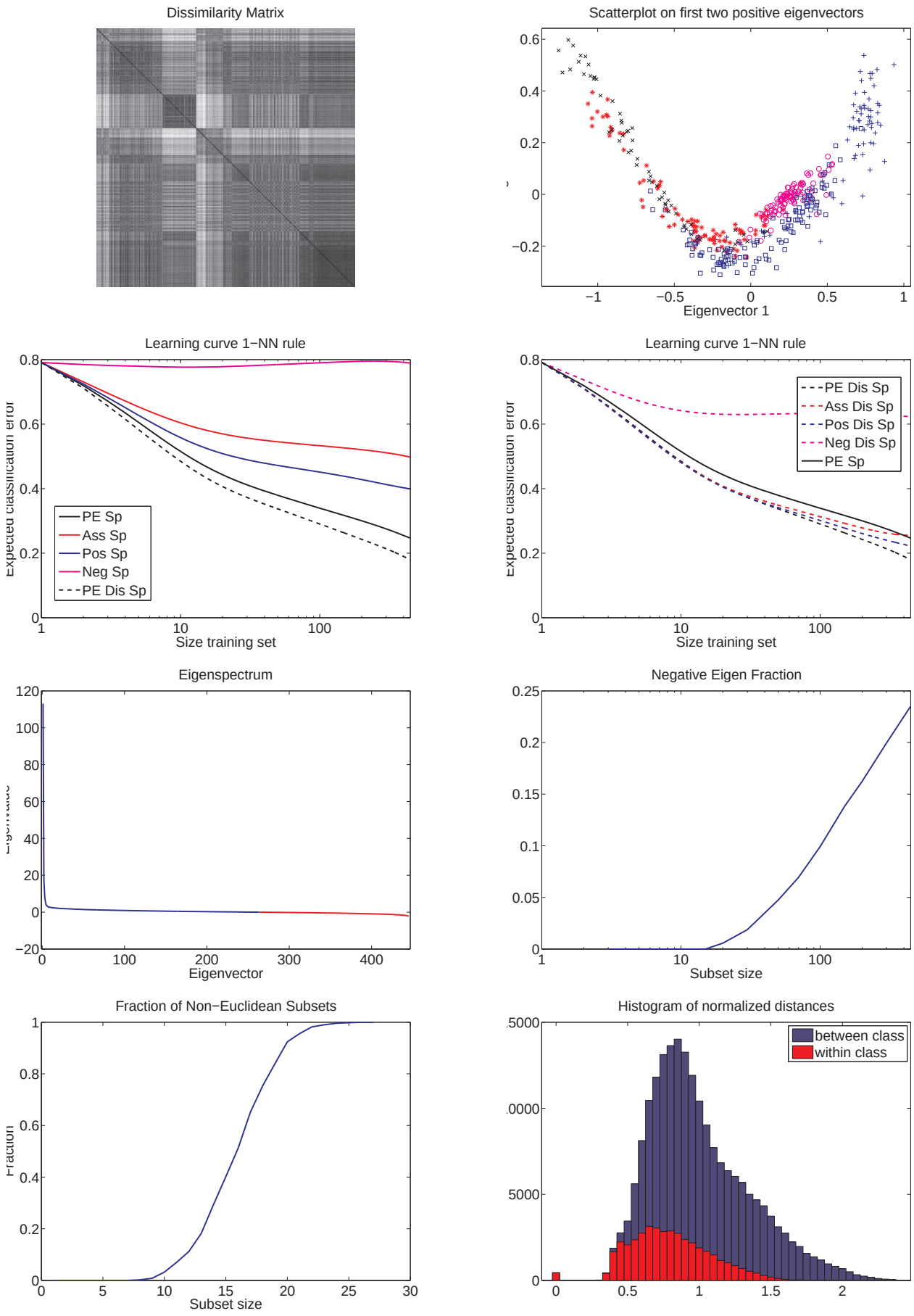


Figure 25: Graphical results for Chickenpieces-7-45.

2.26 Chickenpieces-7-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 7. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.040	asymmetry
446	number of objects
290	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
261, 184	number of positive and negative eigenvalues
0.219	negative eigenfraction
0.014	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.782, 1.057	average within-class and between-class dissimilarity
29.1, 43.7, 20.4	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 45: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	36.6 (0.8)	49.4 (0.9)	44.2 (0.6)	78.0 (0.9)	36.6 (0.8)
Parzen	41.9 (0.7)	41.9 (0.7)	41.8 (0.6)	86.2 (0.6)	41.9 (0.7)
NM	50.5 (5.4)	51.1 (2.0)	44.2 (0.7)	73.8 (0.0)	37.9 (0.9)
SVM-1	31.6 (1.3)	24.7 (0.6)	21.3 (0.7)	70.4 (1.2)	31.6 (0.7)

Table 46: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	29.7 (1.3)	32.4 (1.2)	30.4 (1.7)	68.9 (0.6)	27.3 (1.0)
Parzen	44.0 (0.6)	42.8 (0.7)	43.4 (0.5)	60.7 (1.3)	44.1 (0.5)
NM	28.0 (0.6)	29.9 (0.8)	28.5 (0.8)	67.9 (0.7)	29.6 (0.8)
SVM-1	18.1 (1.0)	22.0 (0.9)	19.6 (0.7)	59.9 (0.7)	24.8 (0.6)

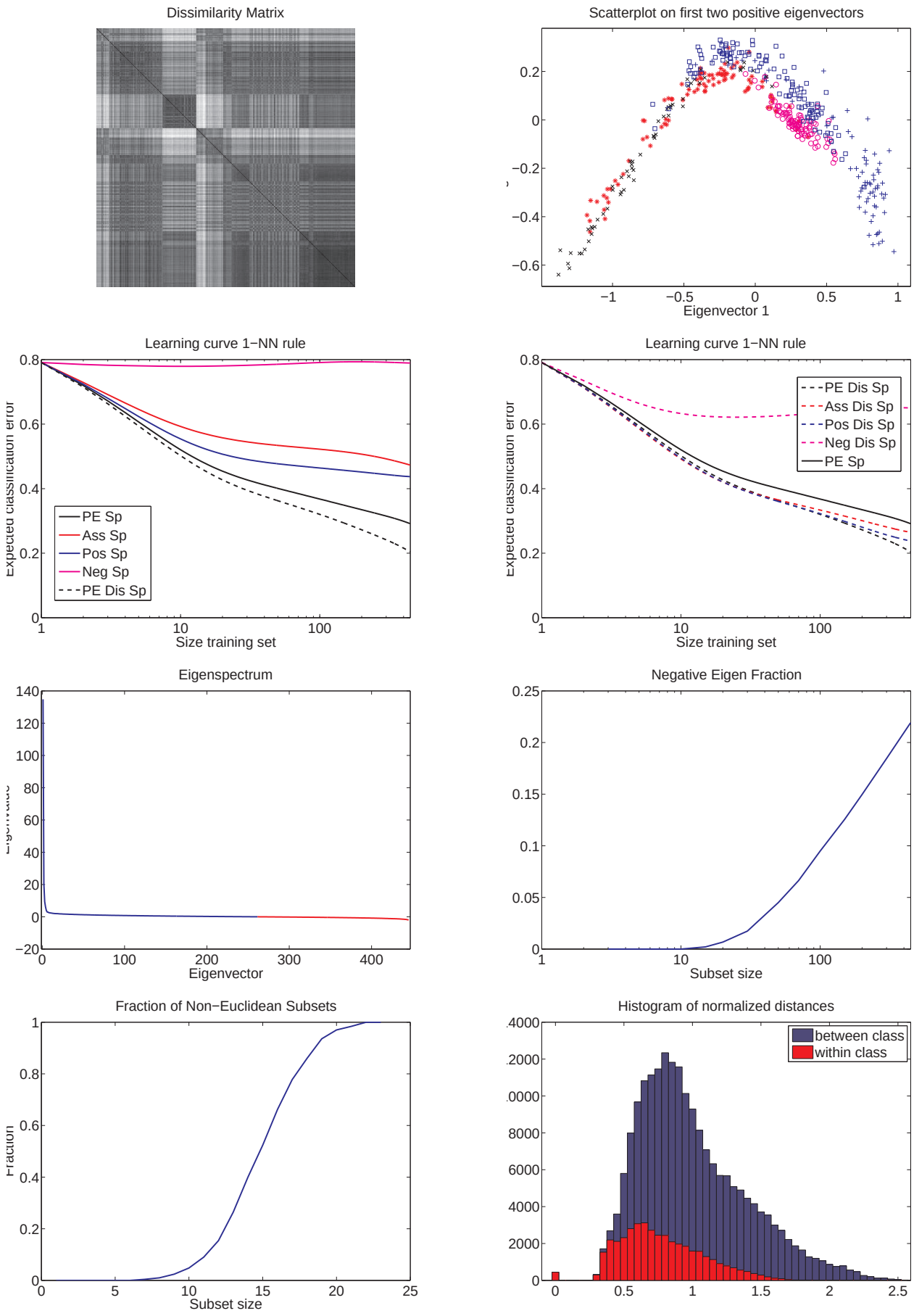


Figure 26: Graphical results for Chickenpieces-7-60.

2.27 Chickenpieces-7-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 7. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.039	asymmetry
446	number of objects
272	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
258, 187	number of positive and negative eigenvalues
0.186	negative eigenfraction
0.010	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.749, 1.066	average within-class and between-class dissimilarity
35.4, 46.9, 22.2	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 47: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	40.9 (0.7)	50.0 (0.5)	47.2 (0.8)	79.6 (0.6)	40.9 (0.7)
Parzen	44.1 (0.4)	43.5 (0.5)	43.8 (0.5)	81.1 (0.8)	44.1 (0.4)
NM	43.2 (2.4)	48.8 (0.8)	73.4 (5.2)	73.8 (0.0)	43.3 (2.2)
SVM-1	32.5 (1.1)	26.4 (0.6)	23.1 (0.9)	72.5 (1.1)	35.3 (0.7)

Table 48: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	32.7 (1.0)	37.0 (1.2)	34.3 (1.2)	71.5 (1.1)	31.3 (1.1)
Parzen	46.2 (0.4)	45.1 (0.5)	45.5 (0.5)	61.8 (0.7)	46.6 (0.4)
NM	31.4 (0.7)	33.0 (0.7)	32.1 (0.6)	70.6 (0.9)	34.5 (0.6)
SVM-1	20.0 (0.9)	26.3 (0.7)	22.8 (1.0)	63.5 (1.0)	26.3 (0.8)

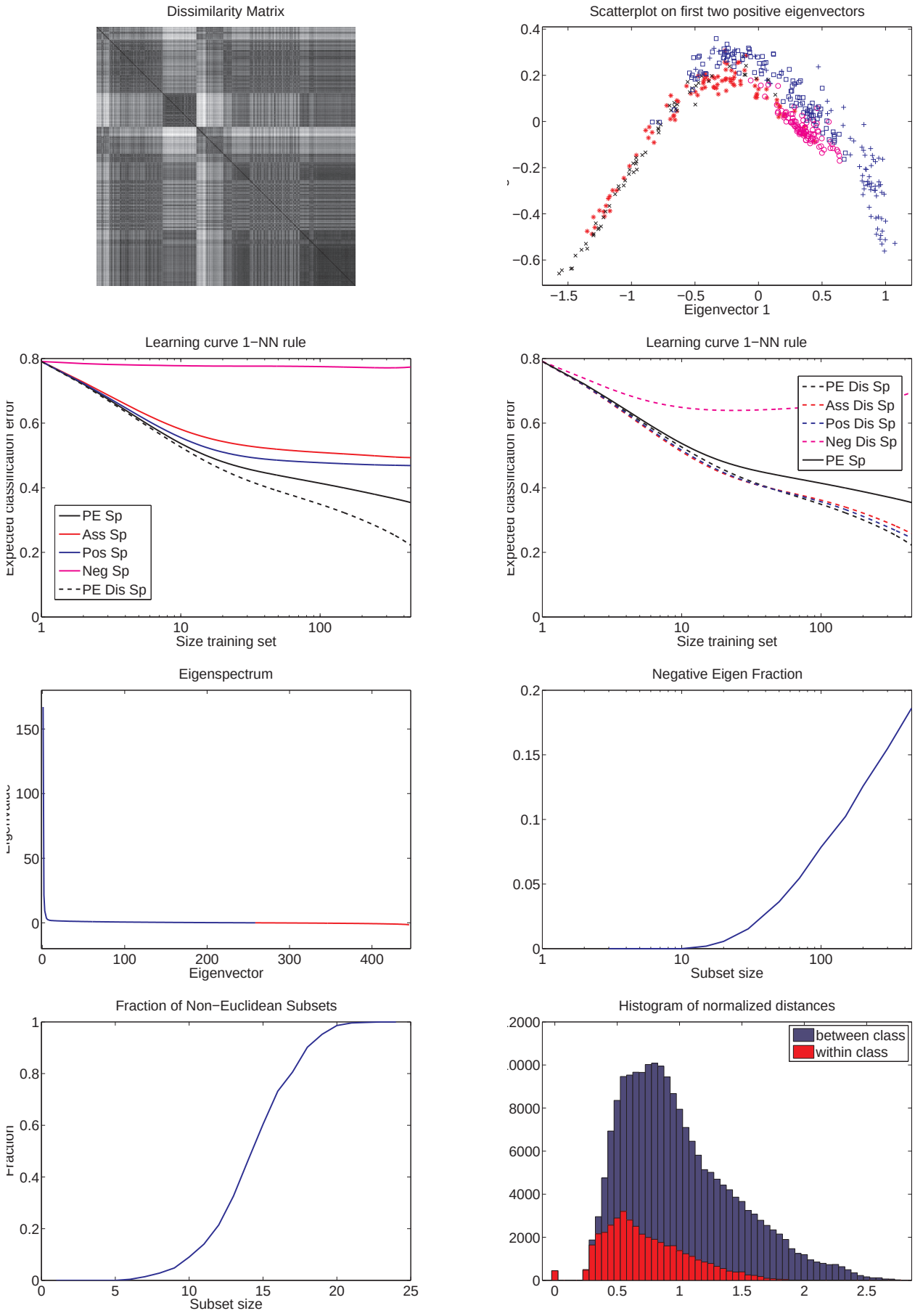


Figure 27: Graphical results for Chickenpieces-7-90.

2.28 Chickenpieces-7-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 7. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.039	asymmetry
446	number of objects
252	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
257, 188	number of positive and negative eigenvalues
0.153	negative eigenfraction
0.008	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.726, 1.072	average within-class and between-class dissimilarity
40.6, 49.1, 28.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 49: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	44.9 (0.6)	49.9 (0.8)	48.6 (0.6)	76.1 (0.5)	44.9 (0.6)
Parzen	45.0 (0.5)	44.5 (0.5)	44.7 (0.6)	78.0 (0.8)	45.0 (0.5)
NM	42.6 (0.9)	76.2 (2.8)	53.7 (3.3)	73.8 (0.0)	42.5 (0.8)
SVM-1	37.4 (1.1)	28.2 (1.0)	26.2 (0.9)	72.4 (0.9)	37.0 (0.8)

Table 50: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	35.0 (1.1)	36.5 (1.3)	35.5 (1.2)	67.9 (1.1)	33.9 (1.2)
Parzen	46.7 (0.6)	46.1 (0.5)	46.5 (0.4)	60.0 (0.8)	46.8 (0.3)
NM	34.3 (0.7)	34.9 (0.6)	33.9 (0.7)	66.7 (1.3)	38.3 (0.8)
SVM-1	22.3 (0.8)	27.1 (1.0)	24.7 (1.1)	63.9 (1.5)	28.7 (1.0)

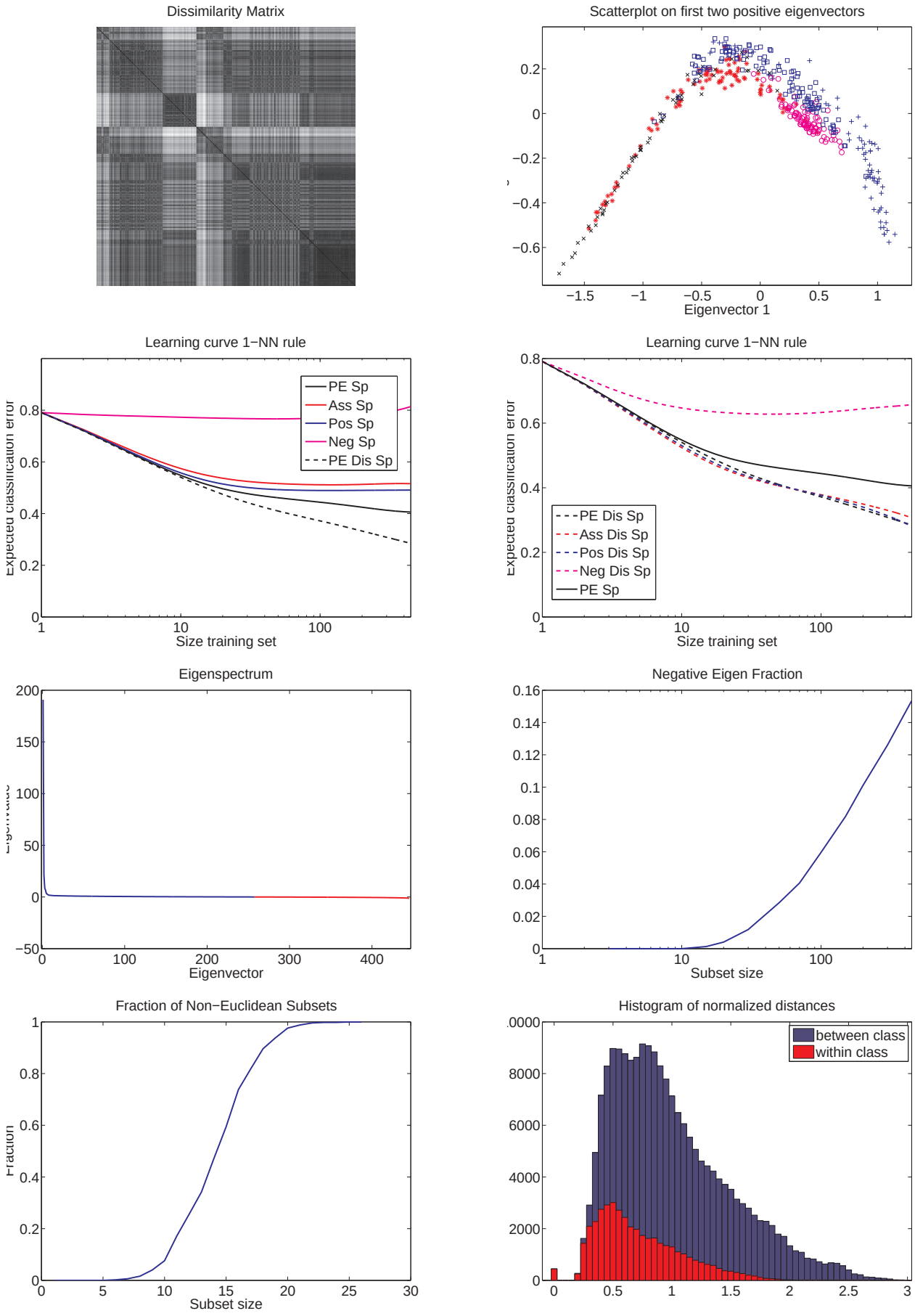


Figure 28: Graphical results for Chickenpieces-7-120.

2.29 Chickenpieces-10-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 10. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.046	asymmetry
446	number of objects
301	number of significant eigenvectors
1	number of triangle inequality violations out of 88120680
255, 190	number of positive and negative eigenvalues
0.257	negative eigenfraction
0.023	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.796, 1.053	average within-class and between-class dissimilarity
16.1, 34.5, 12.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 51: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	25.7 (1.1)	44.2 (0.9)	36.1 (1.1)	79.7 (0.6)	25.7 (1.1)
Parzen	34.3 (0.6)	34.9 (0.7)	34.7 (0.7)	87.3 (0.6)	34.3 (0.6)
NM	27.9 (1.0)	51.1 (2.8)	38.2 (1.2)	73.8 (0.0)	28.1 (1.0)
SVM-1	23.5 (0.7)	15.6 (0.8)	12.2 (0.6)	65.2 (1.4)	19.8 (0.6)

Table 52: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	20.7 (1.2)	24.1 (1.0)	22.1 (1.2)	69.3 (1.2)	19.3 (1.3)
Parzen	39.7 (0.7)	37.9 (0.6)	38.5 (0.7)	63.6 (1.0)	38.7 (0.6)
NM	20.0 (1.2)	22.2 (0.9)	20.6 (1.1)	68.2 (1.0)	21.0 (0.9)
SVM-1	11.5 (0.6)	15.8 (0.7)	12.4 (0.6)	61.3 (1.3)	19.4 (0.8)

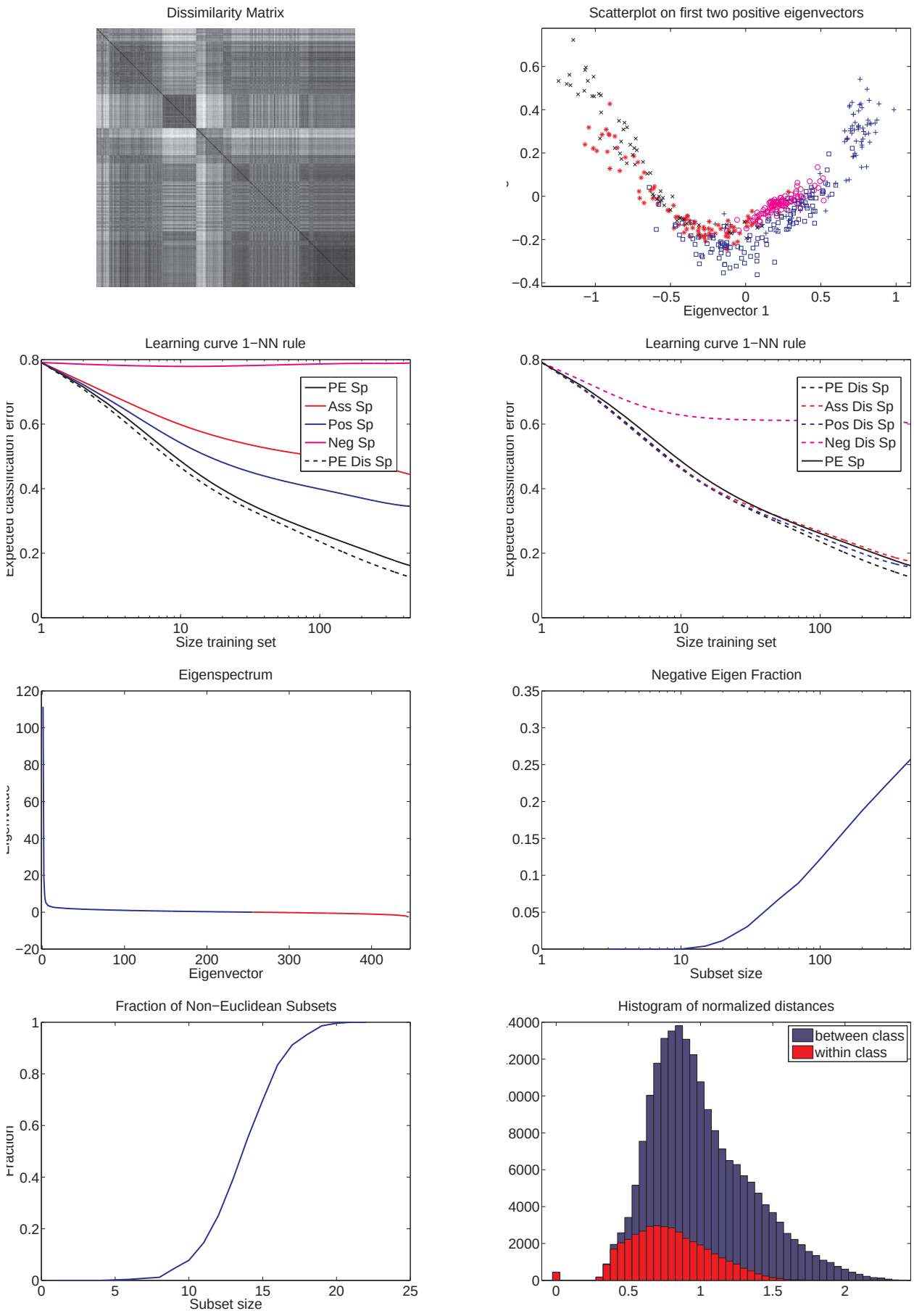


Figure 29: Graphical results for Chickenpieces-10-45.

2.30 Chickenpieces-10-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 10. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.041	asymmetry
446	number of objects
292	number of significant eigenvectors
3	number of triangle inequality violations out of 88120680
253, 192	number of positive and negative eigenvalues
0.244	negative eigenfraction
0.018	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.773, 1.059	average within-class and between-class dissimilarity
17.7, 37.0, 14.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 53: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	28.2 (1.1)	44.9 (0.8)	39.3 (0.8)	77.8 (0.9)	28.2 (1.1)
Parzen	38.8 (0.9)	39.5 (0.8)	39.0 (0.9)	85.5 (0.7)	38.8 (0.9)
NM	28.2 (1.0)	46.4 (0.9)	38.7 (0.9)	73.8 (0.0)	29.3 (1.1)
SVM-1	24.3 (0.5)	16.1 (0.9)	14.0 (0.9)	67.7 (0.5)	23.0 (0.7)

Table 54: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	25.2 (1.1)	29.2 (1.2)	27.5 (1.2)	68.5 (1.2)	23.3 (0.9)
Parzen	42.6 (0.7)	39.2 (0.8)	41.0 (1.0)	60.3 (1.2)	42.4 (0.5)
NM	23.7 (1.0)	26.5 (0.8)	25.5 (1.0)	67.1 (1.1)	25.5 (1.0)
SVM-1	12.6 (0.8)	17.3 (1.0)	13.7 (0.9)	58.4 (1.3)	20.7 (0.9)

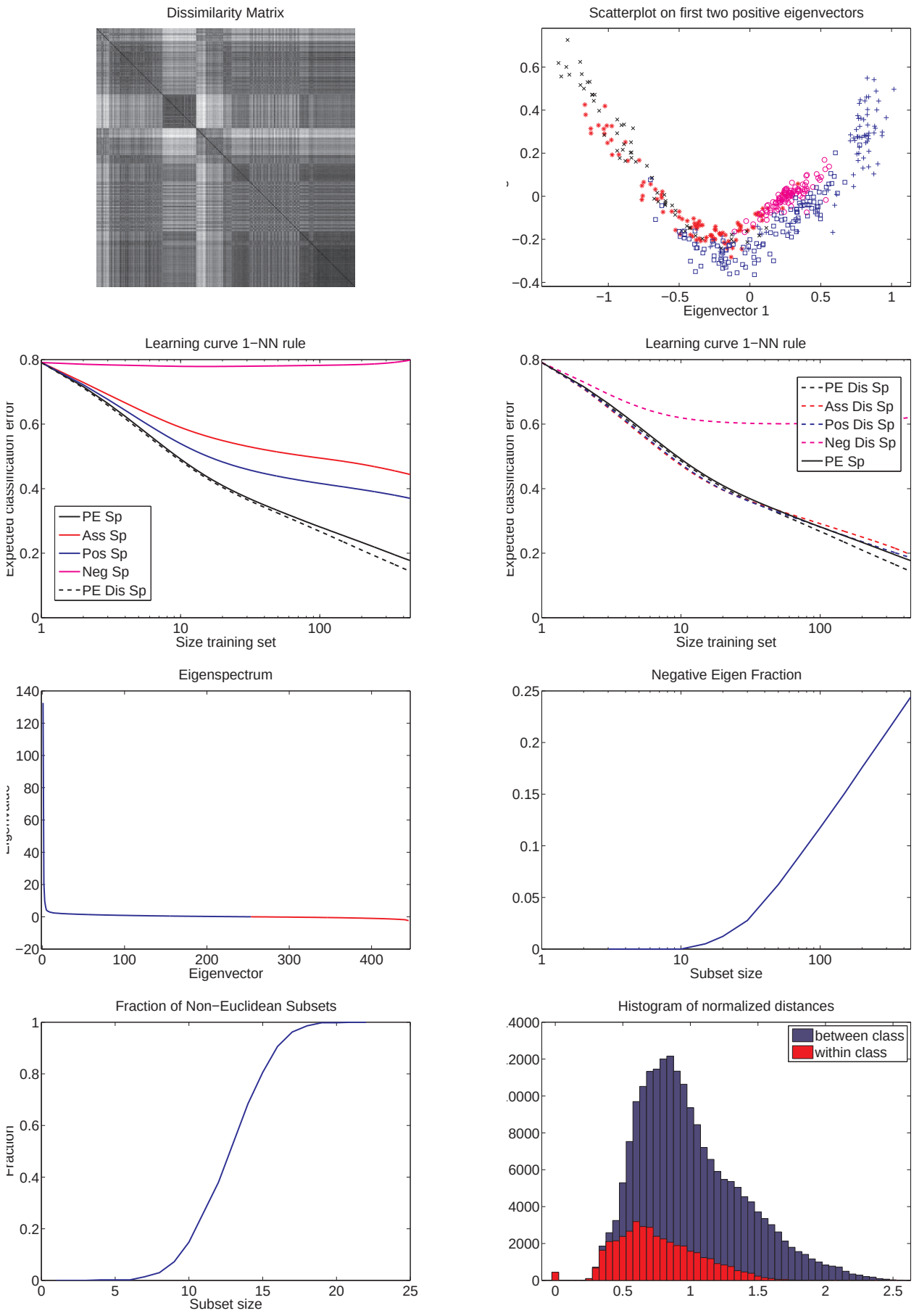


Figure 30: Graphical results for Chickenpieces-10-60.

2.31 Chickenpieces-10-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 10. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.040	asymmetry
446	number of objects
275	number of significant eigenvectors
0	number of triangle inequality violations out of 88120680
251, 194	number of positive and negative eigenvalues
0.211	negative eigenfraction
0.012	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.742, 1.068	average within-class and between-class dissimilarity
22.2, 37.9, 17.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 55: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	31.2 (1.1)	45.7 (0.8)	40.6 (0.7)	77.7 (1.1)	31.2 (1.1)
Parzen	42.2 (0.6)	42.8 (0.6)	42.5 (0.6)	85.4 (0.9)	42.2 (0.6)
NM	33.1 (1.1)	46.5 (0.9)	51.2 (4.8)	73.8 (0.0)	33.1 (1.1)
SVM-1	27.9 (1.0)	20.1 (0.7)	17.2 (0.7)	67.6 (0.8)	27.9 (0.5)

Table 56: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	27.8 (1.2)	32.5 (1.1)	30.8 (1.3)	71.5 (1.3)	27.7 (0.8)
Parzen	45.1 (0.5)	42.7 (0.8)	44.2 (0.5)	60.1 (1.1)	44.8 (0.5)
NM	28.6 (1.2)	30.2 (1.2)	29.1 (1.3)	70.6 (1.1)	32.2 (1.0)
SVM-1	15.3 (0.9)	20.6 (0.9)	17.5 (1.0)	61.9 (1.1)	23.7 (0.8)

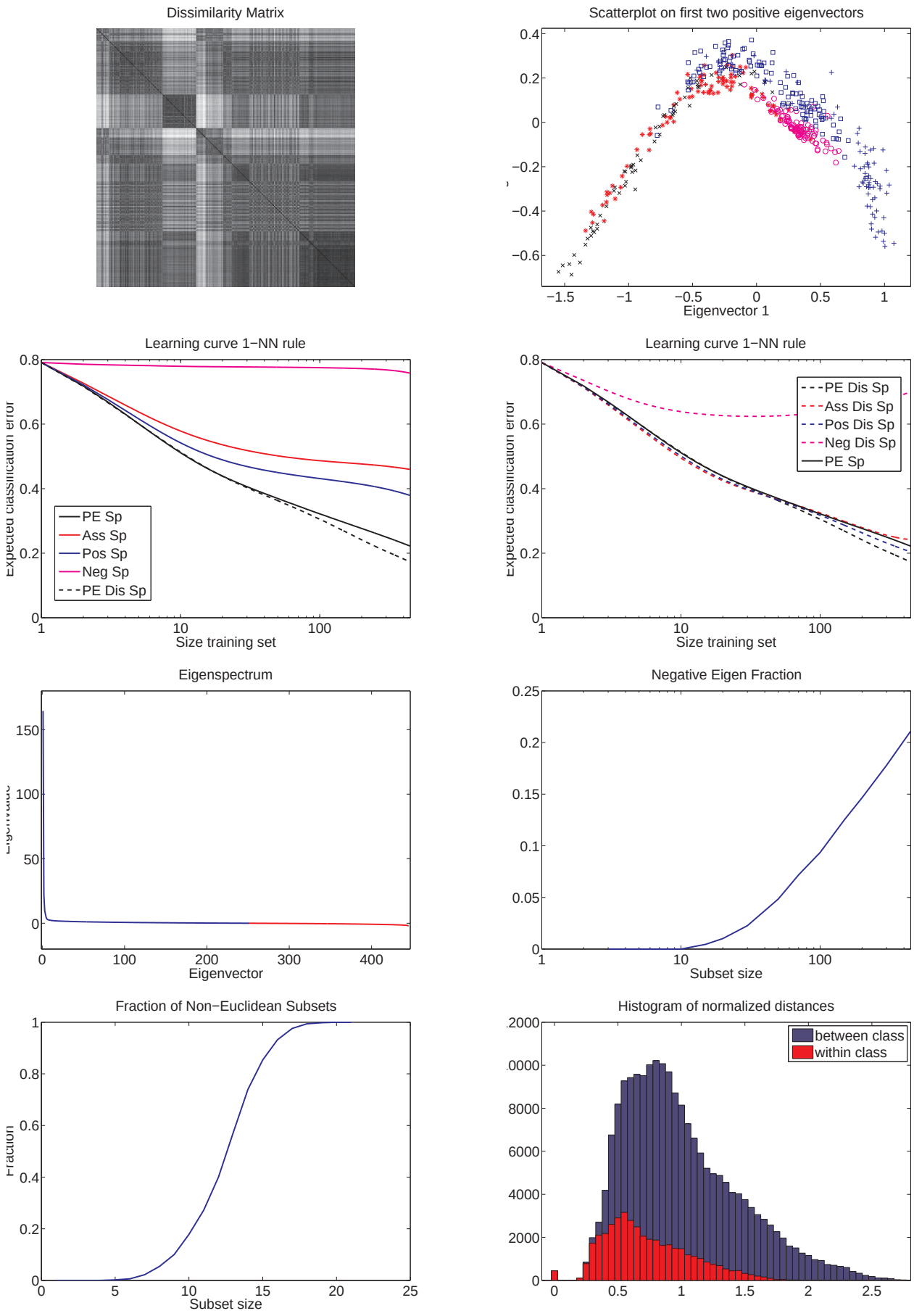


Figure 31: Graphical results for Chickenpieces-10-90.

2.32 Chickenpieces-10-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 10. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.040	asymmetry
446	number of objects
257	number of significant eigenvectors
1	number of triangle inequality violations out of 88120680
250, 195	number of positive and negative eigenvalues
0.178	negative eigenfraction
0.010	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.720, 1.073	average within-class and between-class dissimilarity
28.3, 40.4, 18.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 57: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	36.3 (0.8)	45.3 (0.9)	42.2 (1.1)	76.2 (0.6)	36.3 (0.8)
Parzen	43.9 (0.4)	43.9 (0.3)	43.9 (0.4)	81.9 (0.9)	43.9 (0.4)
NM	36.1 (1.2)	61.4 (4.5)	45.2 (3.1)	73.8 (0.0)	43.4 (3.8)
SVM-1	29.2 (1.1)	21.3 (0.7)	18.1 (0.8)	70.4 (1.0)	31.8 (0.8)

Table 58: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	31.0 (1.0)	34.0 (1.4)	33.1 (1.4)	71.5 (1.1)	30.7 (0.8)
Parzen	45.8 (0.5)	45.0 (0.5)	45.3 (0.5)	63.7 (0.9)	45.9 (0.5)
NM	31.7 (1.0)	32.9 (1.1)	31.9 (0.9)	70.8 (1.3)	36.2 (1.0)
SVM-1	17.8 (0.9)	23.0 (0.9)	19.7 (1.0)	64.2 (1.3)	25.1 (0.8)

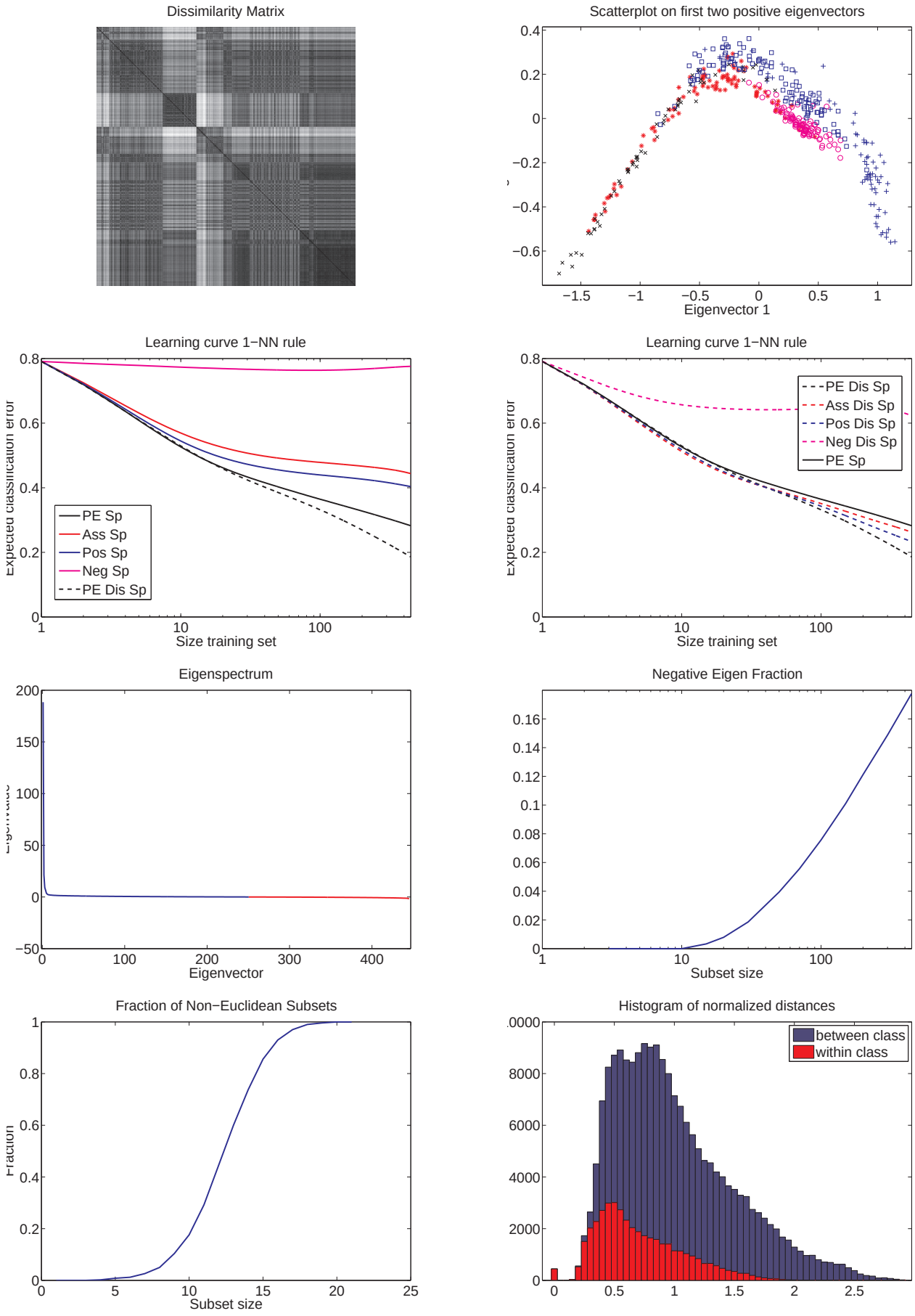


Figure 32: Graphical results for Chickenpieces-10-120.

2.33 Chickenpieces-15-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 15. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.051	asymmetry
446	number of objects
306	number of significant eigenvectors
74	number of triangle inequality violations out of 88120680
246, 199	number of positive and negative eigenvalues
0.286	negative eigenfraction
0.028	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.786, 1.056	average within-class and between-class dissimilarity
7.4, 24.0, 4.9	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 59: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	15.6 (0.5)	37.8 (0.7)	26.8 (0.8)	81.1 (0.8)	15.6 (0.5)
Parzen	26.3 (0.7)	26.7 (0.4)	26.5 (0.5)	91.5 (0.9)	26.3 (0.7)
NM	16.9 (0.7)	39.1 (0.8)	29.2 (0.7)	73.8 (0.0)	17.0 (0.7)
SVM-1	20.5 (0.7)	11.4 (0.6)	9.2 (0.6)	62.4 (0.9)	14.5 (0.6)

Table 60: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	12.3 (0.8)	16.5 (1.2)	13.8 (1.2)	74.6 (1.1)	11.9 (0.8)
Parzen	35.3 (0.6)	35.4 (0.5)	35.3 (0.6)	64.0 (1.2)	33.8 (0.6)
NM	11.6 (0.8)	15.6 (1.0)	13.6 (1.0)	74.3 (1.2)	13.3 (0.7)
SVM-1	8.8 (0.5)	12.1 (0.6)	10.0 (0.5)	60.9 (1.0)	14.9 (0.7)

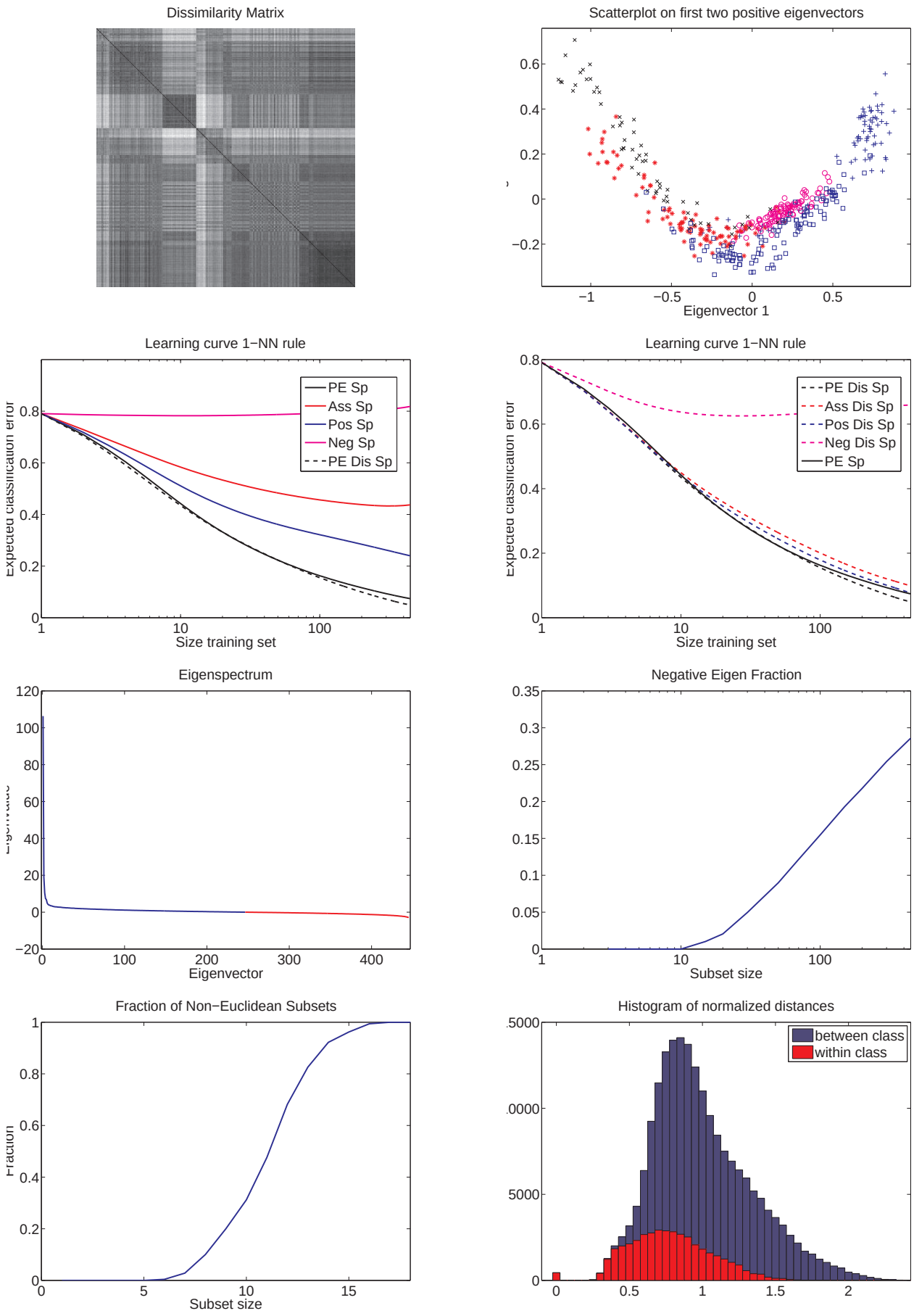


Figure 33: Graphical results for Chickenpieces-15-45.

2.34 Chickenpieces-15-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 15. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.044	asymmetry
446	number of objects
297	number of significant eigenvectors
50	number of triangle inequality violations out of 88120680
245, 200	number of positive and negative eigenvalues
0.274	negative eigenfraction
0.023	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.763, 1.062	average within-class and between-class dissimilarity
7.8, 28.7, 7.0	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 61: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	16.1 (0.6)	38.7 (0.6)	30.1 (0.8)	78.3 (1.1)	16.1 (0.6)
Parzen	31.0 (0.7)	32.3 (0.7)	31.4 (0.7)	89.2 (0.7)	31.0 (0.7)
NM	17.2 (0.6)	41.5 (0.8)	30.8 (0.8)	73.8 (0.0)	17.8 (0.6)
SVM-1	20.0 (0.7)	13.1 (0.6)	10.5 (0.6)	63.4 (0.7)	16.6 (0.5)

Table 62: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	15.0 (0.8)	18.5 (0.9)	16.5 (0.9)	69.3 (1.5)	14.0 (0.8)
Parzen	39.0 (0.6)	36.7 (0.4)	37.6 (0.4)	60.3 (1.1)	38.1 (0.6)
NM	14.6 (0.7)	17.9 (0.8)	16.1 (0.7)	67.9 (1.5)	17.2 (0.8)
SVM-1	8.6 (0.3)	12.7 (0.6)	10.4 (0.5)	57.5 (1.3)	16.1 (0.5)

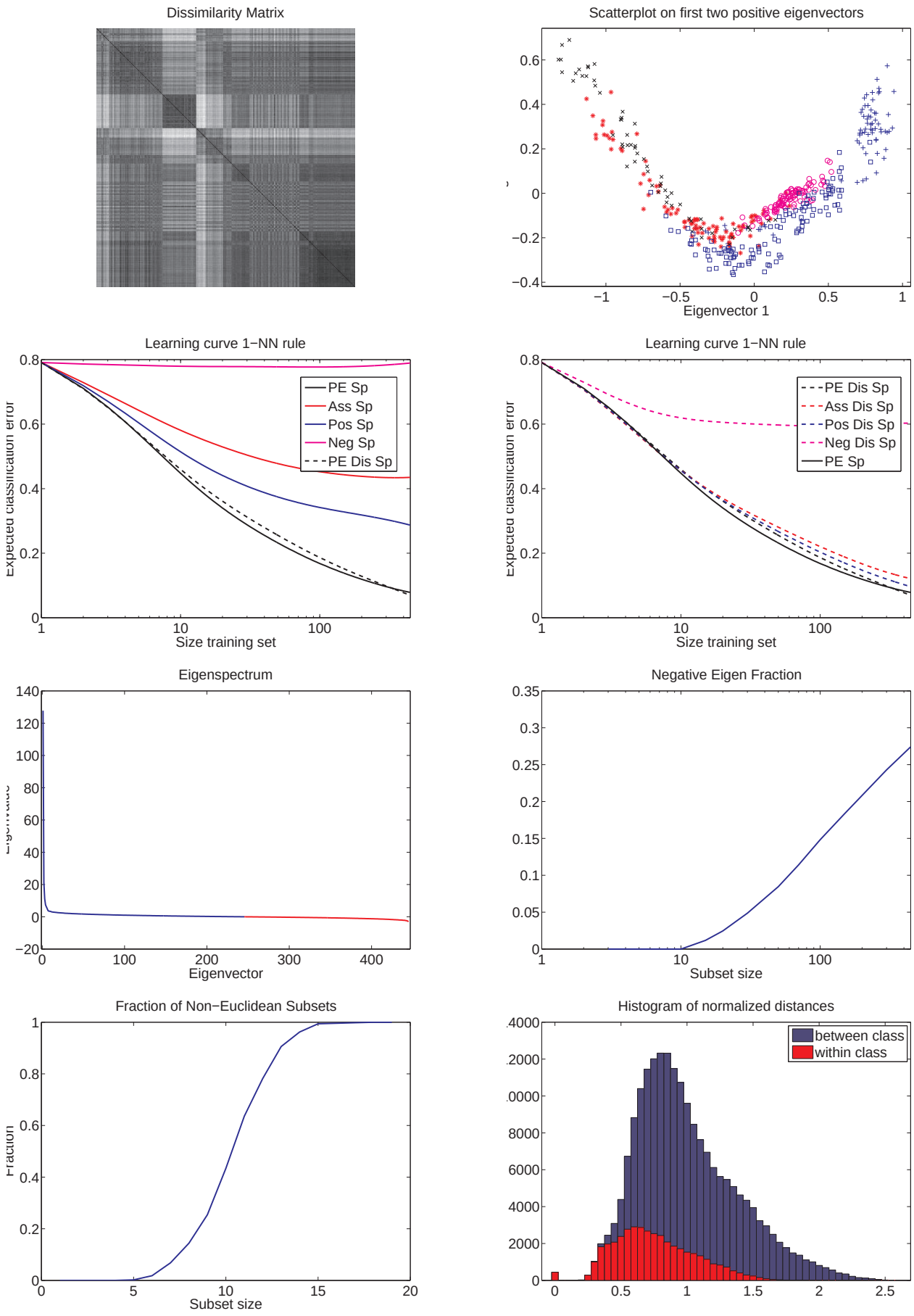


Figure 34: Graphical results for Chickenpieces-15-60.

2.35 Chickenpieces-15-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 15. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.043	asymmetry
446	number of objects
280	number of significant eigenvectors
95	number of triangle inequality violations out of 88120680
244, 201	number of positive and negative eigenvalues
0.243	negative eigenfraction
0.016	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.733, 1.070	average within-class and between-class dissimilarity
10.8, 31.8, 11.2	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 63: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	20.6 (0.7)	41.4 (1.0)	34.1 (1.0)	78.2 (0.6)	20.6 (0.7)
Parzen	37.2 (0.8)	37.8 (0.7)	37.2 (0.7)	88.7 (0.8)	37.2 (0.8)
NM	23.3 (1.0)	42.4 (1.1)	35.2 (1.2)	73.8 (0.0)	22.2 (0.9)
SVM-1	23.3 (0.8)	14.0 (0.6)	12.2 (0.6)	65.3 (0.9)	19.3 (0.5)

Table 64: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	20.7 (0.5)	24.3 (0.6)	21.6 (0.6)	70.0 (1.0)	19.3 (0.7)
Parzen	42.9 (0.6)	39.5 (0.6)	40.8 (0.7)	60.9 (0.6)	42.7 (0.5)
NM	20.8 (0.7)	22.6 (0.8)	21.6 (0.9)	68.2 (0.9)	25.1 (0.7)
SVM-1	10.1 (0.4)	15.3 (0.7)	12.4 (0.5)	58.2 (1.4)	18.3 (0.3)

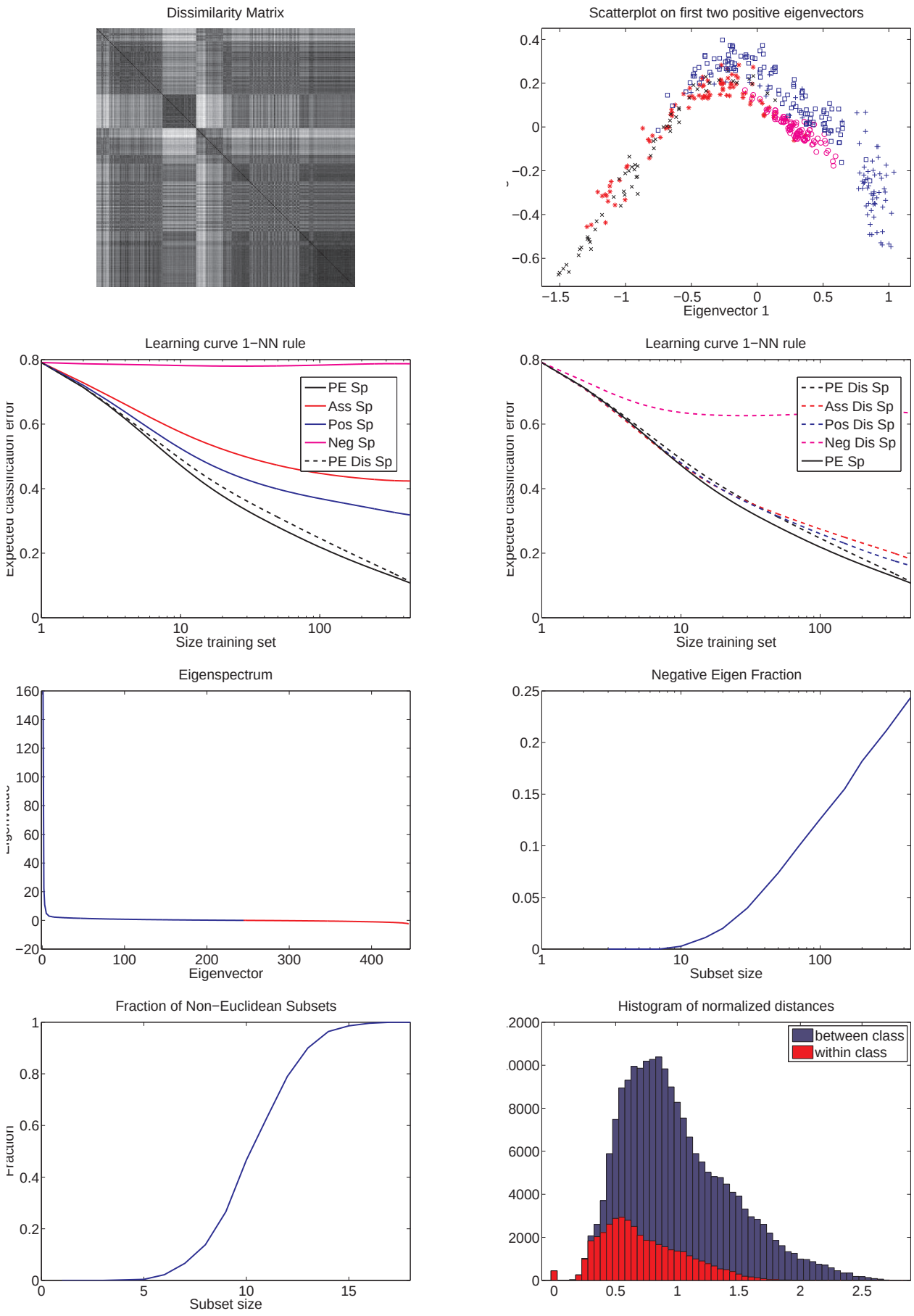


Figure 35: Graphical results for Chickenpieces-15-90.

2.36 Chickenpieces-15-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 15. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.043	asymmetry
446	number of objects
264	number of significant eigenvectors
25	number of triangle inequality violations out of 88120680
244, 201	number of positive and negative eigenvalues
0.208	negative eigenfraction
0.012	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.713, 1.075	average within-class and between-class dissimilarity
15.0, 33.6, 16.8	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 65: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	25.8 (0.9)	41.9 (0.9)	36.4 (0.8)	77.7 (0.8)	25.8 (0.9)
Parzen	40.3 (0.7)	39.9 (0.7)	40.4 (0.6)	86.7 (0.7)	40.3 (0.7)
NM	28.1 (1.3)	42.4 (1.1)	36.4 (1.0)	73.8 (0.0)	28.2 (1.2)
SVM-1	23.1 (0.6)	15.9 (0.5)	12.6 (0.6)	67.6 (0.7)	22.7 (0.6)

Table 66: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	25.4 (0.6)	28.3 (0.9)	27.5 (0.5)	68.5 (1.2)	24.4 (0.6)
Parzen	43.9 (0.5)	41.6 (0.7)	43.4 (0.5)	61.9 (0.6)	44.1 (0.5)
NM	26.0 (0.4)	26.6 (0.7)	26.1 (0.8)	67.9 (1.4)	29.7 (0.5)
SVM-1	11.8 (0.5)	16.8 (0.6)	13.9 (0.5)	59.1 (0.7)	20.2 (0.5)

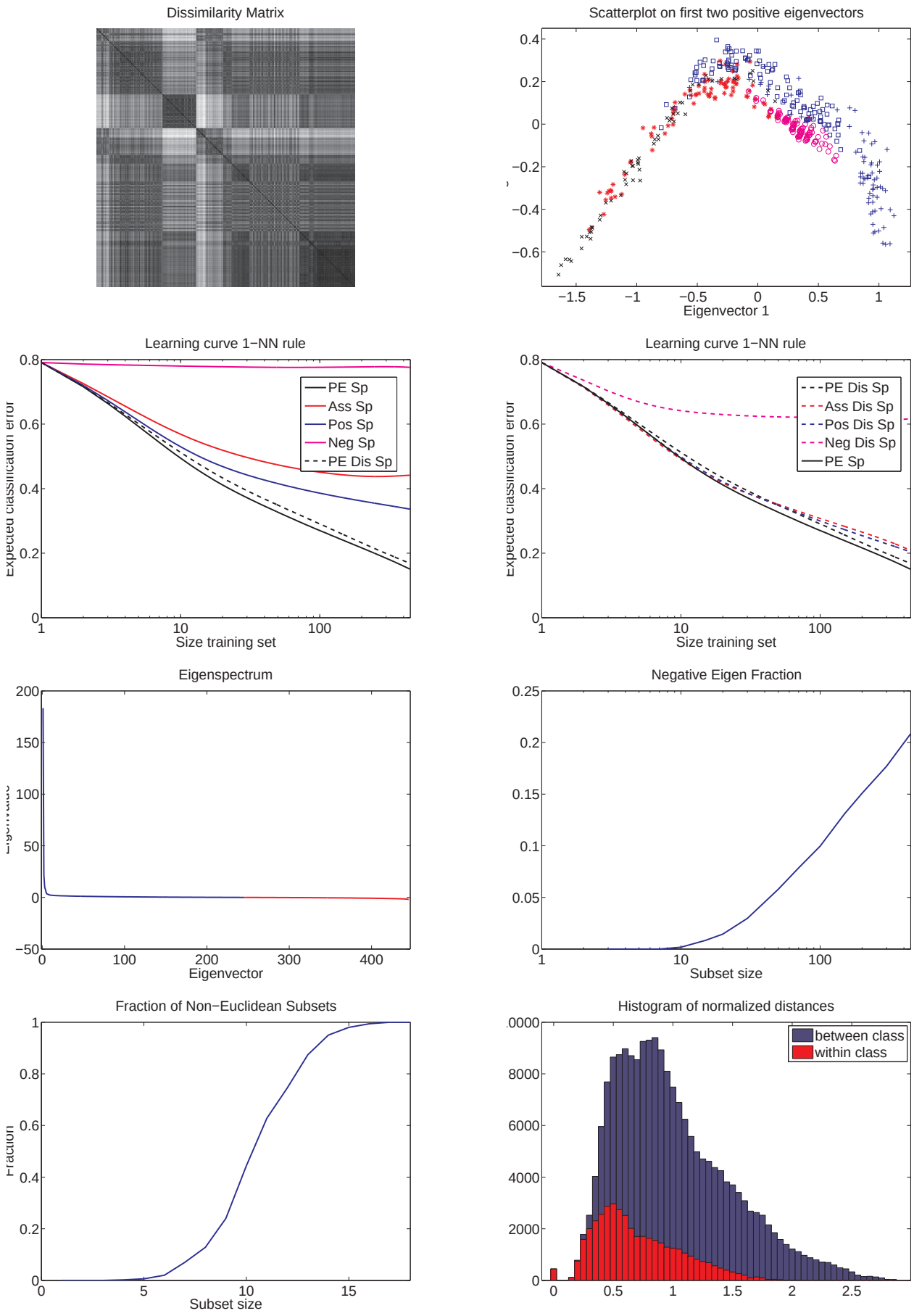


Figure 36: Graphical results for Chickenpieces-15-120.

2.37 Chickenpieces-20-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 20. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.057	asymmetry
446	number of objects
308	number of significant eigenvectors
695	number of triangle inequality violations out of 88120680
243, 202	number of positive and negative eigenvalues
0.307	negative eigenfraction
0.034	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.781, 1.057	average within-class and between-class dissimilarity
6.3, 17.0, 5.4	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 67: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	14.1 (0.6)	34.3 (1.1)	21.7 (0.9)	78.9 (1.0)	14.1 (0.6)
Parzen	20.4 (0.6)	21.8 (0.6)	21.2 (0.7)	90.6 (0.8)	20.4 (0.6)
NM	14.9 (0.8)	37.6 (1.2)	38.1 (6.6)	73.8 (0.0)	14.3 (0.7)
SVM-1	18.9 (0.7)	10.2 (0.6)	8.2 (0.5)	62.0 (1.1)	13.1 (0.4)

Table 68: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	9.6 (0.8)	13.0 (0.8)	10.3 (0.7)	71.0 (1.1)	10.2 (0.7)
Parzen	32.5 (0.7)	34.1 (0.8)	33.4 (0.7)	63.3 (0.6)	31.0 (0.8)
NM	9.3 (0.8)	12.7 (0.8)	10.2 (0.7)	70.7 (1.0)	11.1 (0.7)
SVM-1	6.5 (0.4)	10.0 (0.5)	8.0 (0.5)	64.6 (1.0)	13.5 (0.4)

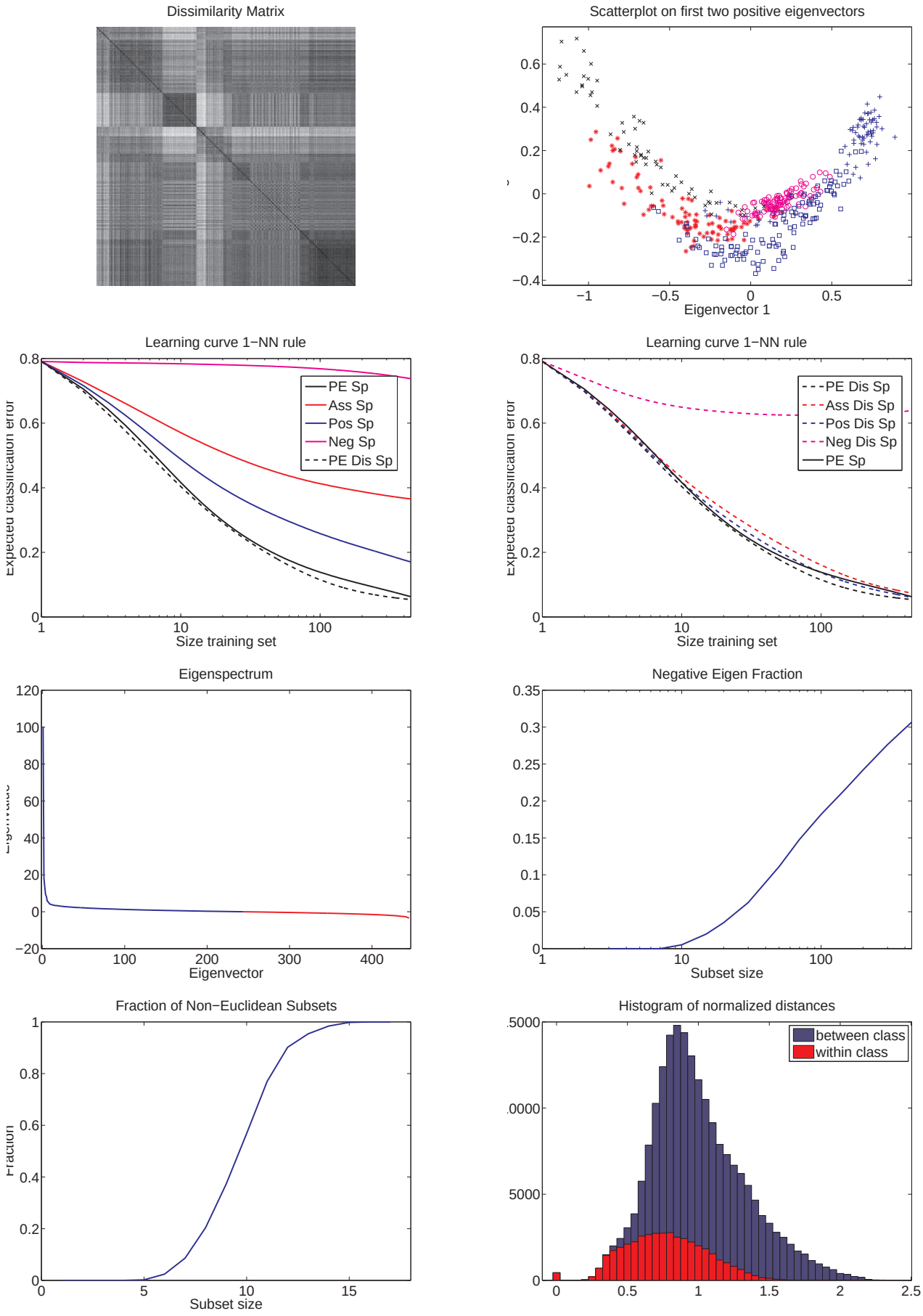


Figure 37: Graphical results for Chickenpieces-20-45.

2.38 Chickenpieces-20-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 20. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.048	asymmetry
446	number of objects
300	number of significant eigenvectors
666	number of triangle inequality violations out of 88120680
240, 205	number of positive and negative eigenvalues
0.298	negative eigenfraction
0.027	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.756, 1.064	average within-class and between-class dissimilarity
6.5, 23.1, 6.1	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 69: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	13.1 (0.7)	36.5 (0.9)	25.0 (1.1)	78.5 (0.7)	13.1 (0.7)
Parzen	25.3 (0.6)	26.3 (0.5)	25.7 (0.5)	88.2 (0.7)	25.3 (0.6)
NM	13.3 (0.5)	53.3 (4.9)	26.3 (1.0)	73.8 (0.0)	13.2 (0.6)
SVM-1	20.2 (0.9)	10.3 (0.4)	8.9 (0.5)	64.1 (1.2)	14.0 (0.5)

Table 70: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	12.3 (0.7)	15.7 (0.6)	14.2 (0.8)	68.9 (1.2)	12.2 (0.8)
Parzen	35.1 (0.6)	34.6 (0.8)	34.4 (0.8)	58.6 (1.0)	33.8 (0.6)
NM	11.7 (0.7)	14.4 (0.6)	13.8 (0.6)	67.3 (1.2)	14.1 (0.7)
SVM-1	7.2 (0.4)	10.5 (0.6)	8.4 (0.4)	59.1 (1.7)	14.2 (0.5)

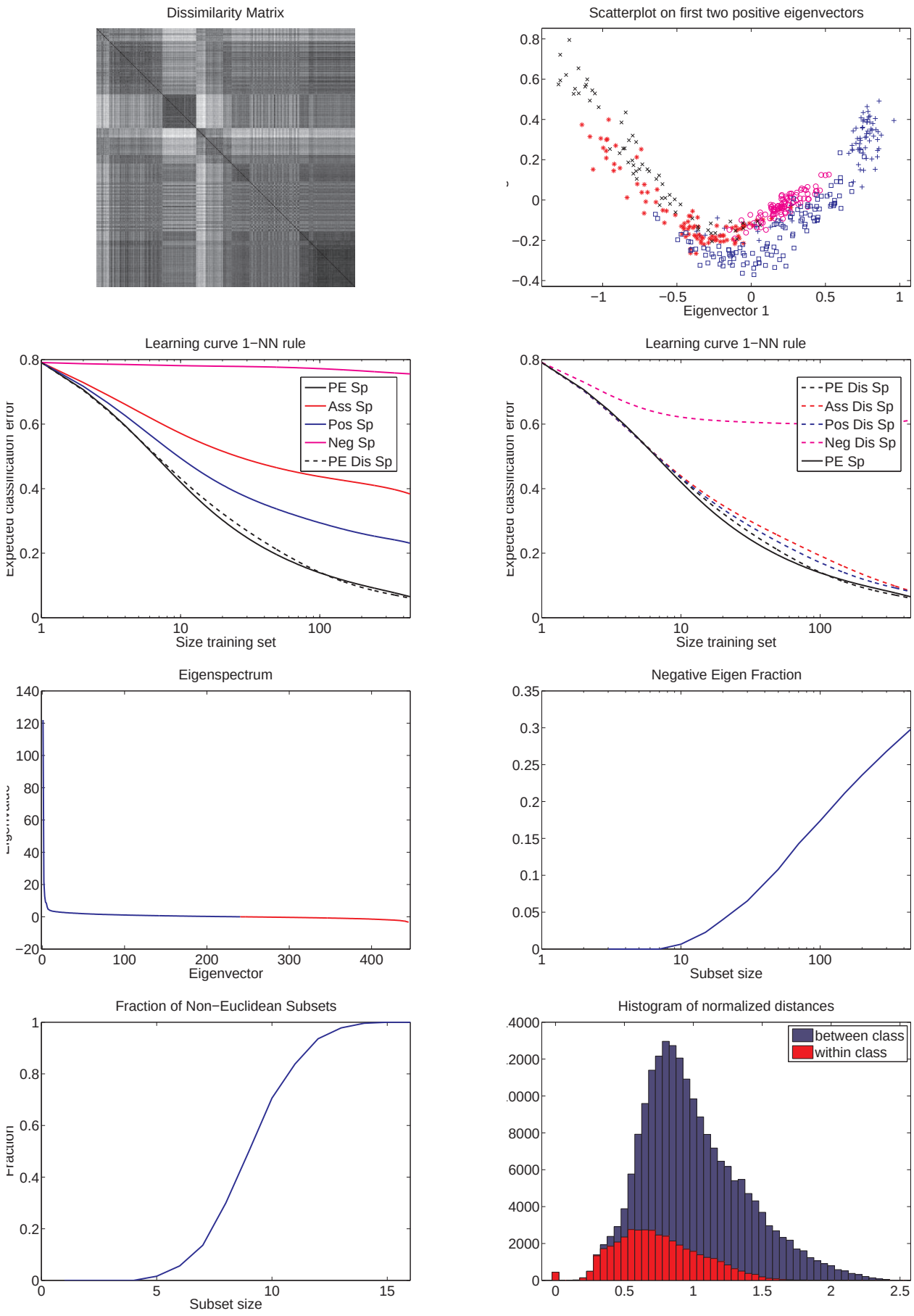


Figure 38: Graphical results for Chickenpieces-20-60.

2.39 Chickenpieces-20-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 20. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.046	asymmetry
446	number of objects
284	number of significant eigenvectors
467	number of triangle inequality violations out of 88120680
239, 206	number of positive and negative eigenvalues
0.269	negative eigenfraction
0.018	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.727, 1.071	average within-class and between-class dissimilarity
7.6, 26.7, 8.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 71: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	17.8 (0.8)	37.1 (1.0)	29.1 (1.3)	76.2 (0.8)	17.8 (0.8)
Parzen	33.0 (0.6)	33.3 (0.8)	33.1 (0.7)	88.6 (0.8)	33.0 (0.6)
NM	18.4 (0.8)	39.6 (1.2)	31.9 (1.2)	73.8 (0.0)	18.1 (0.9)
SVM-1	23.4 (0.5)	11.1 (0.3)	9.5 (0.4)	64.8 (0.9)	16.0 (0.5)

Table 72: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	17.1 (0.6)	19.9 (0.7)	18.6 (0.5)	67.2 (0.6)	17.7 (0.6)
Parzen	40.4 (0.7)	36.6 (0.5)	37.7 (0.6)	59.4 (0.6)	39.7 (0.6)
NM	16.7 (0.4)	19.1 (0.6)	18.1 (0.3)	66.2 (0.6)	20.0 (0.5)
SVM-1	9.1 (0.7)	12.5 (0.5)	10.1 (0.5)	56.9 (0.7)	16.0 (0.6)

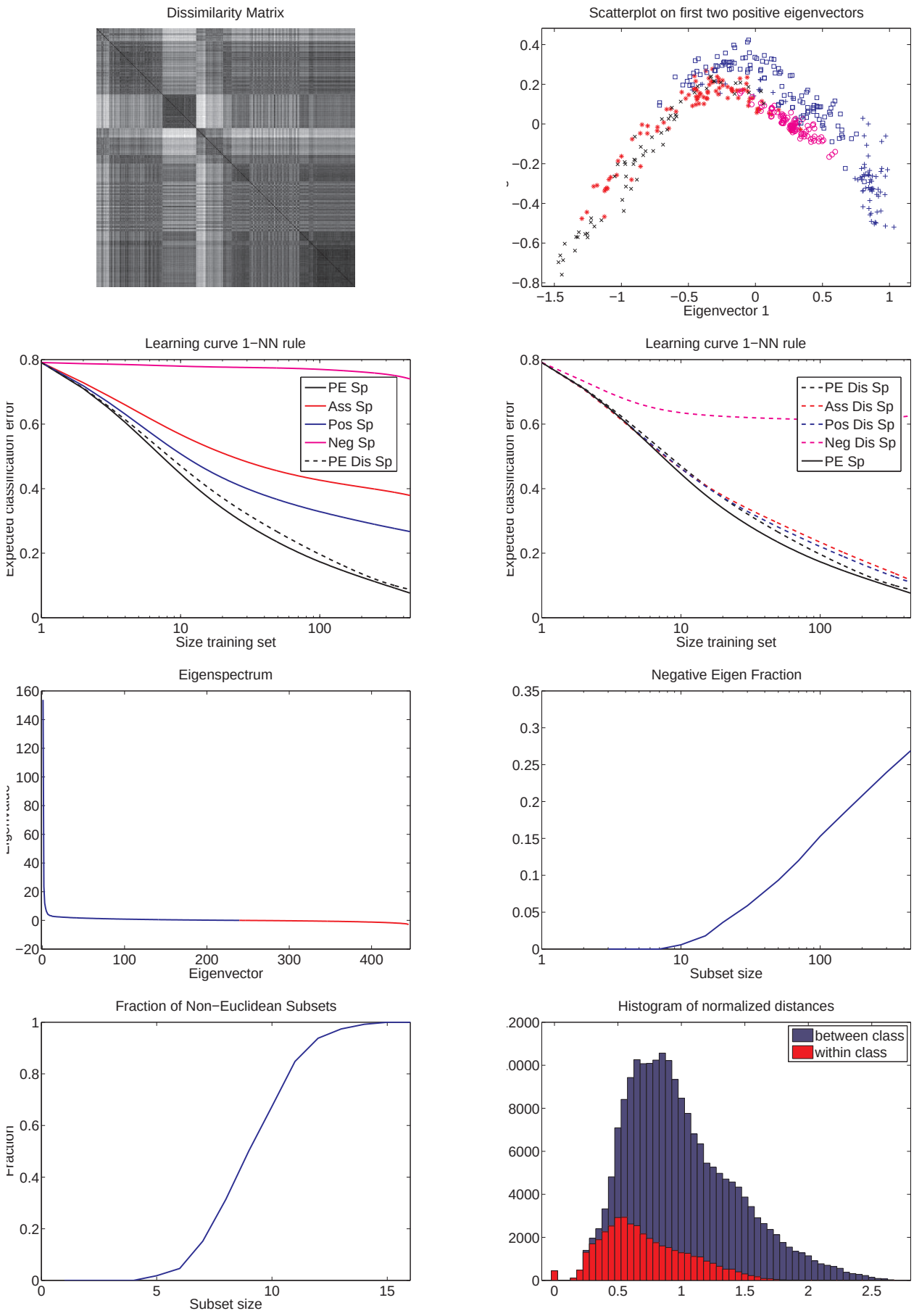


Figure 39: Graphical results for Chickenpieces-20-90.

2.40 Chickenpieces-20-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 20. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.046	asymmetry
446	number of objects
268	number of significant eigenvectors
246	number of triangle inequality violations out of 88120680
239, 206	number of positive and negative eigenvalues
0.234	negative eigenfraction
0.014	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.708, 1.076	average within-class and between-class dissimilarity
10.3, 27.8, 10.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 73: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	21.9 (0.8)	38.0 (0.7)	31.3 (1.1)	75.6 (1.0)	21.9 (0.8)
Parzen	36.9 (0.5)	37.1 (0.6)	36.9 (0.5)	86.6 (0.9)	36.9 (0.5)
NM	23.2 (0.8)	40.7 (1.0)	33.0 (1.1)	73.8 (0.0)	23.5 (0.9)
SVM-1	24.1 (1.0)	12.5 (0.4)	10.9 (0.5)	64.9 (1.0)	19.5 (0.4)

Table 74: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	20.8 (0.6)	23.9 (0.7)	22.5 (0.8)	68.4 (1.1)	20.8 (0.6)
Parzen	42.7 (0.5)	39.4 (0.6)	40.8 (0.7)	62.3 (0.8)	42.8 (0.5)
NM	20.8 (0.5)	23.3 (0.5)	21.5 (0.5)	67.0 (1.0)	26.1 (0.5)
SVM-1	9.6 (0.4)	13.6 (0.5)	10.9 (0.3)	58.4 (0.6)	17.4 (0.5)

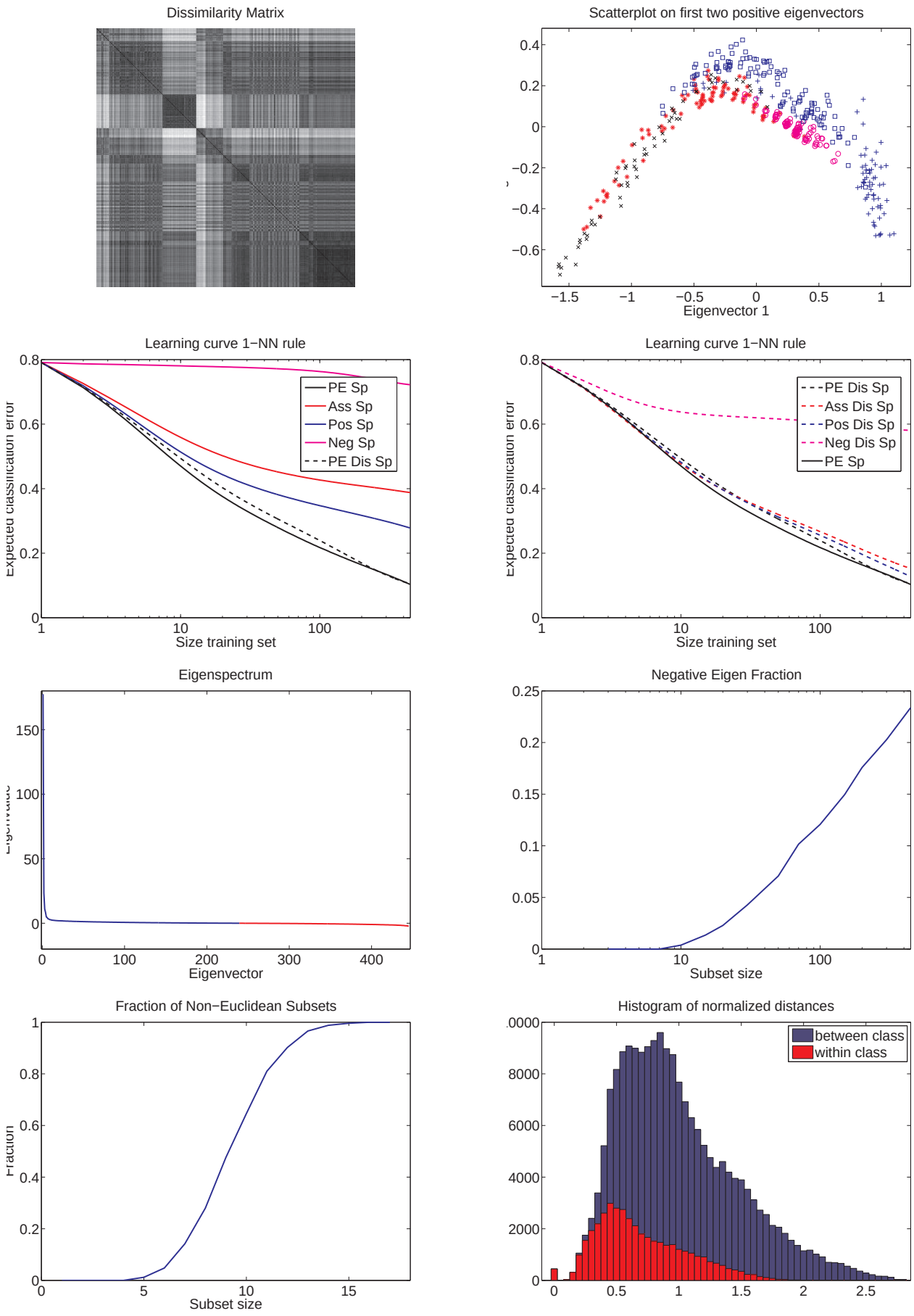


Figure 40: Graphical results for Chickenpieces-20-120.

2.41 Chickenpieces-25-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 25. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.063	asymmetry
446	number of objects
309	number of significant eigenvectors
1375	number of triangle inequality violations out of 88120680
237, 208	number of positive and negative eigenvalues
0.320	negative eigenfraction
0.037	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.780, 1.058	average within-class and between-class dissimilarity
4.3, 14.1, 6.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 75: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	9.9 (0.6)	29.6 (1.1)	17.5 (0.9)	79.6 (1.1)	9.9 (0.6)
Parzen	17.8 (0.6)	19.2 (0.5)	18.5 (0.6)	89.6 (0.7)	17.8 (0.6)
NM	10.9 (0.6)	35.0 (3.2)	17.6 (1.0)	73.8 (0.0)	10.5 (0.5)
SVM-1	16.7 (1.1)	8.2 (0.6)	6.9 (0.8)	62.6 (0.8)	10.8 (0.7)

Table 76: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	8.3 (0.7)	11.7 (0.7)	9.7 (0.8)	72.2 (0.9)	8.6 (0.7)
Parzen	31.3 (0.7)	34.8 (0.8)	33.5 (0.8)	65.5 (0.8)	29.6 (0.8)
NM	7.6 (0.6)	11.5 (0.7)	9.5 (0.8)	70.9 (0.8)	9.7 (0.7)
SVM-1	6.3 (0.6)	9.3 (0.7)	7.4 (0.8)	60.4 (1.1)	12.0 (0.7)

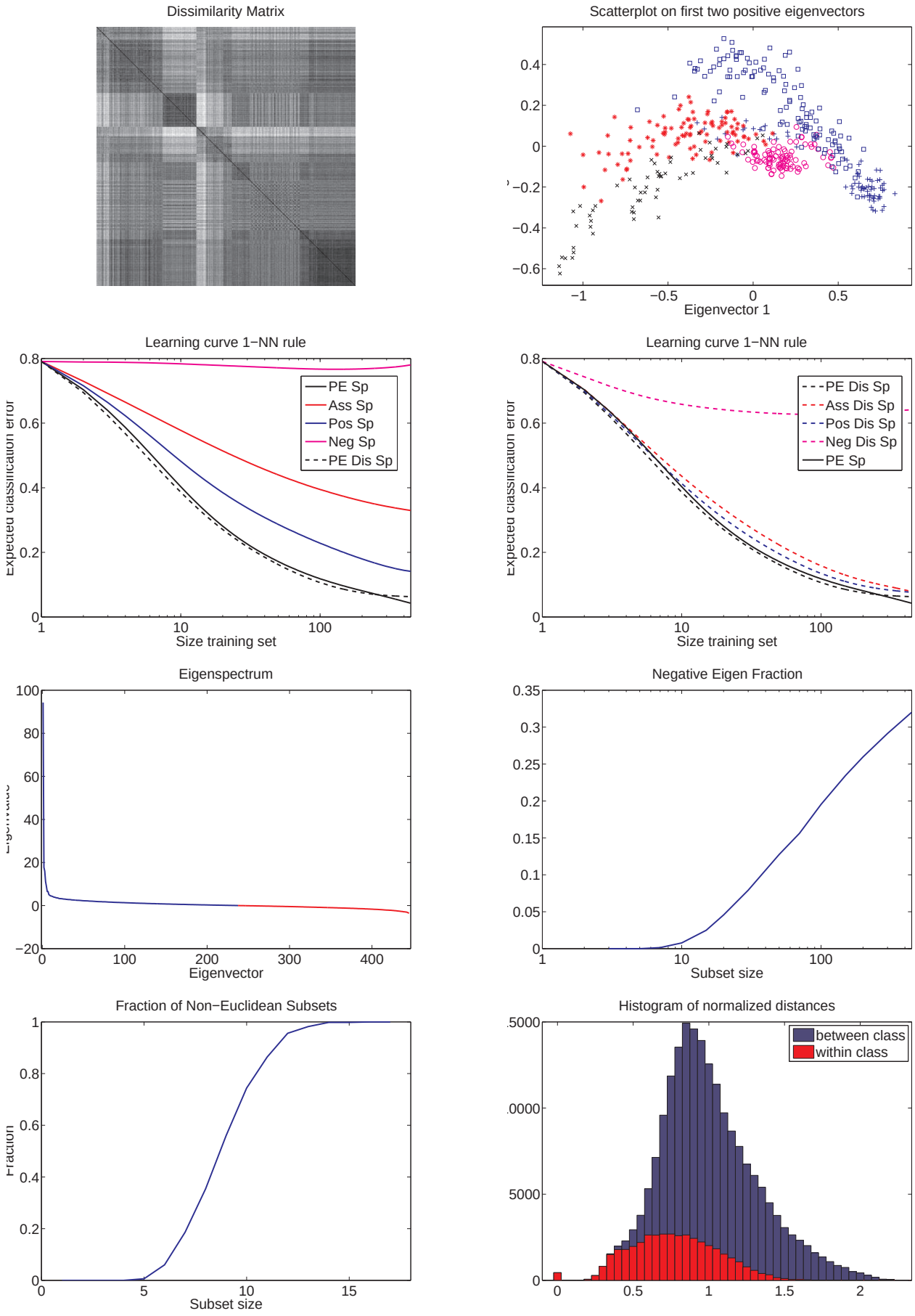


Figure 41: Graphical results for Chickenpieces-25-45.

2.42 Chickenpieces-25-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 25. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.053	asymmetry
446	number of objects
302	number of significant eigenvectors
2793	number of triangle inequality violations out of 88120680
237, 208	number of positive and negative eigenvalues
0.314	negative eigenfraction
0.033	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.753, 1.065	average within-class and between-class dissimilarity
3.8, 17.7, 5.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 77: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	10.8 (0.8)	33.8 (1.1)	21.4 (1.0)	78.0 (1.0)	10.8 (0.8)
Parzen	21.7 (0.5)	22.5 (0.5)	21.9 (0.6)	88.5 (0.6)	21.7 (0.5)
NM	10.9 (0.6)	37.7 (2.8)	28.8 (4.2)	73.8 (0.0)	10.9 (0.8)
SVM-1	17.9 (1.2)	9.6 (0.7)	8.4 (0.9)	62.1 (1.3)	12.7 (0.6)

Table 78: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	9.6 (0.6)	14.7 (0.6)	11.6 (0.7)	70.9 (1.2)	9.4 (0.8)
Parzen	33.6 (0.8)	36.3 (0.7)	35.2 (0.8)	64.0 (0.8)	32.0 (0.7)
NM	9.0 (0.6)	14.4 (0.7)	11.5 (0.8)	69.9 (1.0)	11.3 (0.6)
SVM-1	6.7 (0.6)	11.0 (0.7)	8.2 (0.7)	58.9 (1.1)	13.2 (0.6)

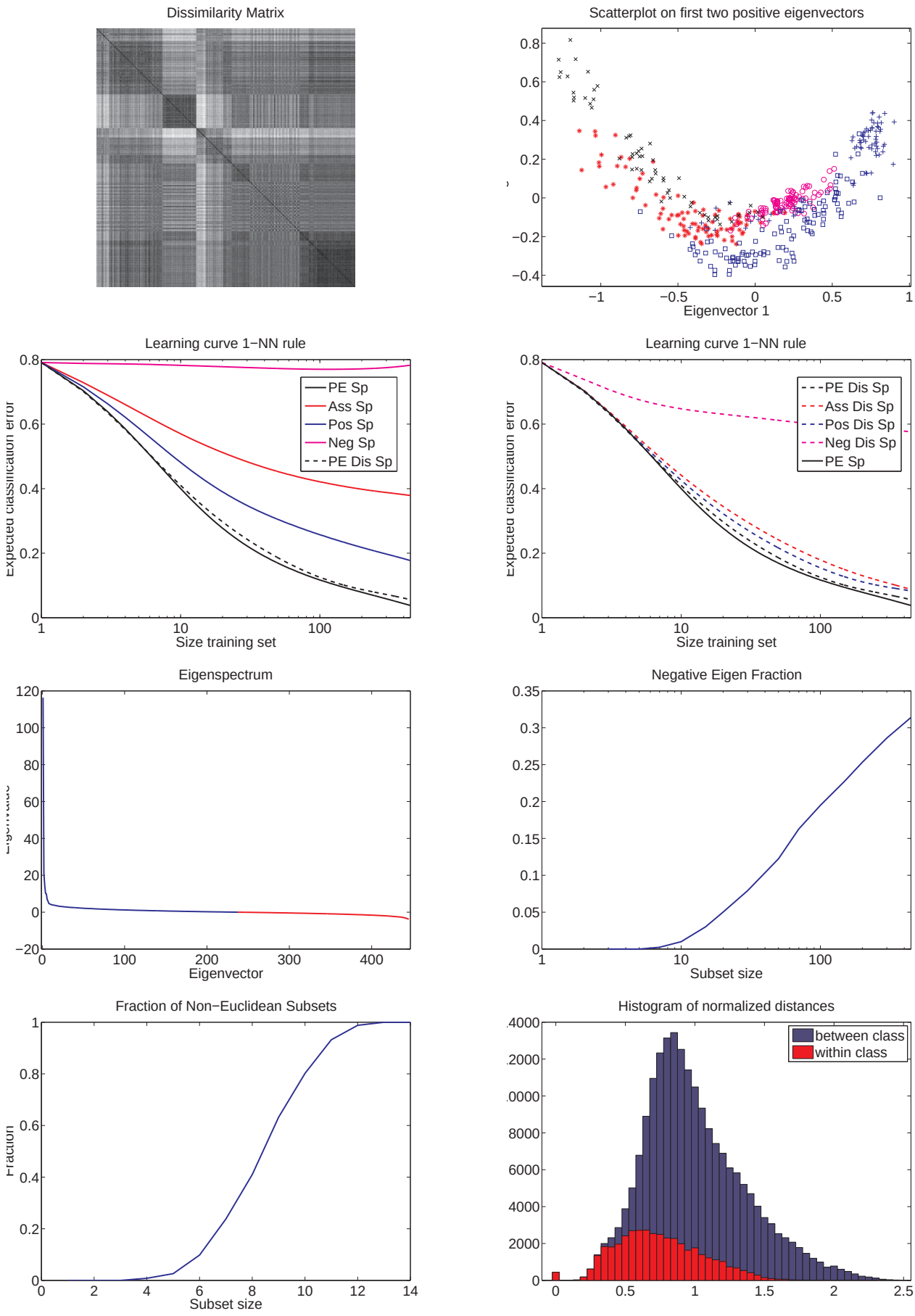


Figure 42: Graphical results for Chickenpieces-25-60.

2.43 Chickenpieces-25-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 25. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.048	asymmetry
446	number of objects
288	number of significant eigenvectors
2354	number of triangle inequality violations out of 88120680
235, 210	number of positive and negative eigenvalues
0.289	negative eigenfraction
0.023	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.724, 1.072	average within-class and between-class dissimilarity
6.5, 22.6, 7.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 79: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	12.7 (0.8)	34.8 (0.8)	24.2 (0.9)	76.7 (0.5)	12.7 (0.8)
Parzen	28.6 (0.7)	29.6 (0.7)	29.2 (0.7)	87.7 (1.0)	28.6 (0.7)
NM	13.5 (0.8)	39.3 (0.9)	27.9 (1.3)	73.8 (0.0)	13.9 (0.8)
SVM-1	20.6 (1.3)	10.7 (0.7)	9.9 (0.8)	65.3 (0.9)	15.2 (0.7)

Table 80: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	12.9 (0.9)	18.3 (0.9)	15.6 (1.1)	65.7 (1.0)	14.3 (0.8)
Parzen	38.7 (0.5)	37.0 (0.7)	37.3 (0.7)	60.2 (0.6)	38.2 (0.6)
NM	12.4 (0.8)	17.7 (1.0)	15.6 (1.1)	64.2 (0.9)	16.3 (1.0)
SVM-1	8.1 (0.8)	11.0 (0.8)	9.3 (0.8)	58.8 (0.8)	14.8 (0.7)

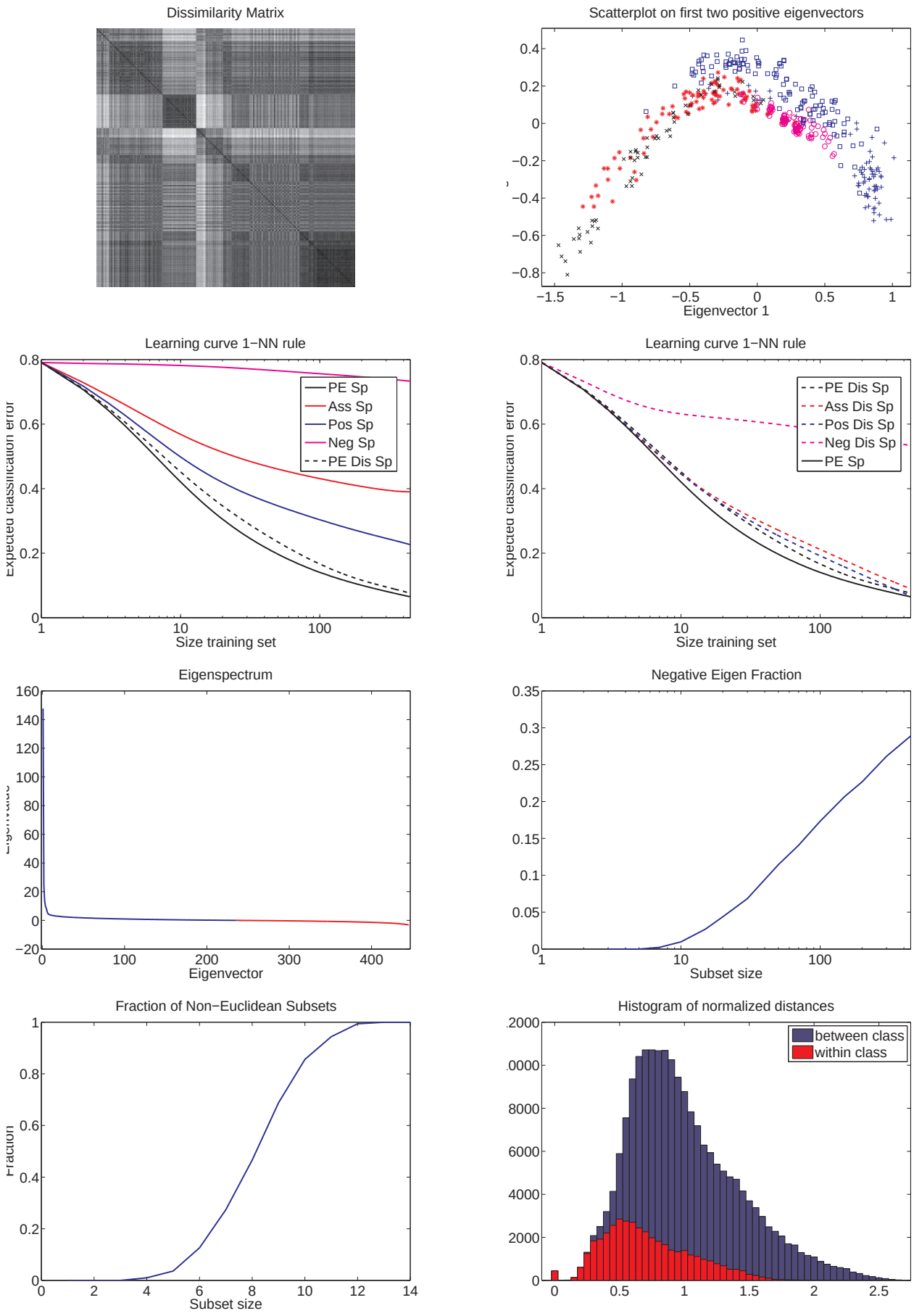


Figure 43: Graphical results for Chickenpieces-25-90.

2.44 Chickenpieces-25-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 25. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.048	asymmetry
446	number of objects
273	number of significant eigenvectors
1146	number of triangle inequality violations out of 88120680
235, 210	number of positive and negative eigenvalues
0.257	negative eigenfraction
0.016	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.706, 1.077	average within-class and between-class dissimilarity
8.7, 23.3, 8.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 81: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	16.9 (0.7)	35.1 (1.2)	26.6 (0.8)	75.9 (0.5)	16.9 (0.7)
Parzen	33.3 (0.7)	33.6 (0.6)	33.5 (0.7)	87.7 (0.9)	33.3 (0.7)
NM	17.9 (0.9)	39.0 (1.1)	30.0 (1.5)	73.8 (0.0)	18.3 (0.7)
SVM-1	21.3 (0.9)	11.3 (0.8)	10.4 (0.6)	65.8 (1.1)	16.8 (0.4)

Table 82: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	17.4 (1.4)	21.3 (1.2)	19.7 (1.4)	65.7 (0.8)	17.6 (1.1)
Parzen	42.1 (0.7)	39.2 (0.6)	40.4 (0.5)	60.9 (0.9)	42.0 (0.6)
NM	17.3 (1.0)	20.3 (1.1)	19.8 (1.0)	64.2 (0.7)	21.5 (1.0)
SVM-1	8.7 (0.7)	12.4 (0.6)	10.1 (0.7)	58.8 (0.8)	16.1 (0.6)

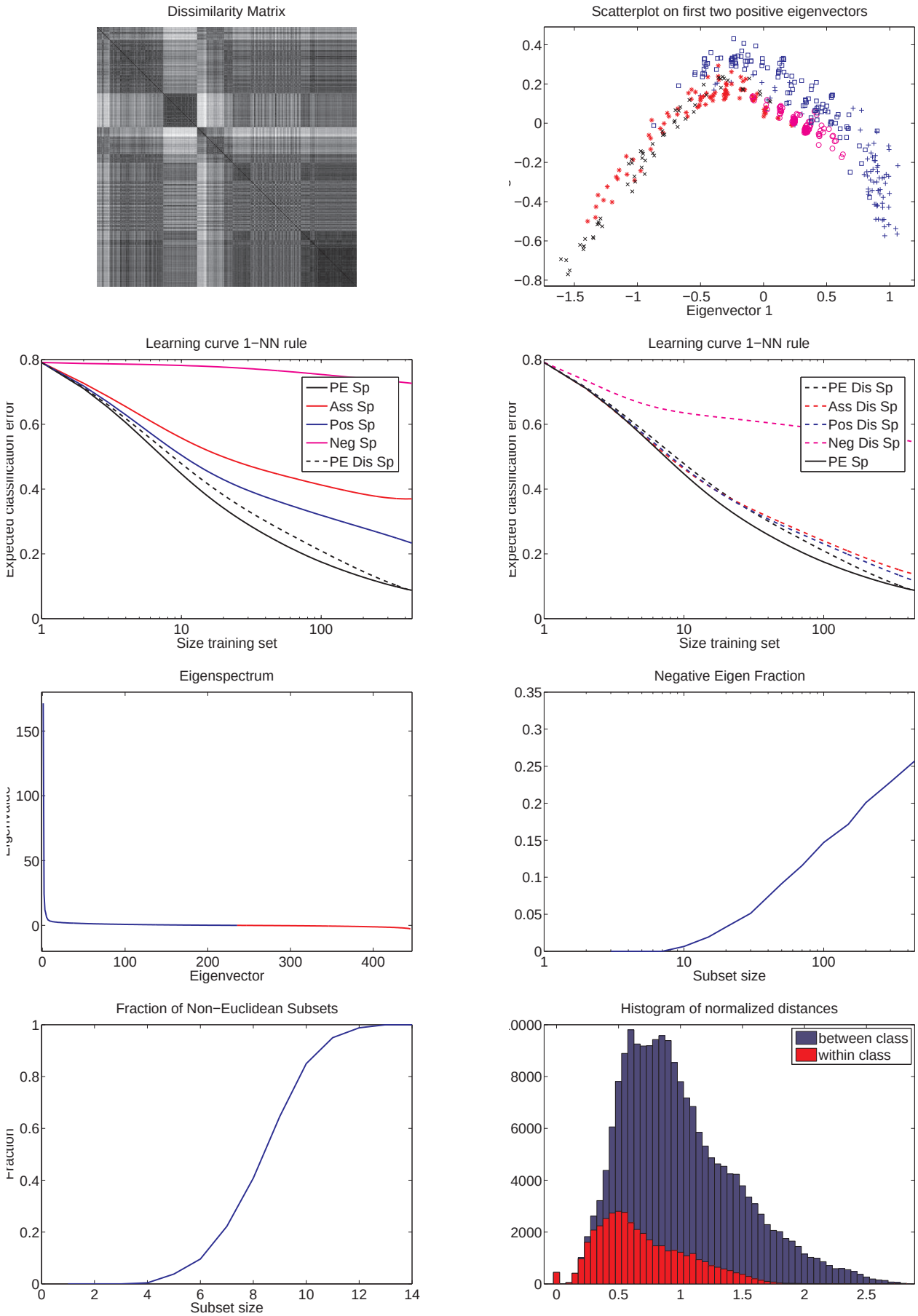


Figure 44: Graphical results for Chickenpieces-25-120.

2.45 Chickenpieces-29-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 29. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.066	asymmetry
446	number of objects
310	number of significant eigenvectors
3360	number of triangle inequality violations out of 88120680
234, 211	number of positive and negative eigenvalues
0.329	negative eigenfraction
0.042	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.777, 1.058	average within-class and between-class dissimilarity
4.0, 14.1, 5.4	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 83: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	9.9 (0.7)	27.4 (1.1)	14.5 (0.6)	79.9 (1.0)	9.9 (0.7)
Parzen	18.3 (0.5)	18.5 (0.4)	18.3 (0.5)	93.5 (0.5)	18.3 (0.5)
NM	9.8 (0.6)	28.0 (1.1)	14.2 (0.6)	73.8 (0.0)	10.2 (0.6)
SVM-1	18.2 (0.8)	8.5 (0.6)	7.0 (0.6)	58.3 (1.2)	10.7 (0.5)

Table 84: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	8.2 (0.6)	10.7 (0.8)	9.1 (0.7)	71.0 (0.8)	8.3 (0.6)
Parzen	30.5 (0.6)	33.2 (0.8)	32.0 (0.6)	66.6 (0.7)	28.4 (0.6)
NM	7.4 (0.6)	10.5 (0.6)	8.7 (0.6)	70.1 (0.9)	9.1 (0.7)
SVM-1	6.1 (0.6)	9.2 (0.4)	7.1 (0.5)	62.0 (1.0)	12.8 (0.6)

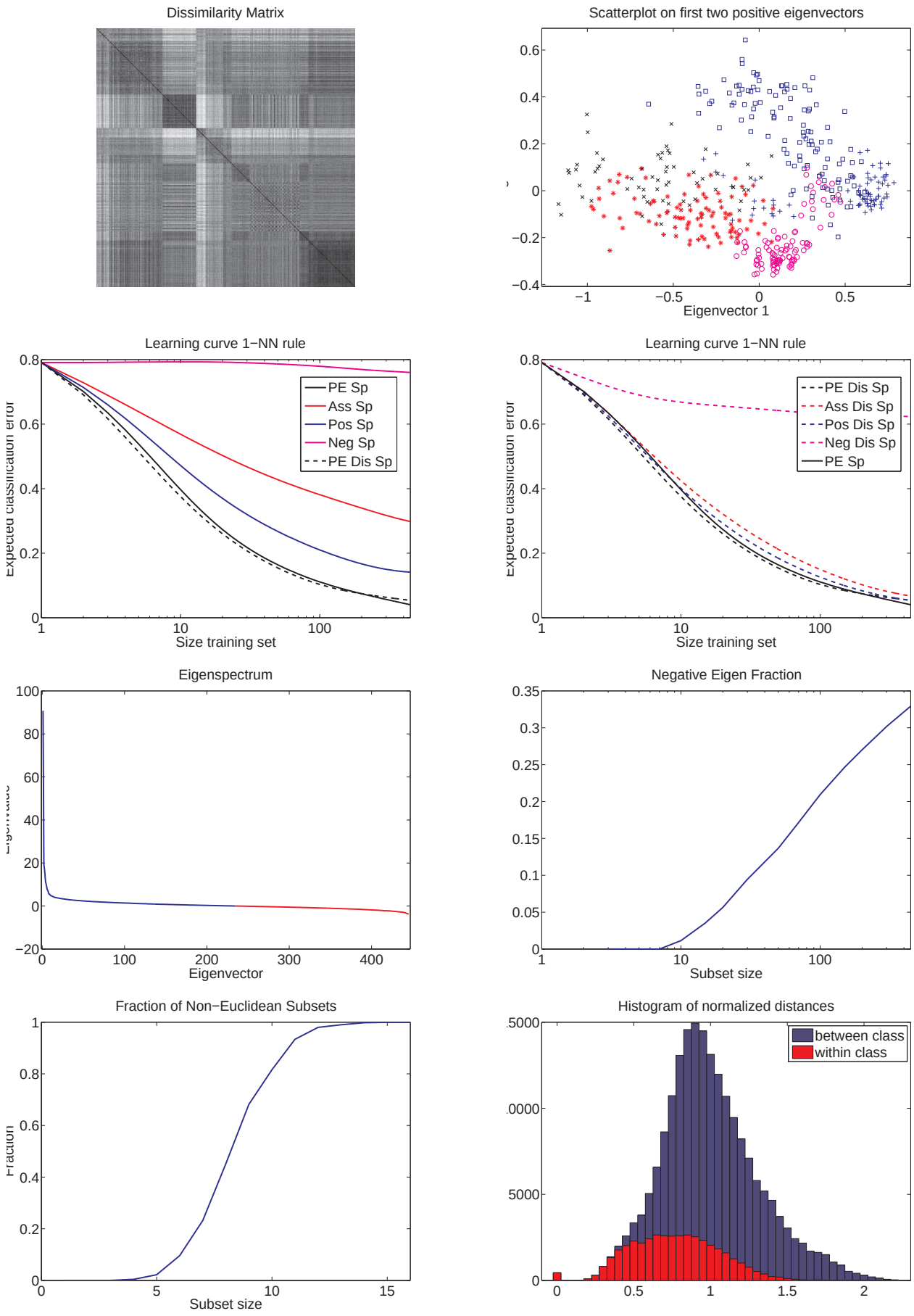


Figure 45: Graphical results for Chickenpieces-29-45.

2.46 Chickenpieces-29-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 29. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.056	asymmetry
446	number of objects
305	number of significant eigenvectors
5221	number of triangle inequality violations out of 88120680
235, 210	number of positive and negative eigenvalues
0.324	negative eigenfraction
0.034	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.749, 1.066	average within-class and between-class dissimilarity
3.4, 15.7, 5.4	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 85: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	9.8 (0.7)	30.5 (1.3)	17.8 (1.0)	79.8 (1.1)	9.8 (0.7)
Parzen	20.9 (0.6)	22.2 (0.6)	21.2 (0.7)	91.3 (0.4)	20.9 (0.6)
NM	10.1 (0.7)	32.0 (1.1)	21.5 (3.3)	73.8 (0.0)	9.9 (0.6)
SVM-1	17.3 (0.8)	9.3 (0.7)	7.2 (0.7)	60.8 (0.7)	13.0 (0.4)

Table 86: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	8.9 (0.7)	11.8 (0.7)	9.6 (0.7)	71.7 (1.4)	9.5 (0.6)
Parzen	32.1 (0.6)	34.2 (0.7)	33.2 (0.6)	66.9 (0.4)	30.2 (0.7)
NM	8.4 (0.8)	11.8 (0.7)	9.5 (0.7)	70.7 (1.4)	11.5 (0.7)
SVM-1	6.5 (0.6)	9.9 (0.6)	7.7 (0.6)	60.3 (1.3)	13.2 (0.6)

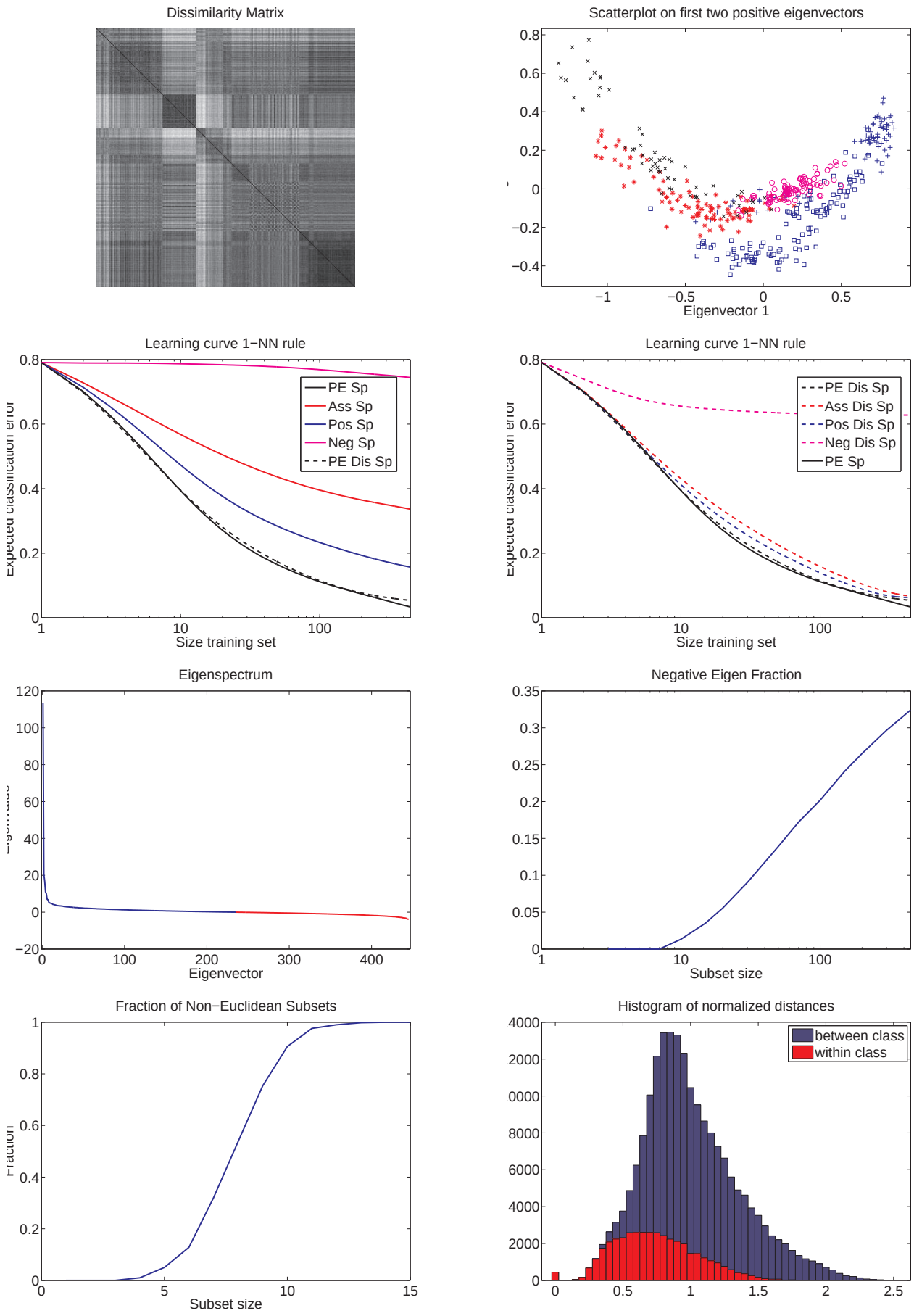


Figure 46: Graphical results for Chickenpieces-29-60.

2.47 Chickenpieces-29-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 29. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.050	asymmetry
446	number of objects
291	number of significant eigenvectors
5634	number of triangle inequality violations out of 88120680
234, 211	number of positive and negative eigenvalues
0.302	negative eigenfraction
0.025	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.718, 1.074	average within-class and between-class dissimilarity
4.7, 18.2, 5.4	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 87: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	11.0 (0.4)	34.6 (1.3)	22.6 (1.2)	76.9 (0.8)	11.0 (0.4)
Parzen	26.3 (0.6)	26.6 (0.6)	26.4 (0.7)	87.4 (0.6)	26.3 (0.6)
NM	11.7 (0.7)	37.8 (2.6)	23.8 (1.4)	73.8 (0.0)	11.5 (0.7)
SVM-1	19.4 (1.1)	10.4 (0.8)	8.4 (1.0)	64.1 (0.9)	14.9 (0.5)

Table 88: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	10.8 (0.8)	14.5 (0.8)	12.9 (0.9)	67.1 (0.8)	11.6 (0.8)
Parzen	36.2 (0.6)	35.3 (0.7)	35.2 (0.5)	63.2 (0.6)	35.2 (0.7)
NM	11.9 (0.7)	14.6 (0.9)	13.7 (0.8)	65.4 (1.0)	15.4 (0.6)
SVM-1	7.2 (0.7)	11.5 (0.7)	8.7 (0.7)	58.9 (1.5)	14.8 (0.4)

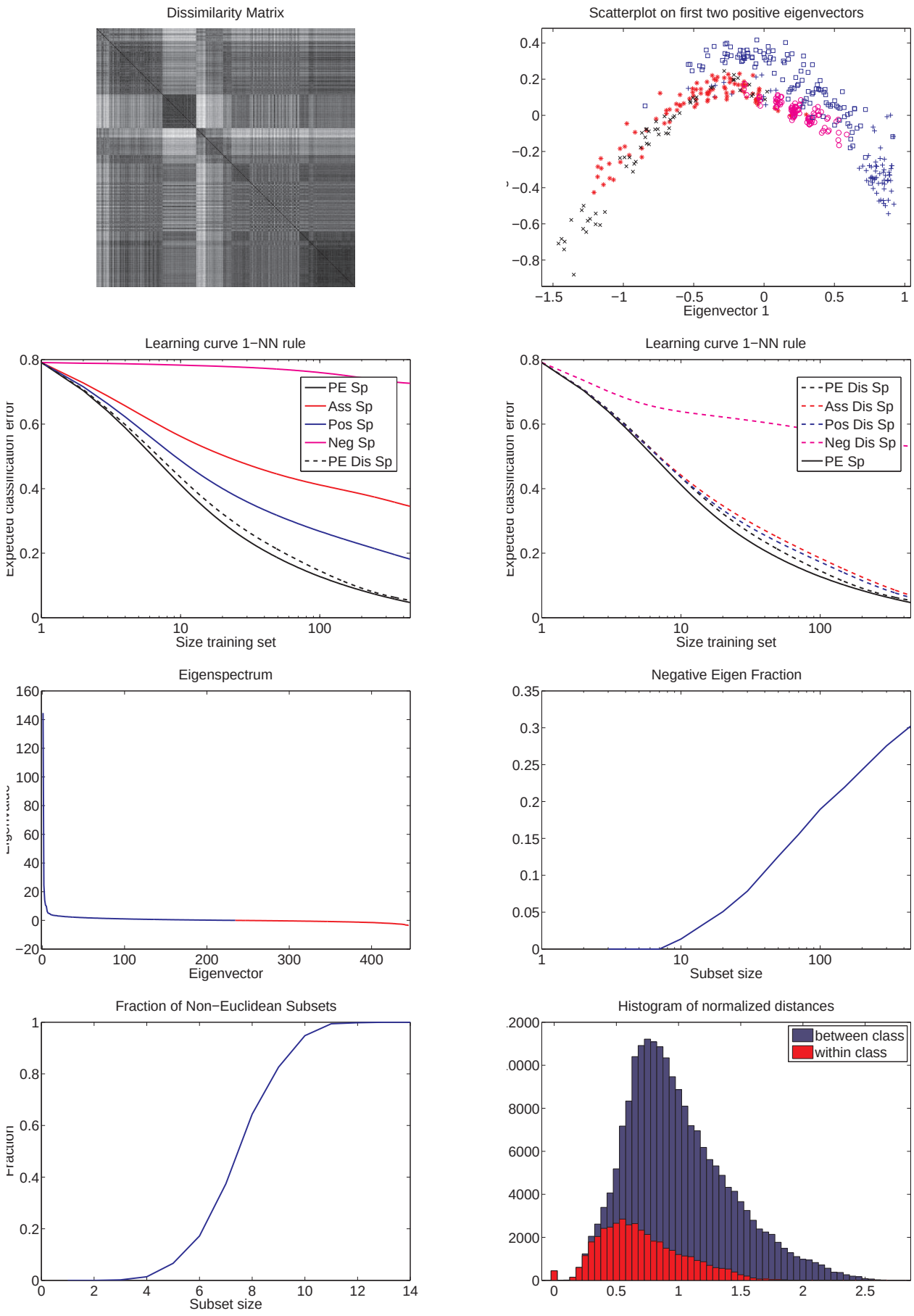


Figure 47: Graphical results for Chickenpieces-29-90.

2.48 Chickenpieces-29-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 29. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.050	asymmetry
446	number of objects
278	number of significant eigenvectors
3419	number of triangle inequality violations out of 88120680
231, 214	number of positive and negative eigenvalues
0.272	negative eigenfraction
0.018	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.700, 1.078	average within-class and between-class dissimilarity
6.3, 20.4, 6.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 89: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	14.0 (0.5)	34.6 (1.2)	24.8 (1.0)	75.5 (0.6)	14.0 (0.5)
Parzen	30.5 (0.7)	31.2 (0.7)	30.9 (0.6)	87.0 (0.7)	30.5 (0.7)
NM	14.4 (0.6)	35.4 (1.4)	27.0 (1.1)	73.8 (0.0)	14.2 (0.5)
SVM-1	20.7 (0.6)	11.8 (0.8)	10.3 (0.8)	64.7 (0.4)	16.5 (0.3)

Table 90: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	14.7 (0.8)	18.0 (0.9)	16.0 (0.8)	62.4 (0.4)	15.3 (0.9)
Parzen	40.0 (0.7)	36.6 (0.9)	37.5 (0.7)	62.3 (0.8)	40.1 (1.0)
NM	15.4 (0.8)	18.3 (0.9)	17.5 (0.8)	62.1 (0.6)	20.2 (0.9)
SVM-1	8.4 (0.7)	12.5 (0.6)	9.9 (0.7)	60.0 (1.6)	15.8 (0.4)

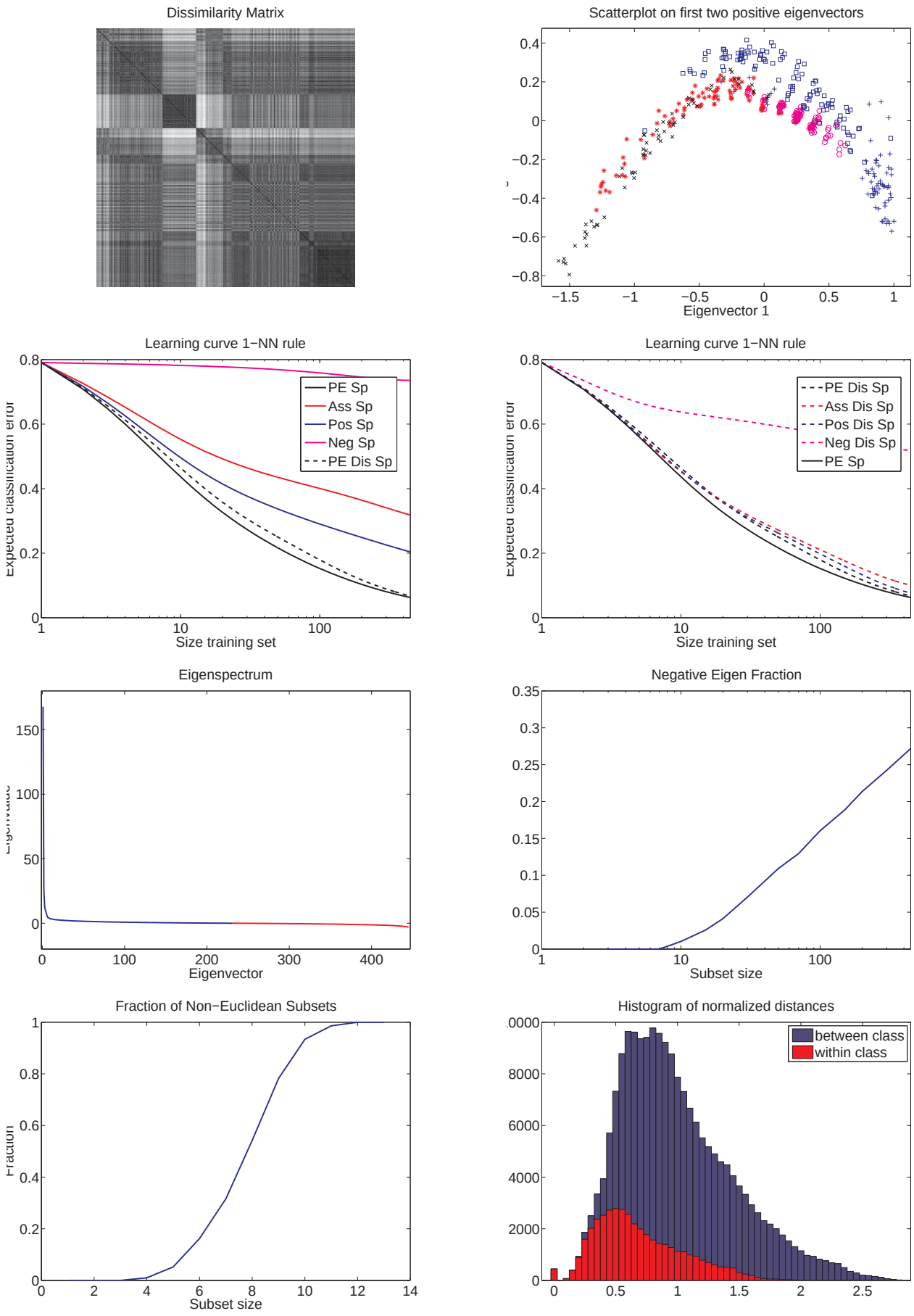


Figure 48: Graphical results for Chickenpieces-29-120.

2.49 Chickenpieces-30-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 30. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.067	asymmetry
446	number of objects
312	number of significant eigenvectors
3188	number of triangle inequality violations out of 88120680
234, 211	number of positive and negative eigenvalues
0.331	negative eigenfraction
0.043	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.775, 1.059	average within-class and between-class dissimilarity
4.5, 13.0, 5.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 91: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	10.7 (0.4)	28.9 (1.0)	16.6 (0.6)	81.9 (0.6)	10.7 (0.4)
Parzen	17.4 (0.6)	18.5 (0.6)	17.9 (0.5)	91.3 (0.7)	17.4 (0.6)
NM	10.3 (0.4)	28.7 (1.0)	16.1 (0.6)	73.8 (0.0)	10.7 (0.5)
SVM-1	20.2 (0.7)	9.6 (0.6)	9.6 (0.4)	61.7 (1.1)	11.6 (0.6)

Table 92: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	7.5 (0.9)	10.4 (0.7)	8.5 (0.6)	72.7 (0.8)	8.2 (0.6)
Parzen	29.8 (0.8)	34.5 (1.0)	32.5 (0.9)	64.7 (0.7)	27.7 (0.9)
NM	6.8 (0.7)	9.9 (0.6)	8.0 (0.6)	71.9 (1.0)	8.7 (0.5)
SVM-1	6.9 (0.6)	9.9 (0.5)	7.9 (0.6)	63.8 (0.9)	12.7 (0.5)

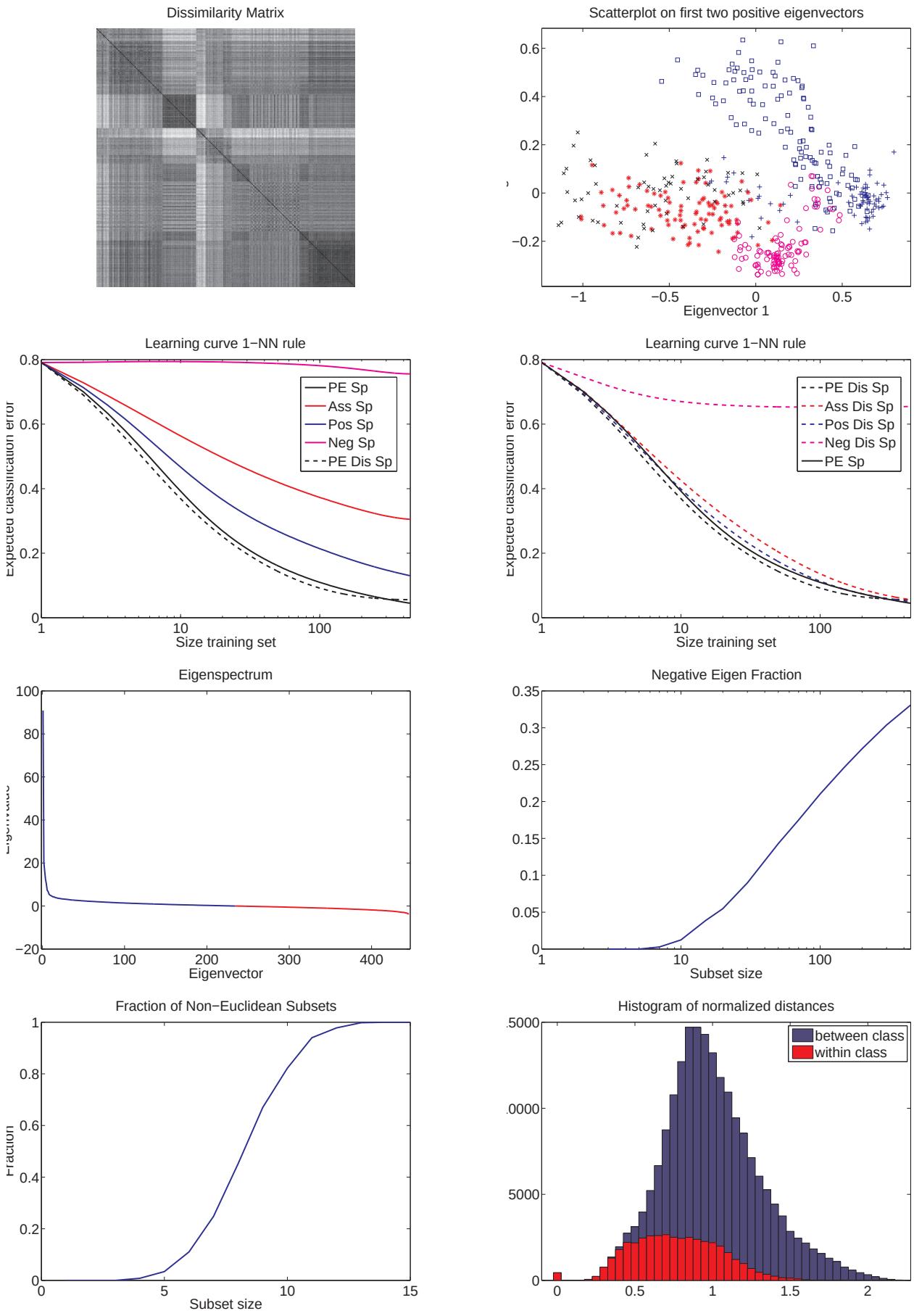


Figure 49: Graphical results for Chickenpieces-30-45.

2.50 Chickenpieces-30-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 30. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.056	asymmetry
446	number of objects
305	number of significant eigenvectors
5722	number of triangle inequality violations out of 88120680
235, 210	number of positive and negative eigenvalues
0.326	negative eigenfraction
0.035	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.746, 1.066	average within-class and between-class dissimilarity
4.9, 13.7, 4.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 93: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	10.1 (0.6)	31.6 (1.6)	16.4 (1.1)	79.1 (0.6)	10.1 (0.6)
Parzen	20.1 (0.5)	20.8 (0.6)	20.5 (0.6)	90.3 (0.8)	20.1 (0.5)
NM	9.5 (0.5)	32.1 (1.3)	16.1 (0.9)	73.8 (0.0)	9.4 (0.4)
SVM-1	18.4 (0.9)	9.6 (0.6)	8.3 (0.5)	63.3 (0.8)	13.0 (0.7)

Table 94: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	8.9 (0.7)	13.1 (0.6)	10.1 (0.6)	69.2 (1.3)	8.7 (0.8)
Parzen	30.9 (0.7)	35.0 (1.0)	33.4 (0.9)	65.2 (0.7)	29.5 (0.9)
NM	7.8 (0.7)	11.9 (0.7)	9.7 (0.6)	68.9 (1.5)	9.9 (0.9)
SVM-1	6.9 (0.5)	10.0 (0.5)	8.1 (0.5)	60.4 (1.1)	13.3 (0.6)

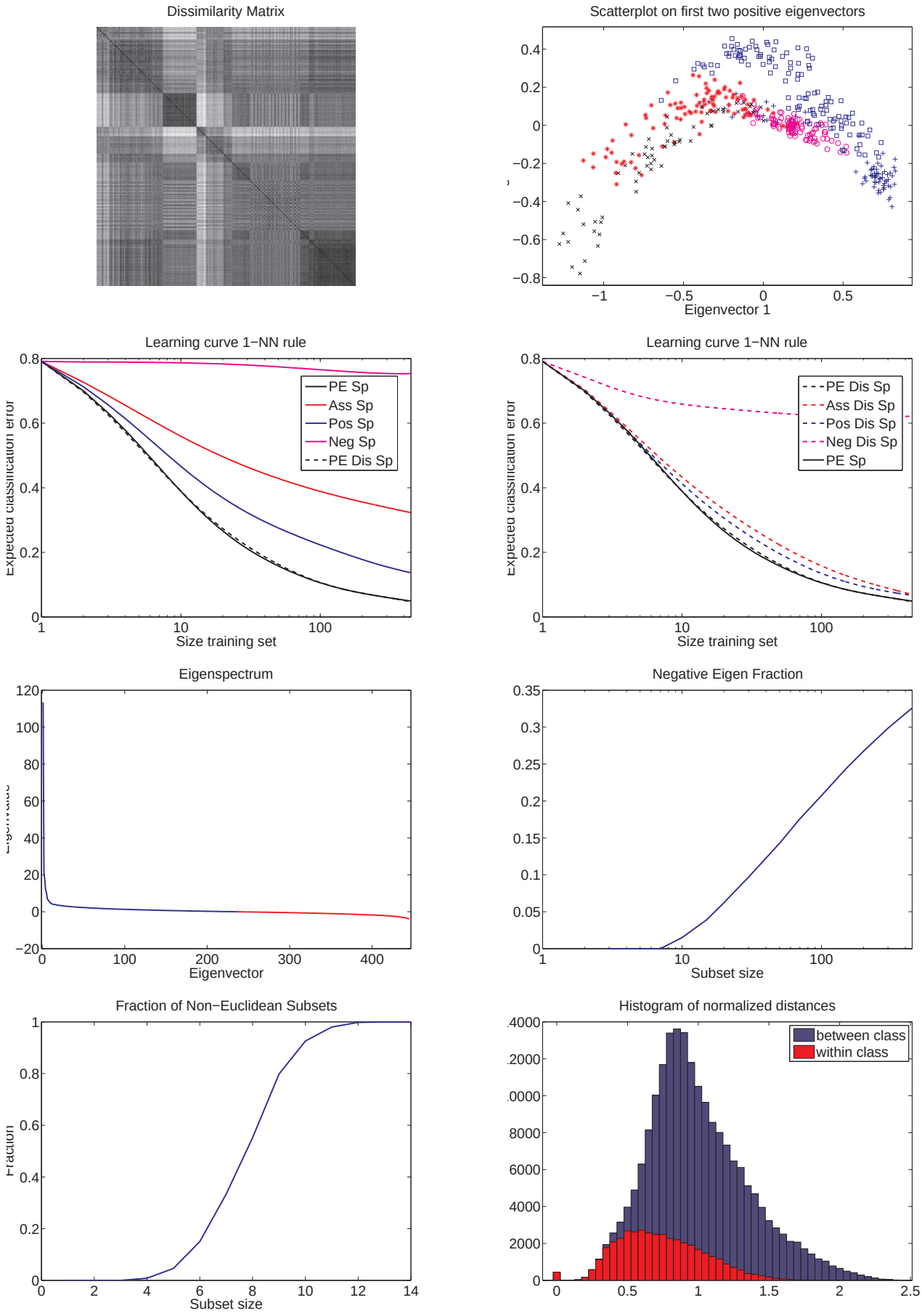


Figure 50: Graphical results for Chickenpieces-30-60.

2.51 Chickenpieces-30-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 30. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.050	asymmetry
446	number of objects
292	number of significant eigenvectors
5918	number of triangle inequality violations out of 88120680
233, 212	number of positive and negative eigenvalues
0.304	negative eigenfraction
0.025	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.716, 1.074	average within-class and between-class dissimilarity
5.8, 24.4, 6.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 95: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	12.1 (0.7)	36.5 (1.2)	23.4 (1.1)	77.4 (0.8)	12.1 (0.7)
Parzen	26.1 (0.5)	27.1 (0.7)	26.6 (0.5)	88.7 (0.5)	26.1 (0.5)
NM	10.8 (0.7)	37.2 (1.3)	23.3 (1.1)	73.8 (0.0)	10.8 (0.8)
SVM-1	20.1 (1.0)	10.4 (1.0)	9.2 (0.8)	63.5 (1.1)	15.8 (0.6)

Table 96: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	12.0 (0.8)	15.3 (0.7)	13.1 (0.7)	67.1 (1.1)	12.8 (0.9)
Parzen	35.6 (0.8)	36.4 (1.0)	36.3 (0.8)	64.2 (1.0)	35.1 (0.8)
NM	11.6 (0.5)	15.2 (0.7)	13.2 (0.6)	66.0 (0.9)	14.5 (0.7)
SVM-1	7.5 (0.5)	11.2 (0.7)	8.8 (0.6)	59.3 (1.0)	14.8 (0.7)

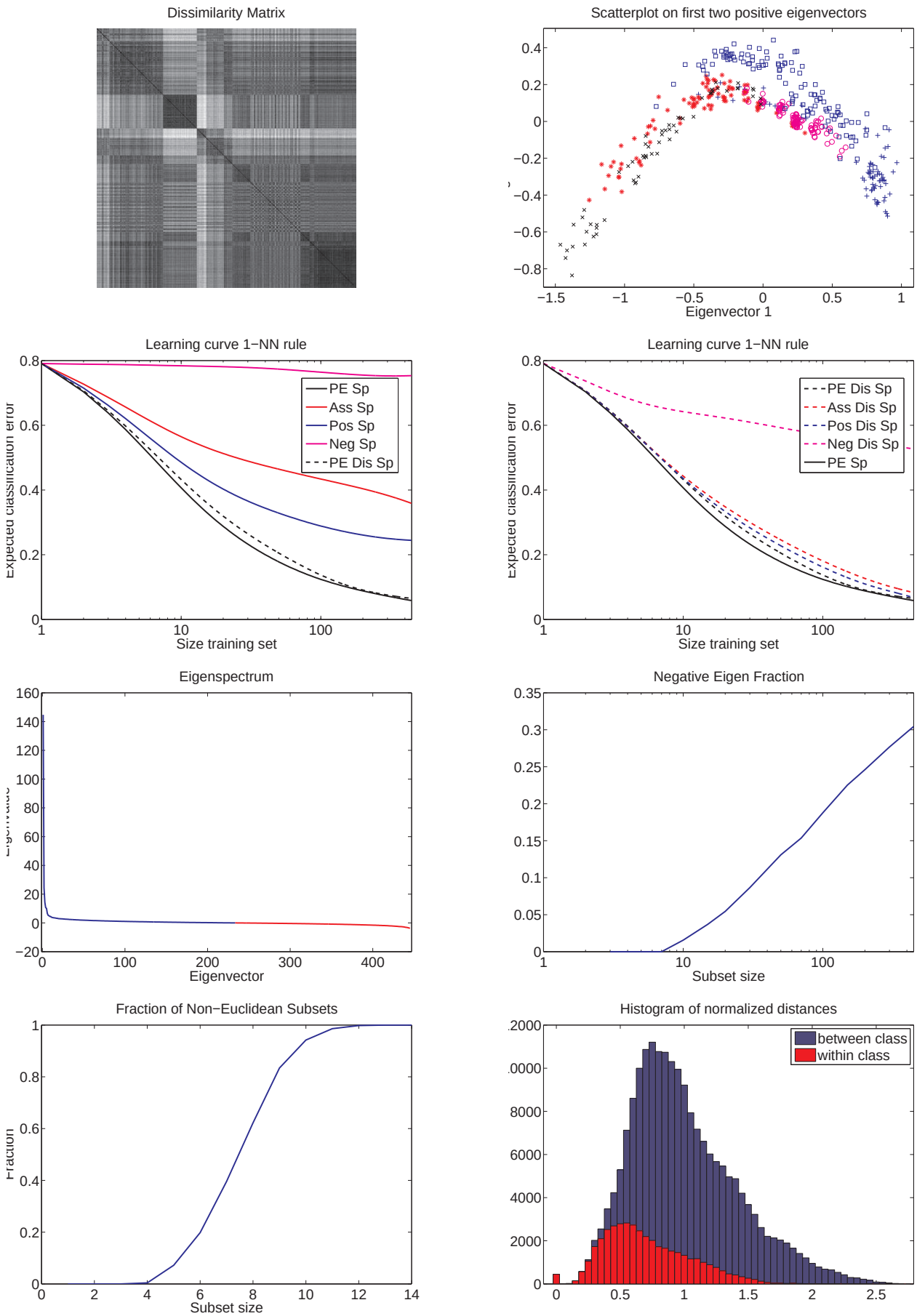


Figure 51: Graphical results for Chickenpieces-30-90.

2.52 Chickenpieces-30-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 30. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.050	asymmetry
446	number of objects
279	number of significant eigenvectors
3510	number of triangle inequality violations out of 88120680
233, 212	number of positive and negative eigenvalues
0.273	negative eigenfraction
0.019	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.698, 1.079	average within-class and between-class dissimilarity
6.7, 21.7, 8.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 97: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	13.3 (0.9)	34.6 (1.3)	25.1 (1.0)	75.8 (0.7)	13.3 (0.9)
Parzen	31.0 (0.5)	31.2 (0.6)	31.4 (0.6)	88.3 (0.7)	31.0 (0.5)
NM	13.7 (0.8)	36.5 (1.4)	25.8 (1.0)	73.8 (0.0)	12.7 (0.8)
SVM-1	19.9 (1.0)	9.9 (0.7)	8.9 (0.7)	66.0 (1.0)	16.5 (0.6)

Table 98: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	14.5 (1.1)	17.8 (0.7)	15.9 (0.9)	63.7 (1.0)	15.2 (0.9)
Parzen	39.6 (0.7)	38.1 (0.7)	38.4 (0.8)	64.9 (0.9)	40.2 (0.8)
NM	15.6 (0.7)	17.0 (0.8)	16.7 (0.6)	62.8 (0.9)	19.7 (0.9)
SVM-1	7.8 (0.6)	11.1 (0.6)	9.0 (0.6)	60.1 (1.0)	15.6 (0.7)

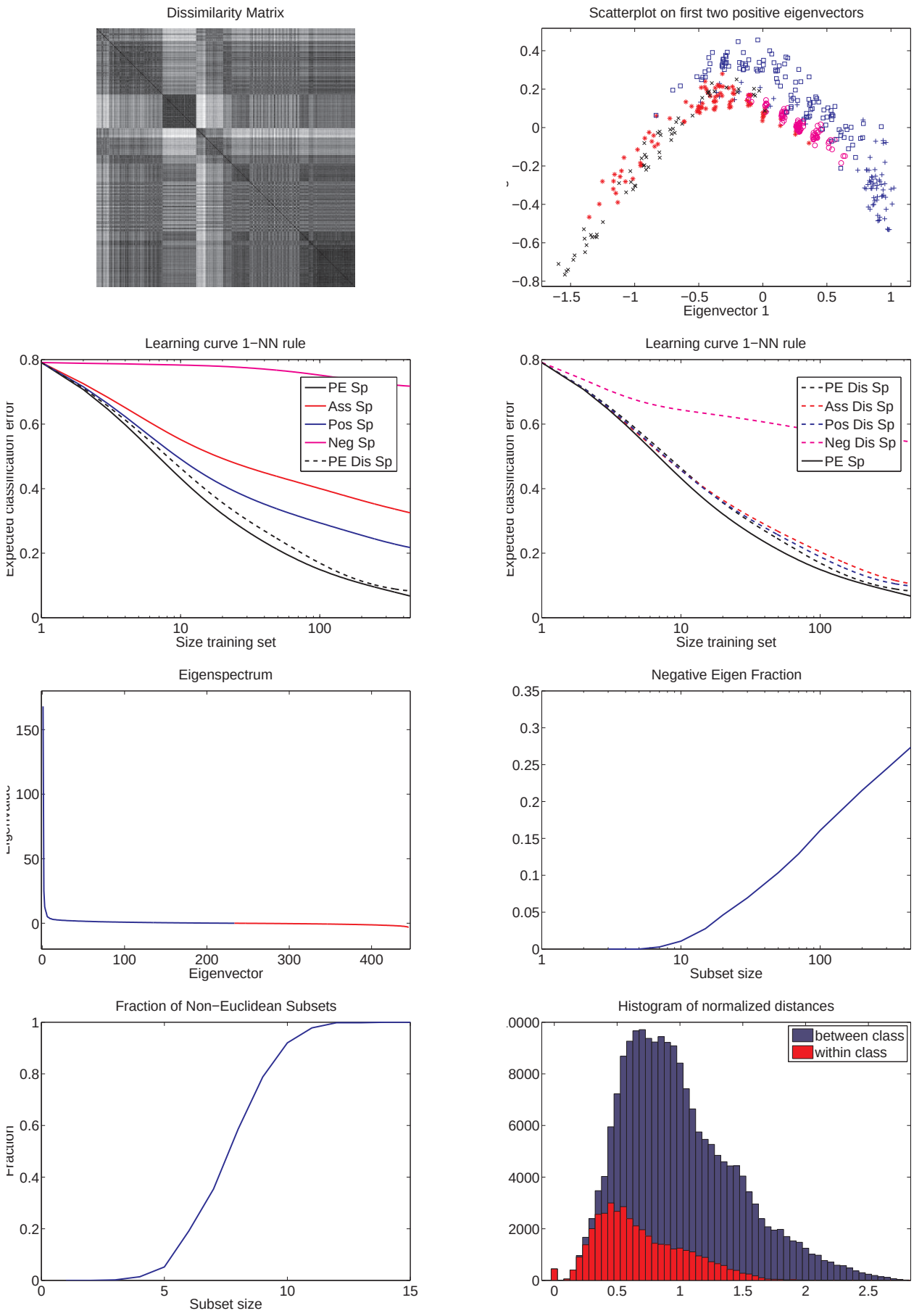


Figure 52: Graphical results for Chickenpieces-30-120.

2.53 Chickenpieces-31-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 31. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.068	asymmetry
446	number of objects
312	number of significant eigenvectors
3832	number of triangle inequality violations out of 88120680
235, 210	number of positive and negative eigenvalues
0.333	negative eigenfraction
0.043	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.775, 1.059	average within-class and between-class dissimilarity
6.5, 13.7, 5.8	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 99: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	10.3 (0.4)	31.8 (1.1)	18.9 (1.2)	82.1 (0.8)	10.3 (0.4)
Parzen	16.9 (0.5)	18.4 (0.7)	17.5 (0.5)	92.4 (0.6)	16.9 (0.5)
NM	10.3 (0.5)	31.7 (1.4)	18.4 (1.2)	73.8 (0.0)	10.4 (0.6)
SVM-1	18.9 (0.5)	9.7 (0.8)	7.8 (0.7)	60.6 (0.9)	11.0 (0.6)

Table 100: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	9.0 (0.6)	11.6 (1.0)	10.0 (0.7)	73.6 (1.1)	8.7 (0.7)
Parzen	30.7 (1.0)	33.8 (0.9)	32.7 (0.8)	64.7 (0.7)	28.9 (1.0)
NM	8.6 (0.6)	11.0 (0.8)	9.5 (0.8)	72.9 (1.3)	9.1 (0.6)
SVM-1	6.9 (0.5)	9.6 (0.5)	7.9 (0.4)	62.6 (1.2)	11.7 (0.7)

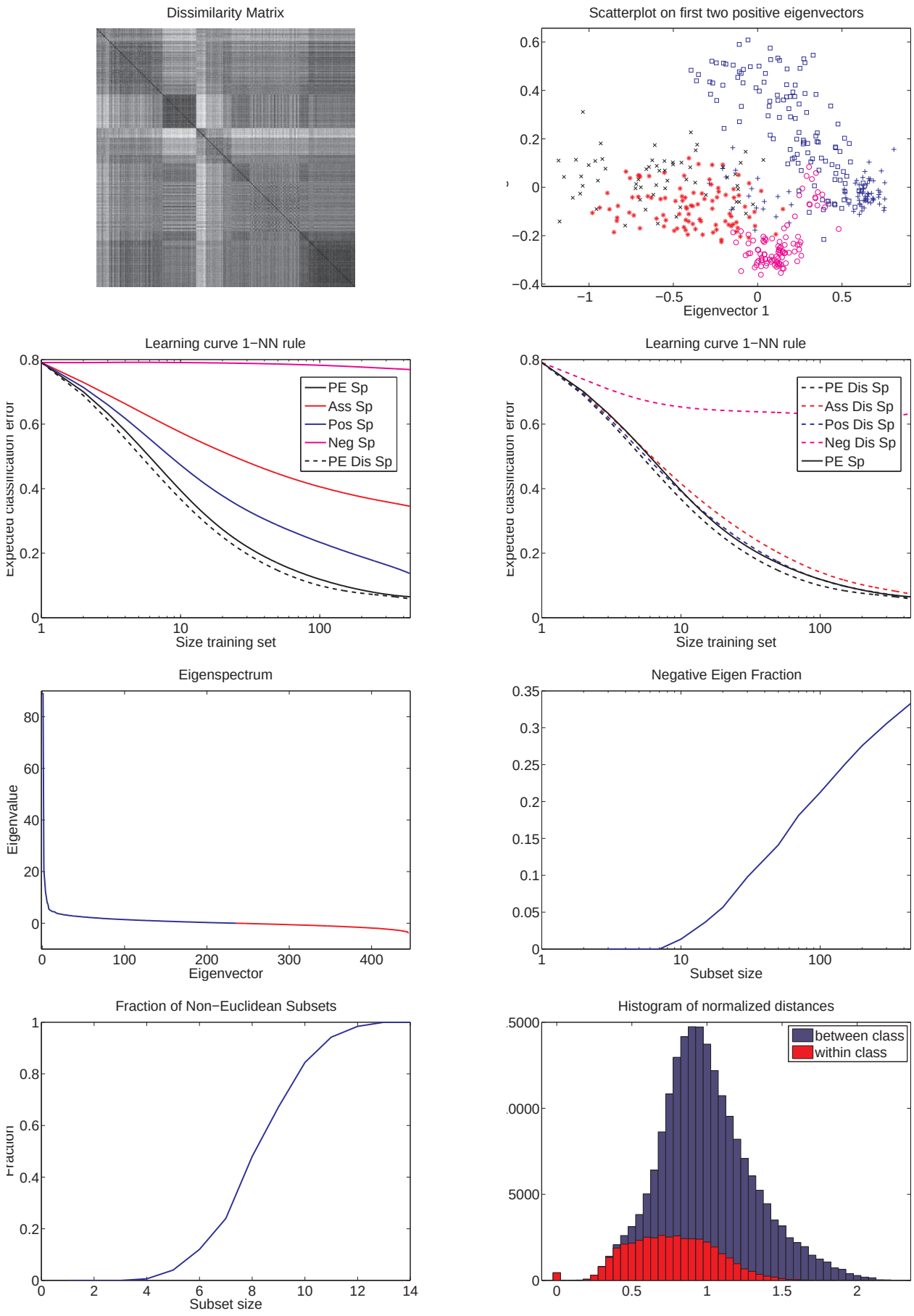


Figure 53: Graphical results for Chickenpieces-31-45.

2.54 Chickenpieces-31-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 31. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.058	asymmetry
446	number of objects
306	number of significant eigenvectors
6867	number of triangle inequality violations out of 88120680
234, 211	number of positive and negative eigenvalues
0.329	negative eigenfraction
0.039	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.747, 1.066	average within-class and between-class dissimilarity
5.4, 15.5, 5.6	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 101: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	11.2 (0.5)	33.0 (0.9)	20.5 (1.1)	79.3 (0.8)	11.2 (0.5)
Parzen	19.8 (0.5)	21.1 (0.7)	20.2 (0.6)	88.7 (0.8)	19.8 (0.5)
NM	10.4 (0.4)	32.8 (1.1)	26.4 (4.3)	73.8 (0.0)	10.8 (0.4)
SVM-1	19.2 (0.9)	9.3 (1.0)	8.8 (0.9)	64.4 (0.8)	12.9 (0.6)

Table 102: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	10.6 (0.7)	14.8 (0.6)	12.6 (0.7)	69.7 (1.3)	9.9 (0.5)
Parzen	32.0 (1.0)	35.4 (0.9)	33.9 (0.9)	64.7 (0.8)	30.3 (0.8)
NM	9.7 (0.8)	14.2 (0.7)	11.8 (0.8)	68.9 (1.2)	11.5 (0.7)
SVM-1	6.8 (0.5)	10.4 (1.0)	8.1 (0.7)	61.8 (1.5)	13.6 (0.7)

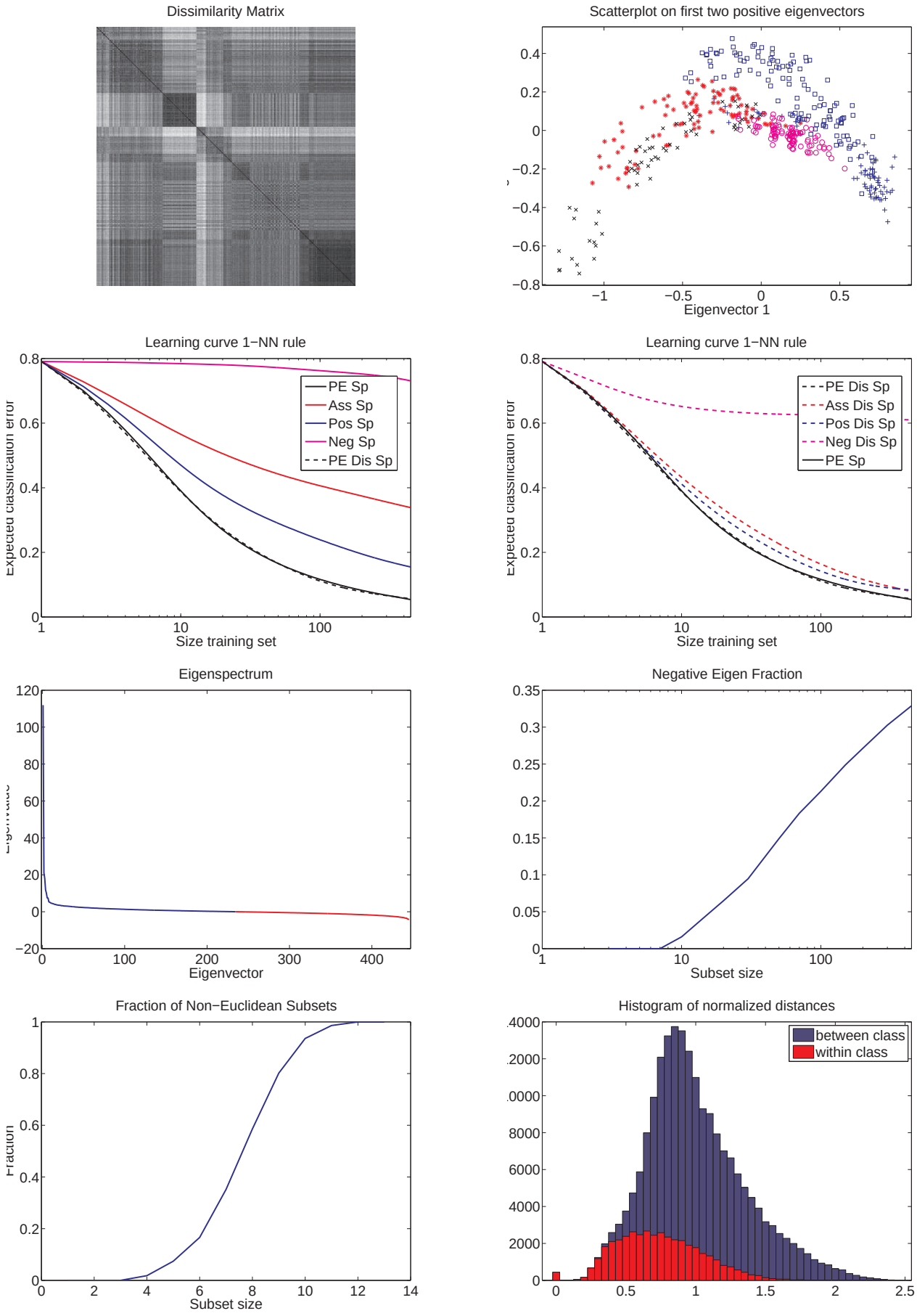


Figure 54: Graphical results for Chickenpieces-31-60.

2.55 Chickenpieces-31-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 31. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.051	asymmetry
446	number of objects
293	number of significant eigenvectors
7314	number of triangle inequality violations out of 88120680
233, 212	number of positive and negative eigenvalues
0.307	negative eigenfraction
0.029	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.715, 1.075	average within-class and between-class dissimilarity
6.5, 21.7, 6.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 103: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	12.8 (0.7)	34.6 (1.0)	24.6 (1.1)	76.9 (0.8)	12.8 (0.7)
Parzen	25.2 (0.8)	25.8 (0.7)	25.7 (0.9)	88.3 (0.8)	25.2 (0.8)
NM	12.2 (0.8)	36.5 (1.1)	22.9 (1.1)	73.8 (0.0)	12.4 (0.6)
SVM-1	19.6 (1.1)	10.4 (0.9)	8.9 (0.7)	66.0 (0.9)	15.1 (0.6)

Table 104: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	14.1 (0.8)	17.3 (0.7)	15.1 (0.7)	65.6 (1.2)	13.5 (0.7)
Parzen	34.7 (0.9)	36.2 (0.6)	36.0 (0.7)	64.1 (0.7)	33.4 (0.9)
NM	13.1 (0.7)	16.2 (0.8)	15.1 (0.7)	64.8 (1.1)	14.9 (1.0)
SVM-1	7.6 (0.8)	10.7 (0.7)	8.9 (0.7)	60.4 (1.4)	15.3 (0.7)

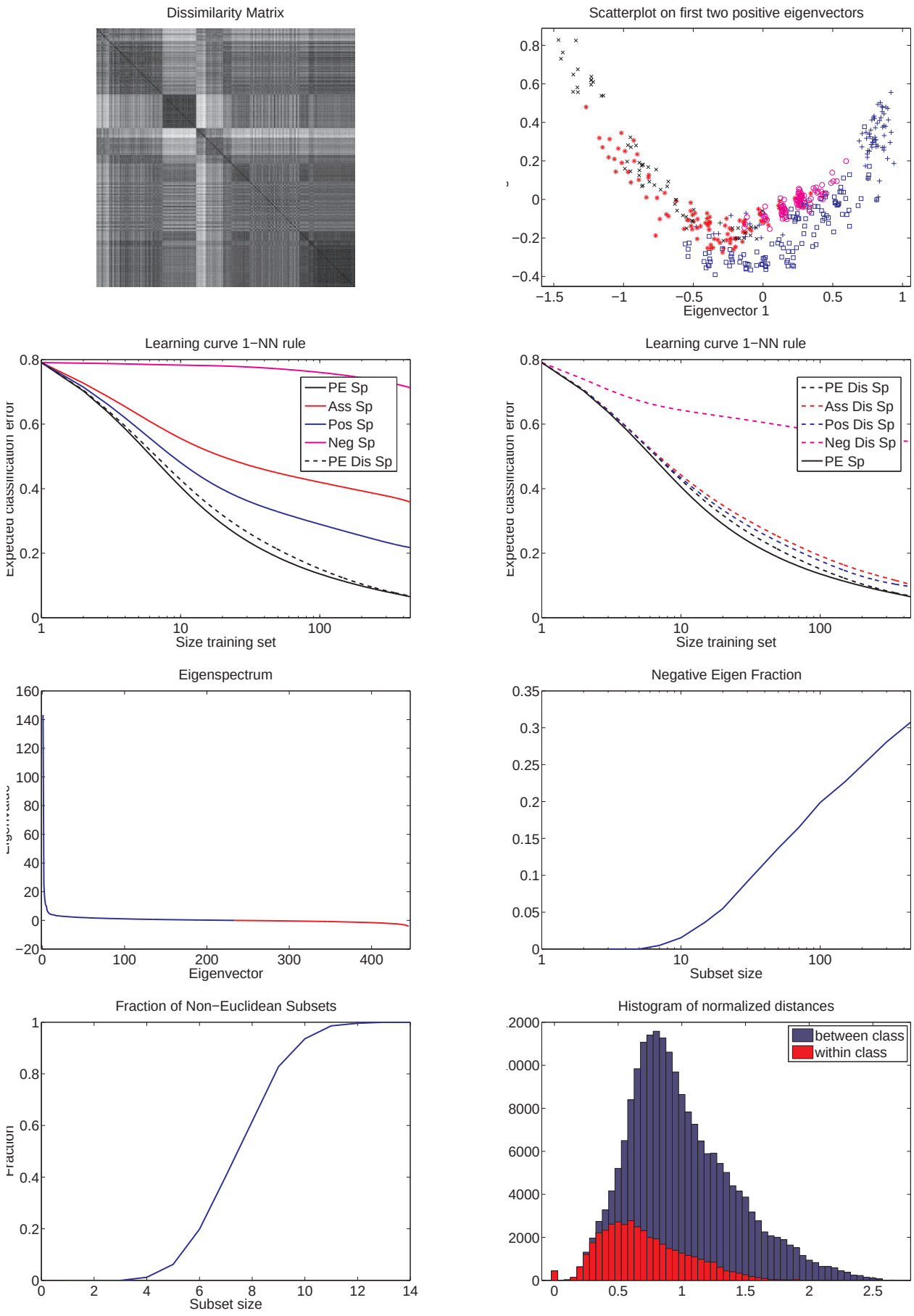


Figure 55: Graphical results for Chickenpieces-31-90.

2.56 Chickenpieces-31-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 31. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.051	asymmetry
446	number of objects
279	number of significant eigenvectors
4240	number of triangle inequality violations out of 88120680
233, 212	number of positive and negative eigenvalues
0.277	negative eigenfraction
0.019	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.696, 1.080	average within-class and between-class dissimilarity
6.7, 24.0, 8.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 105: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	14.5 (0.9)	36.2 (0.9)	27.6 (0.8)	76.9 (0.7)	14.5 (0.9)
Parzen	29.4 (0.9)	30.0 (0.9)	29.7 (0.9)	88.4 (0.8)	29.4 (0.9)
NM	14.3 (1.0)	36.7 (1.2)	27.4 (0.9)	73.8 (0.0)	14.9 (1.0)
SVM-1	21.4 (0.9)	11.2 (0.8)	9.9 (0.7)	66.9 (0.9)	16.0 (0.5)

Table 106: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	16.5 (1.0)	20.2 (1.5)	17.9 (1.2)	65.9 (1.2)	16.2 (1.1)
Parzen	38.0 (1.1)	36.9 (0.5)	37.3 (0.6)	59.1 (0.8)	38.1 (0.9)
NM	16.7 (1.0)	19.2 (1.2)	18.3 (1.1)	64.6 (1.3)	20.1 (1.2)
SVM-1	8.8 (0.6)	12.3 (0.7)	10.2 (0.7)	61.4 (1.5)	15.8 (0.8)

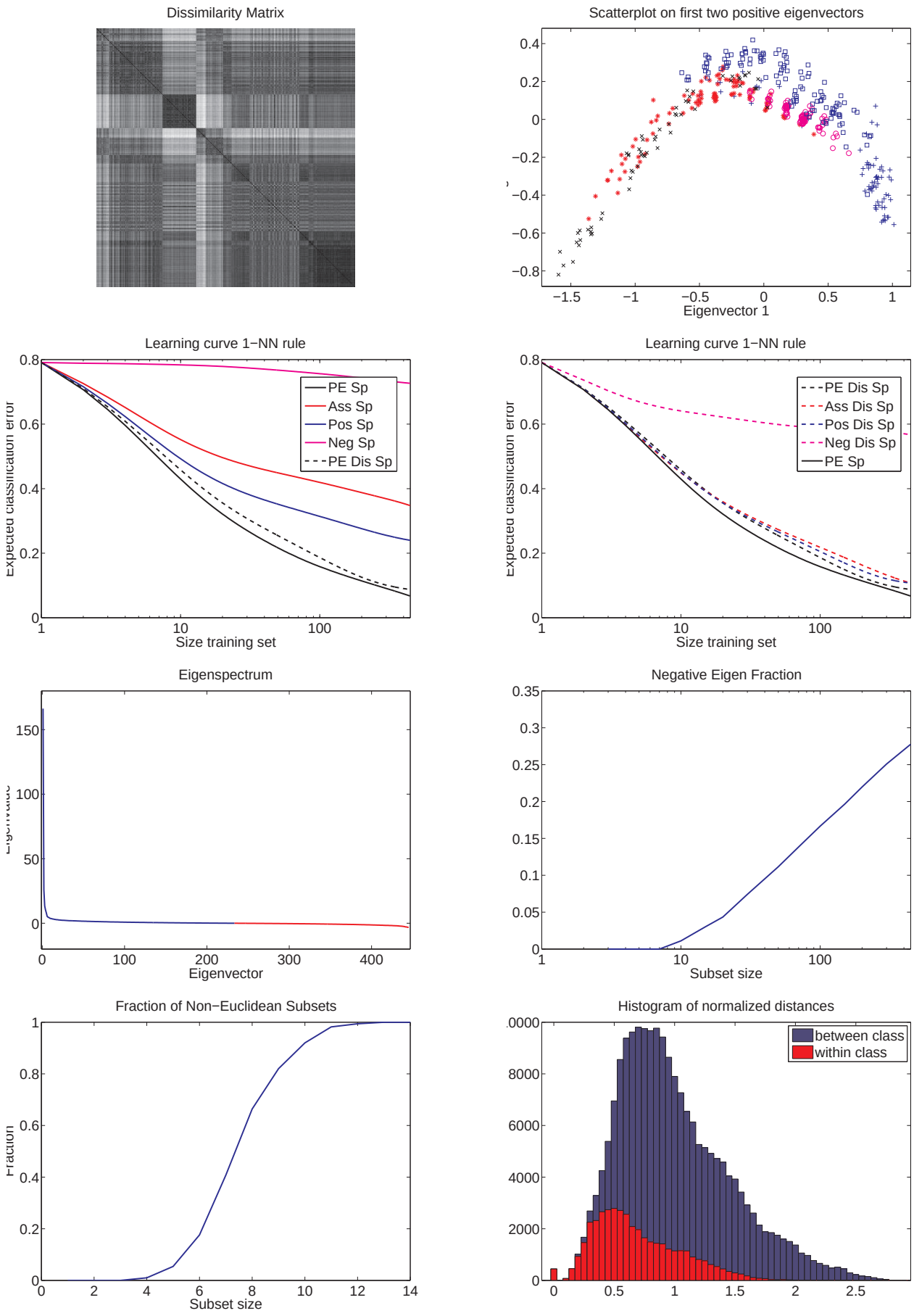


Figure 56: Graphical results for Chickenpieces-31-120.

2.57 Chickenpieces-35-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 35. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.073	asymmetry
446	number of objects
313	number of significant eigenvectors
4834	number of triangle inequality violations out of 88120680
233, 212	number of positive and negative eigenvalues
0.339	negative eigenfraction
0.046	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.782, 1.057	average within-class and between-class dissimilarity
6.3, 15.0, 3.8	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 107: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	10.4 (0.5)	30.1 (0.9)	17.1 (0.8)	82.3 (0.9)	10.4 (0.5)
Parzen	18.1 (0.4)	18.6 (0.5)	18.1 (0.5)	91.2 (0.6)	18.1 (0.4)
NM	10.0 (0.7)	31.0 (0.9)	17.0 (0.7)	73.8 (0.0)	10.0 (0.7)
SVM-1	20.6 (0.7)	9.3 (0.5)	8.2 (0.6)	62.6 (1.0)	11.4 (0.4)

Table 108: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	7.6 (0.6)	11.6 (0.6)	9.6 (0.7)	72.7 (1.0)	8.3 (0.7)
Parzen	30.8 (0.7)	34.4 (0.7)	33.1 (0.6)	63.3 (0.8)	28.8 (0.5)
NM	7.3 (0.5)	11.1 (0.6)	9.4 (0.6)	72.0 (1.0)	8.4 (0.5)
SVM-1	6.3 (0.4)	9.2 (0.4)	7.5 (0.4)	62.0 (0.7)	11.4 (0.4)

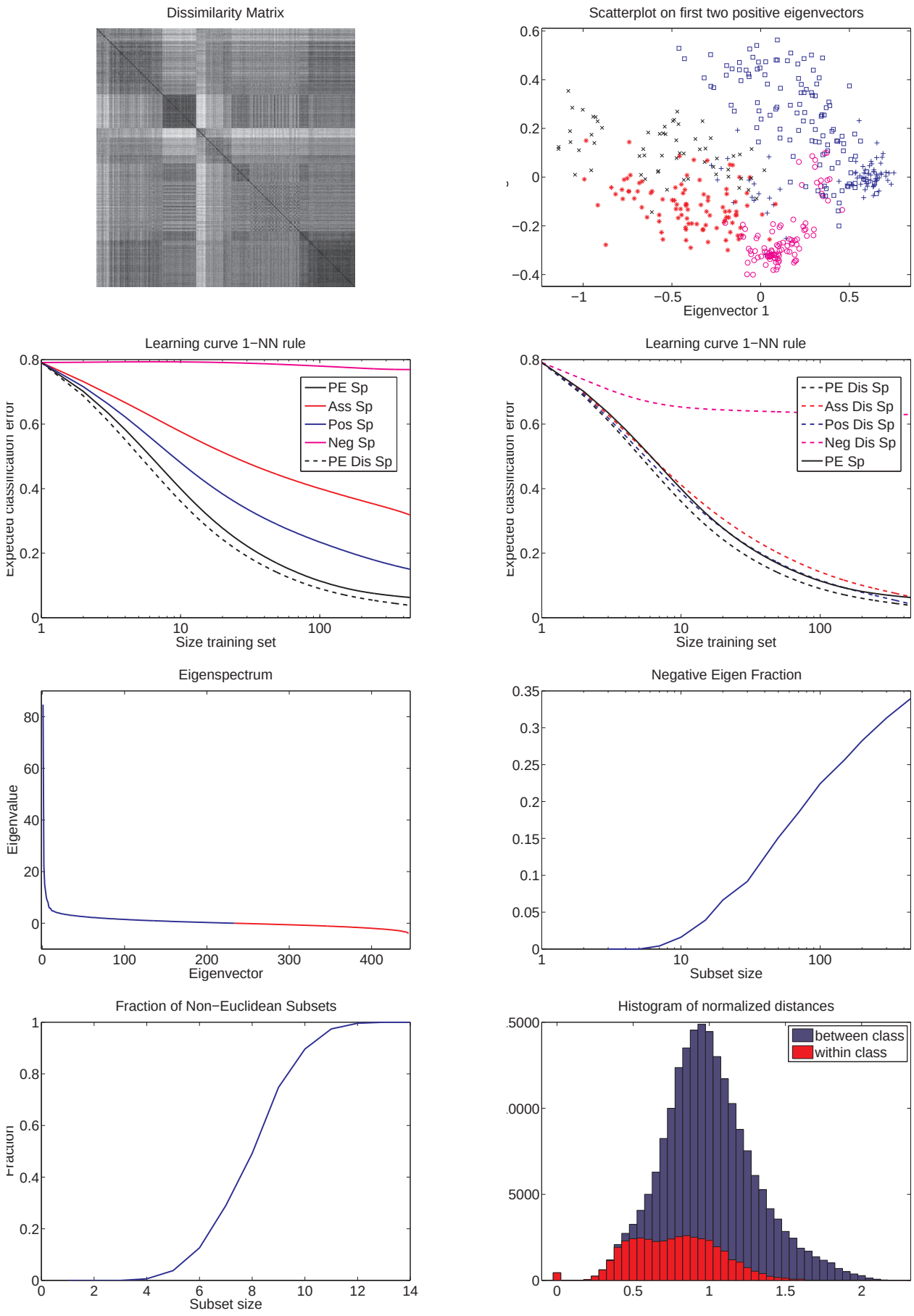


Figure 57: Graphical results for Chickenpieces-35-45.

2.58 Chickenpieces-35-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 35. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.061	asymmetry
446	number of objects
306	number of significant eigenvectors
9416	number of triangle inequality violations out of 88120680
232, 213	number of positive and negative eigenvalues
0.336	negative eigenfraction
0.041	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.751, 1.065	average within-class and between-class dissimilarity
5.6, 13.2, 5.2	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 109: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	10.1 (0.6)	30.5 (1.2)	18.1 (1.0)	77.9 (0.6)	10.1 (0.6)
Parzen	19.4 (0.6)	20.2 (0.7)	20.2 (0.7)	90.2 (0.4)	19.4 (0.6)
NM	9.0 (0.6)	31.5 (1.4)	24.1 (4.4)	73.8 (0.0)	9.1 (0.6)
SVM-1	18.3 (1.2)	9.0 (0.8)	8.4 (0.8)	64.4 (0.8)	12.2 (0.6)

Table 110: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	10.2 (0.7)	12.5 (0.9)	11.0 (0.6)	69.4 (0.7)	10.5 (0.6)
Parzen	31.1 (0.8)	35.0 (0.7)	33.7 (0.6)	64.5 (0.7)	29.4 (0.7)
NM	8.8 (0.7)	12.0 (0.7)	10.6 (0.5)	69.0 (0.9)	10.3 (0.8)
SVM-1	6.8 (0.5)	9.6 (0.6)	8.0 (0.6)	60.9 (0.8)	11.7 (0.7)

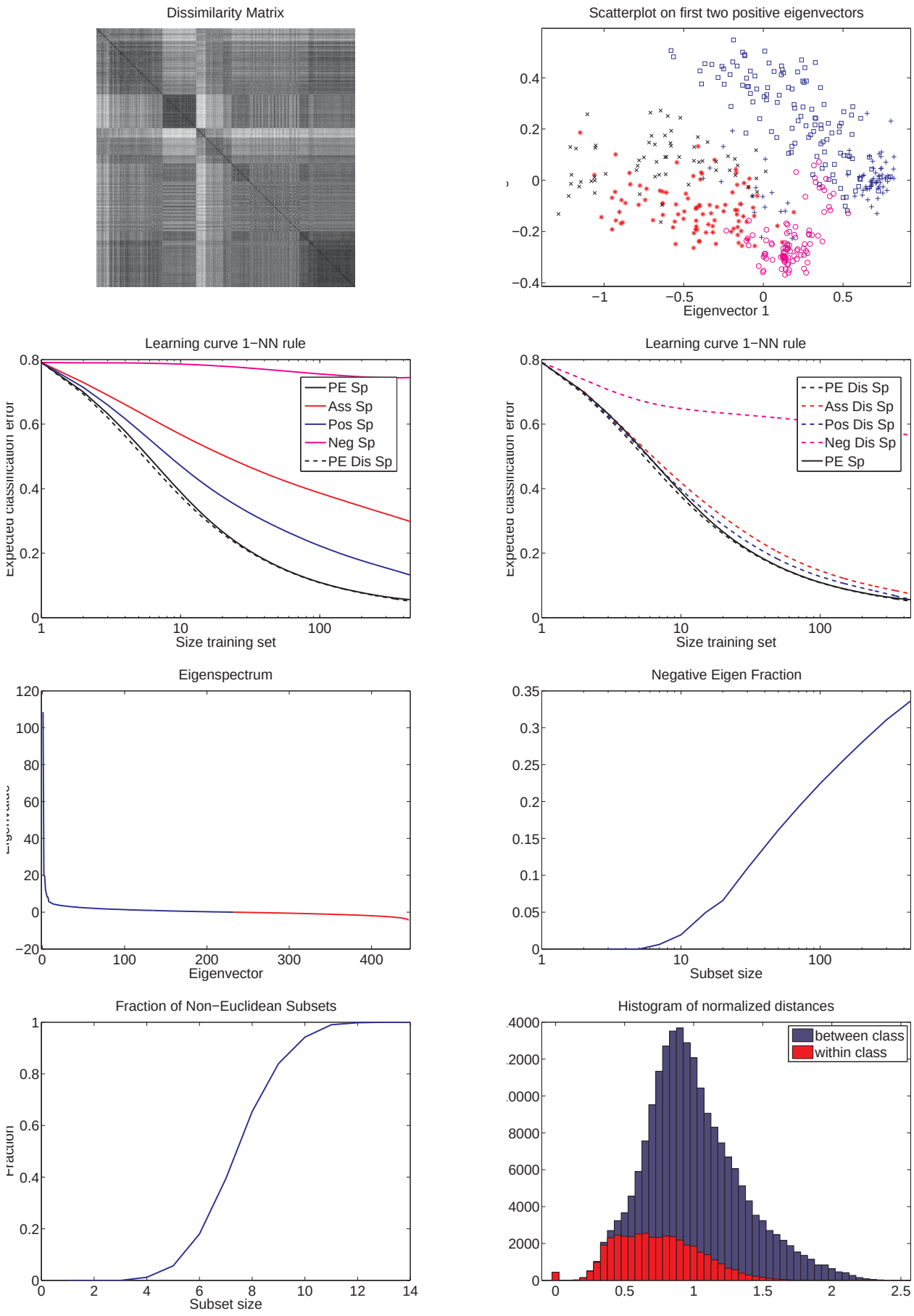


Figure 58: Graphical results for Chickenpieces-35-60.

2.59 Chickenpieces-35-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 35. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.053	asymmetry
446	number of objects
295	number of significant eigenvectors
10457	number of triangle inequality violations out of 88120680
229, 216	number of positive and negative eigenvalues
0.318	negative eigenfraction
0.029	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.716, 1.074	average within-class and between-class dissimilarity
6.1, 19.3, 6.7	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 111: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	11.9 (0.6)	35.2 (1.0)	23.1 (0.9)	76.6 (0.4)	11.9 (0.6)
Parzen	24.1 (0.6)	25.0 (0.7)	24.6 (0.7)	88.9 (0.9)	24.1 (0.6)
NM	10.4 (0.6)	49.0 (4.5)	22.5 (0.9)	73.8 (0.0)	11.0 (0.7)
SVM-1	19.1 (0.7)	10.0 (0.8)	8.5 (0.4)	63.1 (0.9)	14.2 (0.7)

Table 112: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	11.7 (0.6)	14.9 (0.9)	13.6 (0.7)	67.1 (1.1)	13.9 (0.6)
Parzen	34.9 (0.9)	36.3 (0.6)	35.9 (0.6)	63.1 (1.0)	33.8 (0.8)
NM	11.8 (0.8)	14.1 (0.7)	13.4 (0.6)	66.3 (1.0)	14.7 (1.0)
SVM-1	8.0 (0.6)	10.8 (0.7)	8.6 (0.5)	59.8 (1.3)	13.7 (0.7)

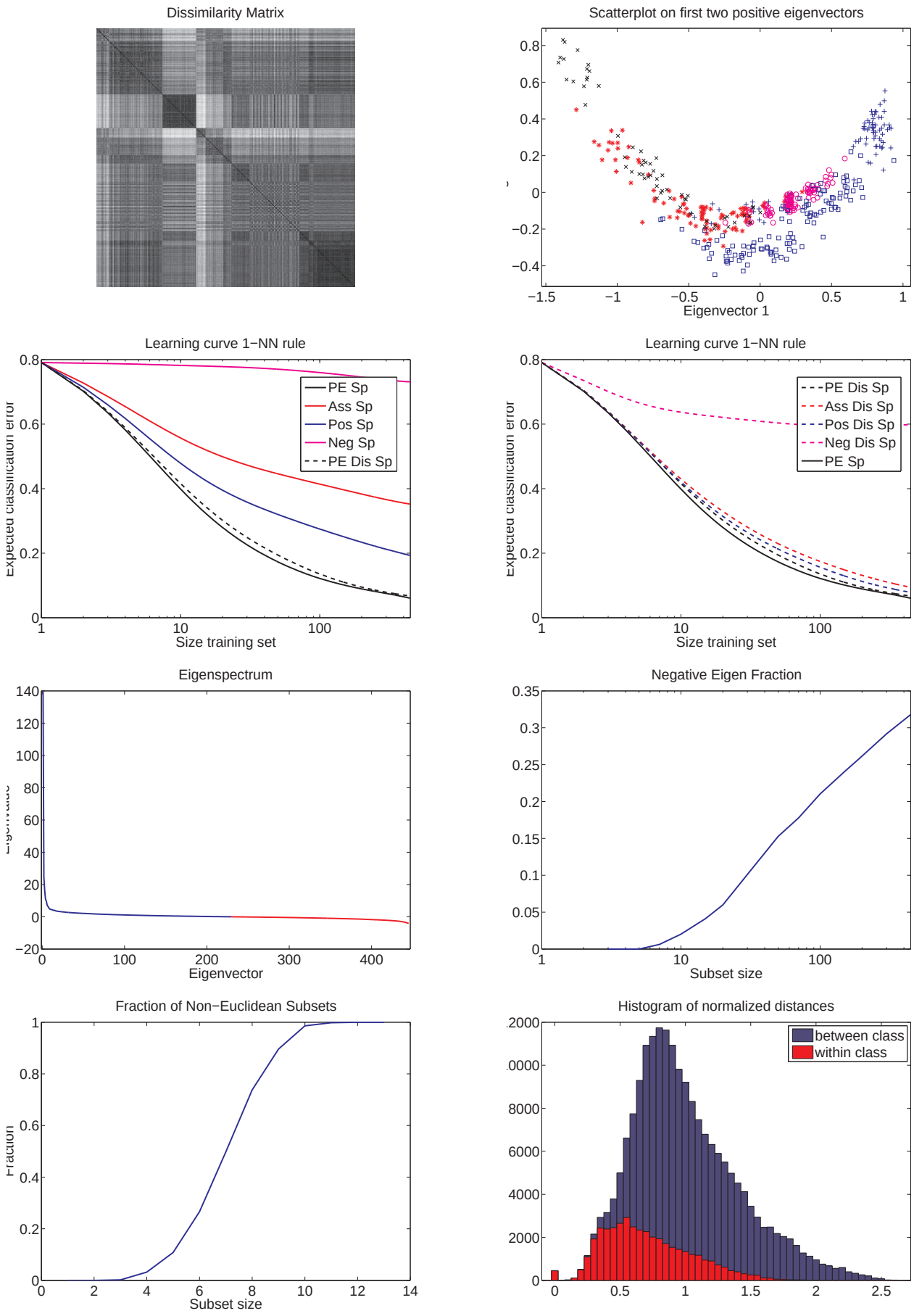


Figure 59: Graphical results for Chickenpieces-35-90.

2.60 Chickenpieces-35-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 35. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.052	asymmetry
446	number of objects
283	number of significant eigenvectors
6971	number of triangle inequality violations out of 88120680
232, 213	number of positive and negative eigenvalues
0.291	negative eigenfraction
0.021	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.697, 1.079	average within-class and between-class dissimilarity
6.5, 22.0, 6.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 113: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	14.1 (0.7)	36.2 (1.1)	26.1 (0.9)	76.2 (0.7)	14.1 (0.7)
Parzen	29.1 (0.9)	29.5 (0.8)	28.8 (0.9)	86.8 (0.4)	29.1 (0.9)
NM	13.3 (0.8)	39.1 (1.0)	27.2 (0.8)	73.8 (0.0)	13.3 (0.8)
SVM-1	20.5 (0.6)	11.1 (0.9)	9.0 (0.6)	64.9 (0.7)	15.5 (0.8)

Table 114: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	15.0 (0.9)	17.9 (1.0)	16.7 (0.9)	65.0 (1.2)	16.1 (0.7)
Parzen	39.2 (1.0)	37.9 (0.5)	38.3 (0.6)	64.4 (0.9)	38.9 (1.1)
NM	15.2 (1.0)	17.6 (0.6)	17.1 (0.8)	64.0 (1.3)	18.9 (1.1)
SVM-1	7.9 (0.6)	11.6 (0.7)	9.1 (0.5)	58.5 (1.1)	15.0 (0.7)

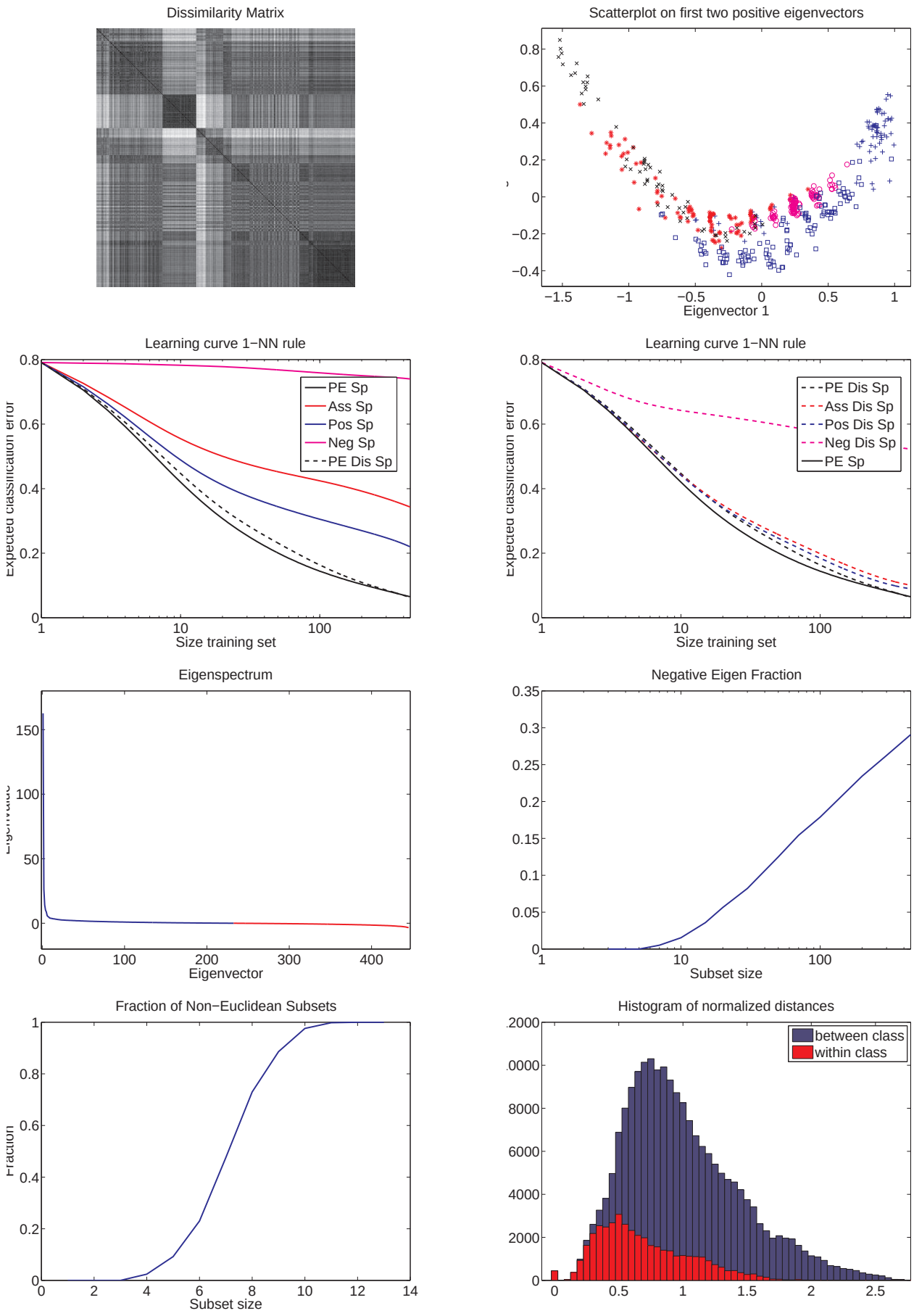


Figure 60: Graphical results for Chickenpieces-35-120.

2.61 Chickenpieces-40-45

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 40. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 45

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.076	asymmetry
446	number of objects
311	number of significant eigenvectors
7549	number of triangle inequality violations out of 88120680
232, 213	number of positive and negative eigenvalues
0.345	negative eigenfraction
0.053	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.780, 1.057	average within-class and between-class dissimilarity
8.7, 15.2, 4.3	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 115: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	12.5 (0.4)	28.4 (1.1)	17.2 (0.6)	81.2 (0.8)	12.5 (0.4)
Parzen	16.8 (0.7)	18.1 (0.6)	17.7 (0.7)	92.8 (0.4)	16.8 (0.7)
NM	11.4 (0.3)	28.5 (0.8)	17.3 (0.3)	73.8 (0.0)	11.8 (0.5)
SVM-1	21.2 (1.2)	10.7 (0.8)	10.8 (0.4)	61.4 (0.8)	11.8 (0.5)

Table 116: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	10.3 (0.9)	12.6 (0.8)	11.6 (0.7)	70.9 (0.7)	10.9 (0.6)
Parzen	27.0 (0.6)	31.2 (0.7)	29.8 (0.6)	62.0 (0.6)	26.2 (0.8)
NM	9.5 (0.8)	11.9 (0.6)	10.7 (0.5)	70.6 (0.8)	10.9 (0.6)
SVM-1	9.1 (0.7)	11.3 (0.6)	10.3 (0.4)	63.8 (1.0)	12.7 (0.4)

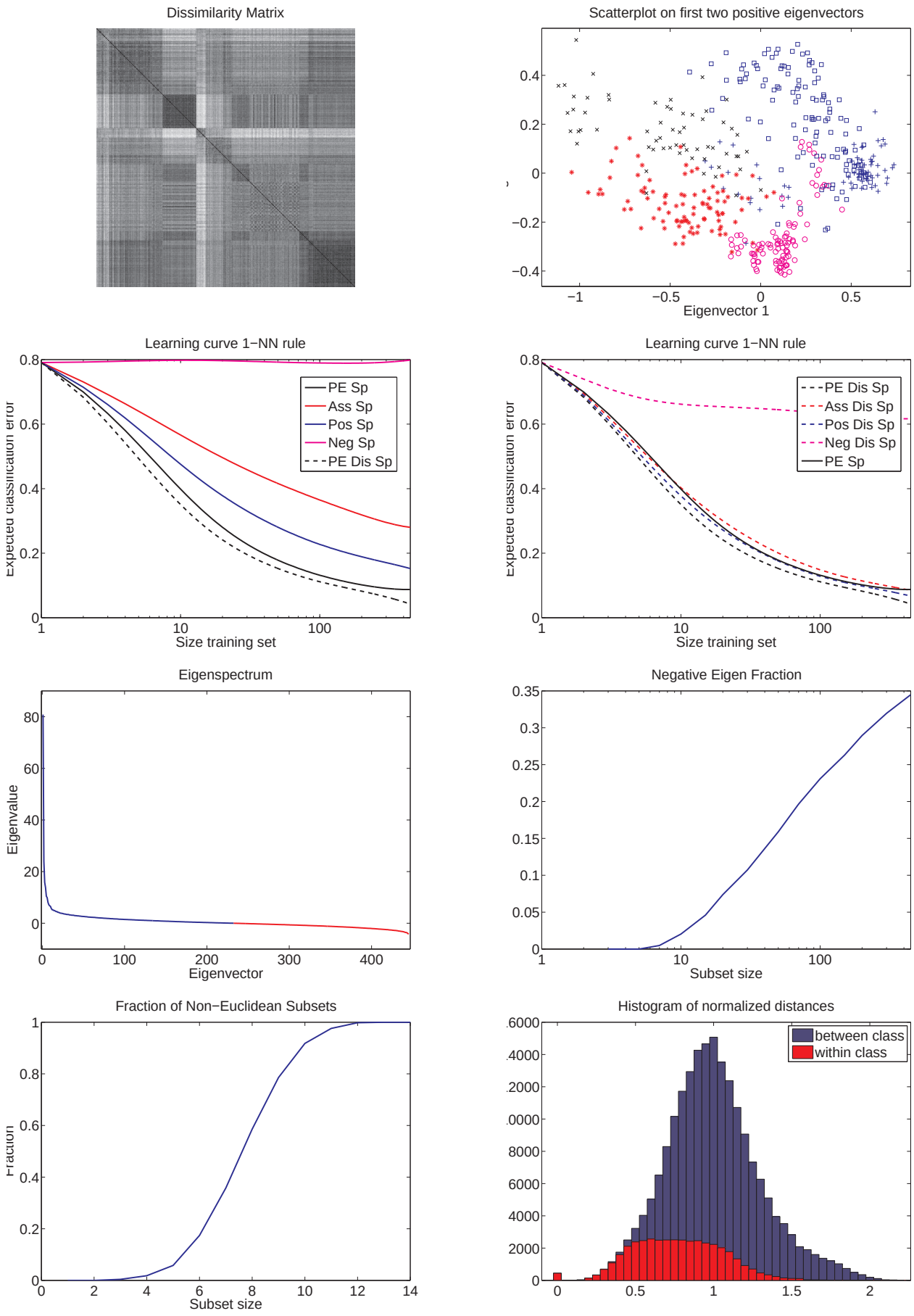


Figure 61: Graphical results for Chickenpieces-40-45.

2.62 Chickenpieces-40-60

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 40. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 60

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.064	asymmetry
446	number of objects
307	number of significant eigenvectors
15838	number of triangle inequality violations out of 88120680
230, 215	number of positive and negative eigenvalues
0.343	negative eigenfraction
0.044	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.746, 1.066	average within-class and between-class dissimilarity
7.8, 12.6, 7.2	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 117: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	11.1 (0.4)	27.1 (1.2)	17.6 (0.9)	79.5 (0.7)	11.1 (0.4)
Parzen	17.4 (0.8)	18.5 (0.9)	18.1 (0.8)	90.3 (0.6)	17.4 (0.8)
NM	10.6 (0.4)	27.8 (0.9)	16.6 (0.8)	73.8 (0.0)	10.8 (0.4)
SVM-1	20.2 (1.3)	10.4 (0.4)	9.4 (0.4)	63.2 (0.7)	12.5 (0.6)

Table 118: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	10.9 (1.0)	14.9 (1.1)	13.3 (1.0)	70.4 (0.6)	11.3 (0.8)
Parzen	29.6 (0.9)	33.6 (0.9)	32.5 (0.9)	64.0 (0.7)	28.5 (0.7)
NM	10.7 (0.9)	13.9 (0.9)	12.7 (0.8)	69.7 (0.7)	11.5 (0.5)
SVM-1	8.7 (0.6)	11.8 (0.5)	9.8 (0.4)	62.1 (1.1)	12.6 (0.5)

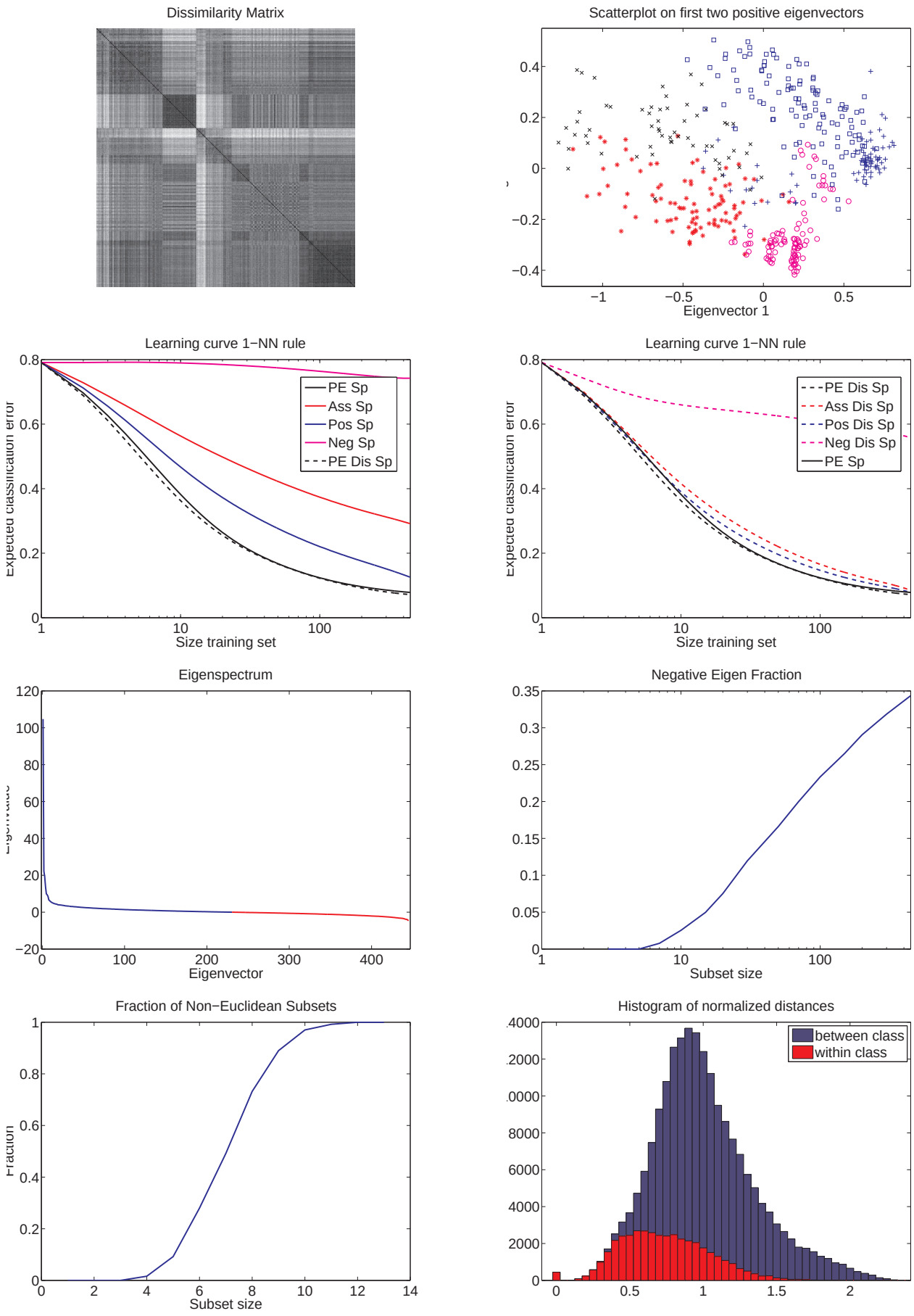


Figure 62: Graphical results for Chickenpieces-40-60.

2.63 Chickenpieces-40-90

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 40. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 90

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.055	asymmetry
446	number of objects
297	number of significant eigenvectors
20504	number of triangle inequality violations out of 88120680
230, 215	number of positive and negative eigenvalues
0.327	negative eigenfraction
0.034	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.709, 1.076	average within-class and between-class dissimilarity
9.4, 18.8, 6.1	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 119: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	12.2 (0.5)	34.0 (1.2)	20.7 (0.9)	78.1 (0.7)	12.2 (0.5)
Parzen	22.1 (0.6)	22.0 (0.5)	22.1 (0.6)	90.0 (0.8)	22.1 (0.6)
NM	10.5 (0.5)	43.7 (3.7)	20.0 (0.9)	73.8 (0.0)	10.9 (0.4)
SVM-1	20.9 (0.8)	11.6 (0.6)	11.4 (0.6)	61.0 (0.8)	14.2 (0.5)

Table 120: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	12.2 (0.7)	15.4 (0.7)	13.7 (0.4)	67.8 (1.1)	13.4 (0.5)
Parzen	31.9 (0.7)	36.9 (1.0)	34.9 (0.8)	62.2 (1.3)	31.4 (0.8)
NM	11.6 (0.6)	14.5 (0.6)	13.2 (0.5)	67.6 (1.2)	13.3 (0.5)
SVM-1	8.8 (0.5)	11.9 (0.4)	10.4 (0.5)	66.4 (0.7)	14.5 (0.4)

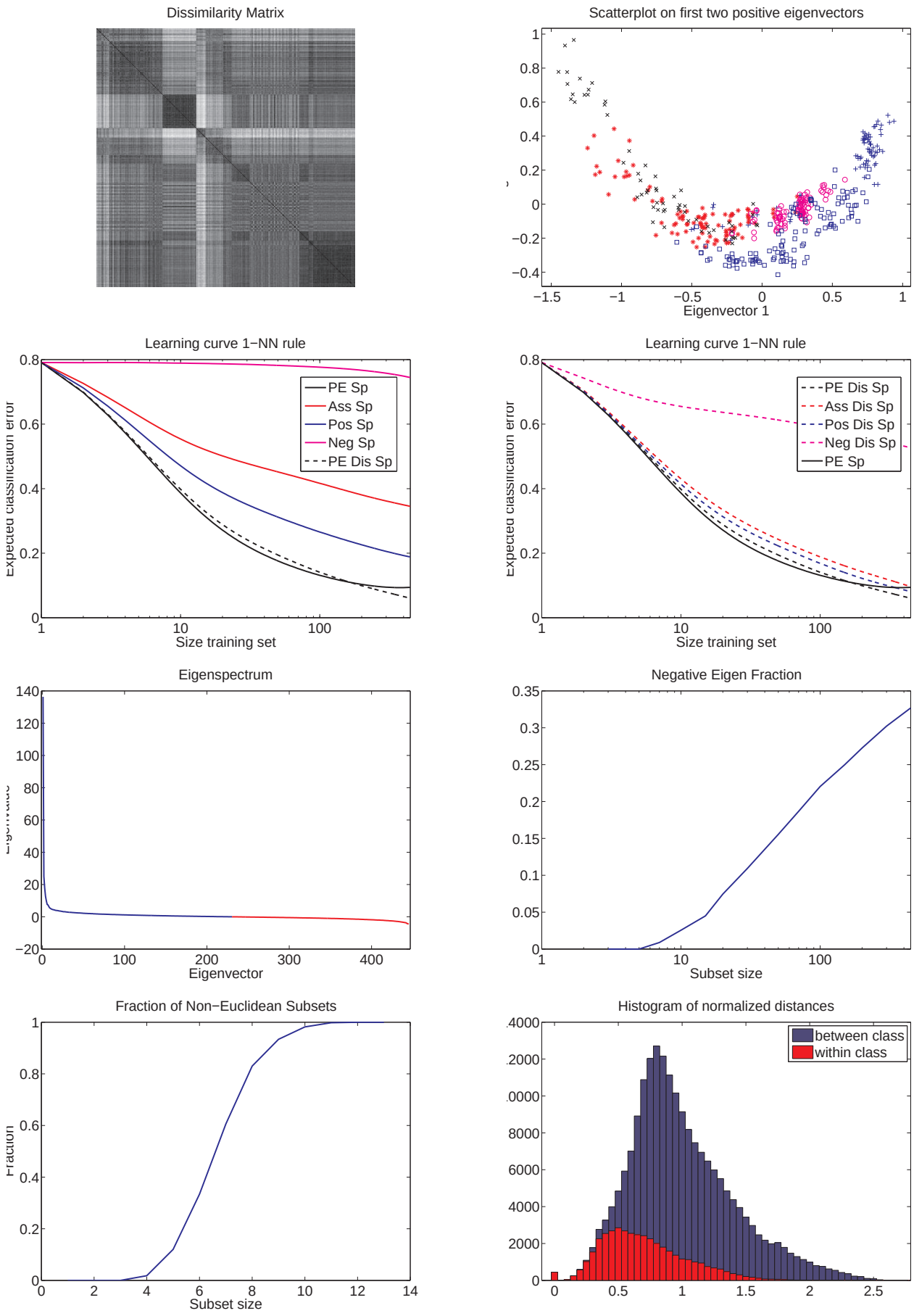


Figure 63: Graphical results for Chickenpieces-40-90.

2.64 Chickenpieces-40-120

Description Dissimilarity Dataset

This is one of the chickenpieces dissimilarity matrices as made available by Bunke et.al. Every entry is a weighted edit distance between two strings representing the contours of 2D blobs. Contours are approximated by vectors of length 40. Angles between vectors are used as replacement costs. The costs for insertion and deletion are 120

Reference(s)

H. Bunke, H., U. Buhler, Applications of approximate string matching to 2D shape recognition, Pattern recognition 26 (1993) 1797-1812

Web page(s)

The original dissimilarity datasets http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices

The original images http://algoval.essex.ac.uk/data/sequence/chicken/

The PRTools versions http://prlab.tudelft.nl/data/disdatasets.html

0.054	asymmetry
446	number of objects
284	number of significant eigenvectors
16471	number of triangle inequality violations out of 88120680
228, 217	number of positive and negative eigenvalues
0.302	negative eigenfraction
0.026	negative eigenratio
5	number of classes, class sizes [76 96 96 61 117]
0.690, 1.081	average within-class and between-class dissimilarity
9.9, 22.9, 6.5	LOO nearest neighbor errors (%) of the PE, Pos and Dis Spaces

Table 121: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the embedded spaces. The st. dev. of the means are in ().

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	13.9 (0.6)	34.2 (1.3)	24.2 (1.2)	74.0 (0.5)	13.9 (0.6)
Parzen	25.6 (0.7)	26.0 (0.8)	25.6 (0.7)	88.9 (0.6)	25.6 (0.7)
NM	12.6 (0.4)	37.5 (1.0)	23.8 (0.9)	73.8 (0.0)	12.5 (0.4)
SVM-1	20.6 (0.7)	12.0 (0.9)	10.9 (0.8)	61.8 (0.7)	15.1 (0.4)

Table 122: Mean 2-fold cross-validation errors (%) of 10 random subsets of 50 objects per class for four classifiers in the dissimilarity spaces. The st. dev. of the means are in ().

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	13.8 (0.6)	17.7 (0.9)	15.4 (0.6)	61.5 (0.8)	15.0 (0.7)
Parzen	34.0 (0.8)	36.1 (0.7)	35.1 (0.6)	62.0 (0.6)	34.3 (0.8)
NM	14.4 (0.7)	16.2 (0.5)	15.6 (0.6)	60.7 (0.6)	16.7 (0.6)
SVM-1	9.4 (0.7)	13.2 (0.6)	11.0 (0.6)	62.5 (1.2)	15.4 (0.5)

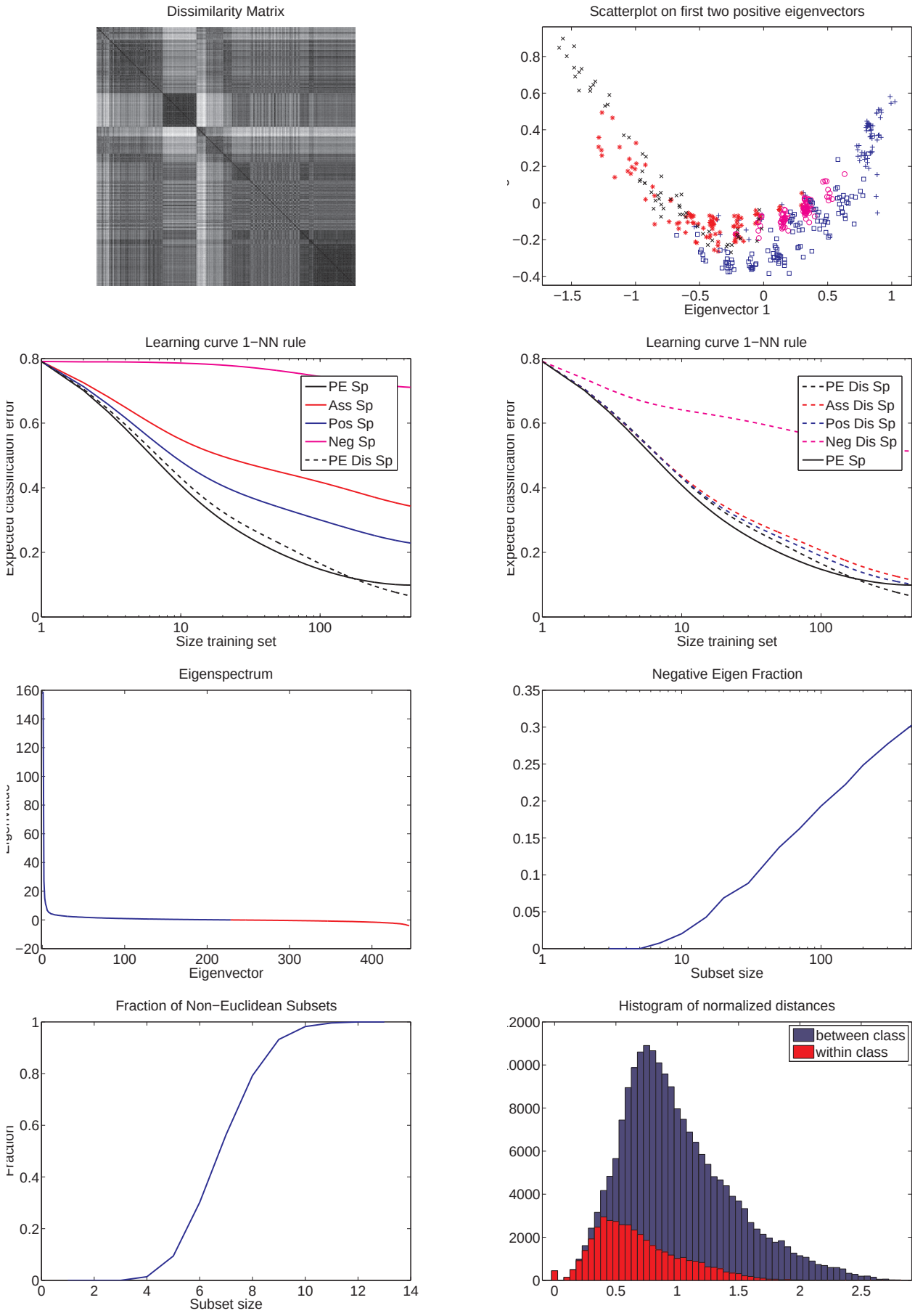


Figure 64: Graphical results for Chickenpieces-40-120.

3 Scatterplots

In this section we present various 2D scatterplots derived from the tables of the individual datasets. The axes of each scatterplot represent classification errors for the same classifier in two different spaces. They, thereby, summerize the relations between the various spaces. On the left we will always present the result for the 44 Chickenpieces datasets, while on the right we will focus on the remaining datasets. For the Chickenpieces data just four different markers are used for the four different cost factors to increase the visibility.

In the coming four subsections we will make four types of comparisons.

(a) The nearest neighbor errors in various spaces as estimated by the leave-one-out estimators over the entire datasets.

(b) The classification errors, per classifier, in the various Euclidean correction spaces (the Ass Space, the Pos Space, the Neg Space and the Cor Space) compared with the error in the PE Space, based on the 2-fold crossvalidation experiments repeated for 10 subsets of 50 objects per class.

(c) A similar comparison as above, but now for the corresponding dissimilarity spaces.

(d) A comparison between the classification errors in the variants of the PE-embedded spaces and the corresponding dissimilarity spaces.

When comparing two spaces, we will also write

$$\text{Space-1} \succeq_{\mathtt{cf}} \text{Space-2}$$

to indicate that the classifier `cf` performs similarly or better in the Space-1 than in the Space-2 as judged on the collection of available datasets. In addition, we will also use the symbols of $\approx_{\mathtt{cf}}$ or $=_{\mathtt{cf}}$ to indicate when the performances of the classifier `cf` are about similar, or, respectively, identical in both spaces.

3.1 1-NN error on the entire datasets

The next page presents a comparison between the nearest neighbor errors on the entire datasets embedded into various spaces and the nearest neighbor errors on the original datasets (in the complete PE space). By studying the plots, the following observations can be made:

(a) Neg Space is not useful for the 1-NN rule.

(b) In general, we have:
Dis Space $\succ_{\mathtt{1-NN}}$ Pos Space $\succ_{\mathtt{1-NN}}$ Ass Space $\succ_{\mathtt{1-NN}}$ Neg Space.

(c) Concerning the Chickenpieces data, we see that the 1-NN rule

- performs much better in the PE space than in the Ass Space. This holds for all sets.
- performs better in the PE space than in the Pos Space. This holds for all sets.
- performs similarly or worse in the PE space than in the Dis Space.

(d) Concerning all other data, we observe that:

- The 1-NN rule performs usually better or somewhat better in the PE Space than in the Ass Space and Pos Space.
- In half of the examples, the 1-NN rule performs (somewhat) better in the Dis Space than in PE Space.
- performs similarly or worse in the complete PE space than in the Dis Space.

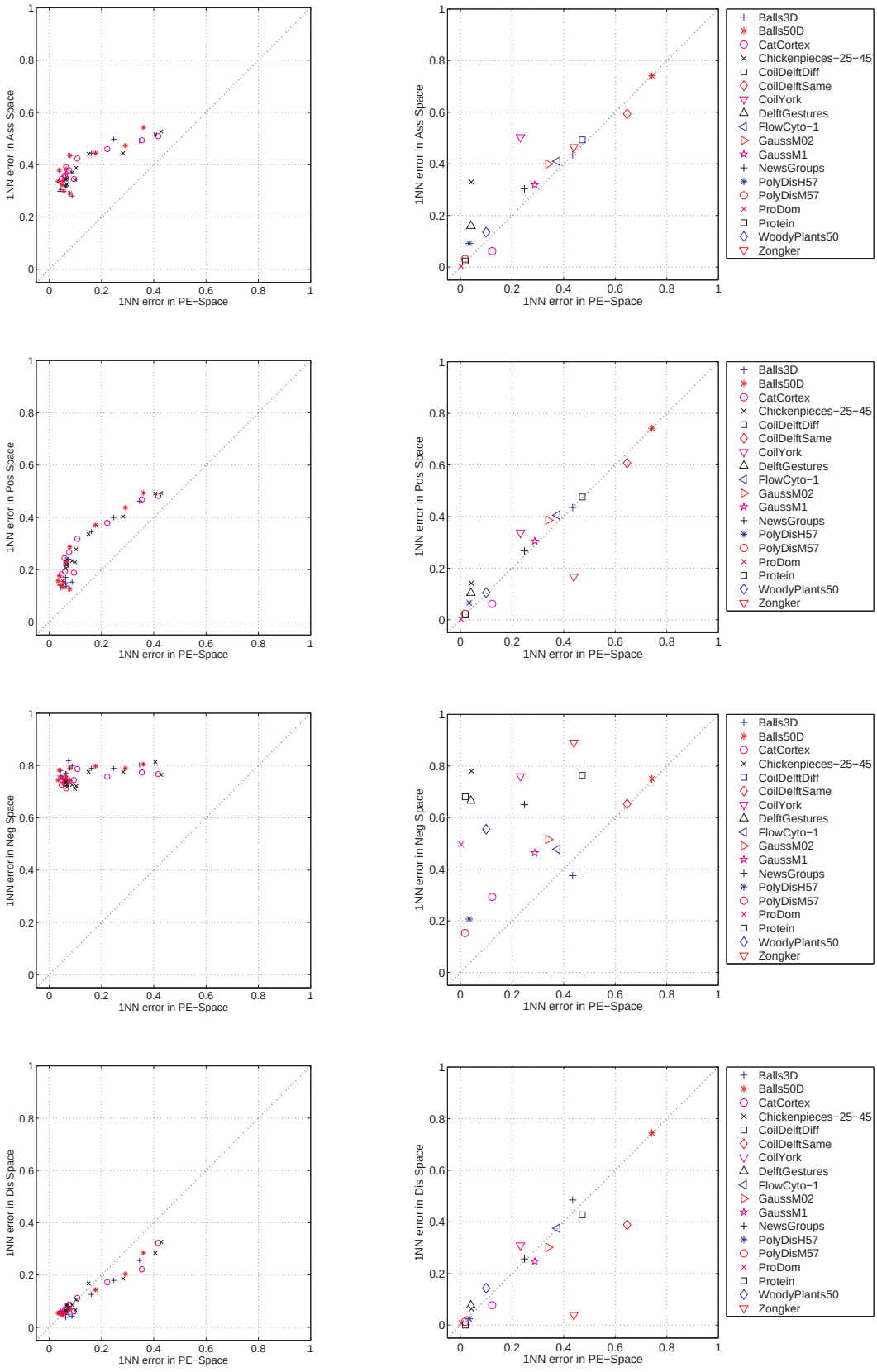


Figure 65: Scatterplots of the leave-one-out 1-NN errors in various spaces versus the original PE Space (defined by the given dissimilarities). The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.2 Classification errors in the PE-related embedded spaces

Here, we consider the performance of various classifiers in the PE space and its modified variants. The scatterplots below enable the comparison of the PE space versus its modified spaces for the given classifier. Each subsection is devoted to an individual classifier.

3.2.1 1-NN errors in the PE Space and its modified spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) Neg Space is not useful for the 1-NN rule.

(b) Obviously (as it is evident from the construction of the spaces), the 1-NN performance is the same in the Cor and PE Spaces.

(c) Although the Neg Space shows bad to random performances, removing or inverting its contribution is counterproductive as follows from the deteriorating performances of the Pos Space and Ass Space.

(d) In general, we have:
PE Space $=_{1-\mathsf{NN}}$ Cor Space $\succeq_{1-\mathsf{NN}}$ Pos Space $\succeq_{1-\mathsf{NN}}$ Ass Space $\succ_{1-\mathsf{NN}}$ Neg Space.

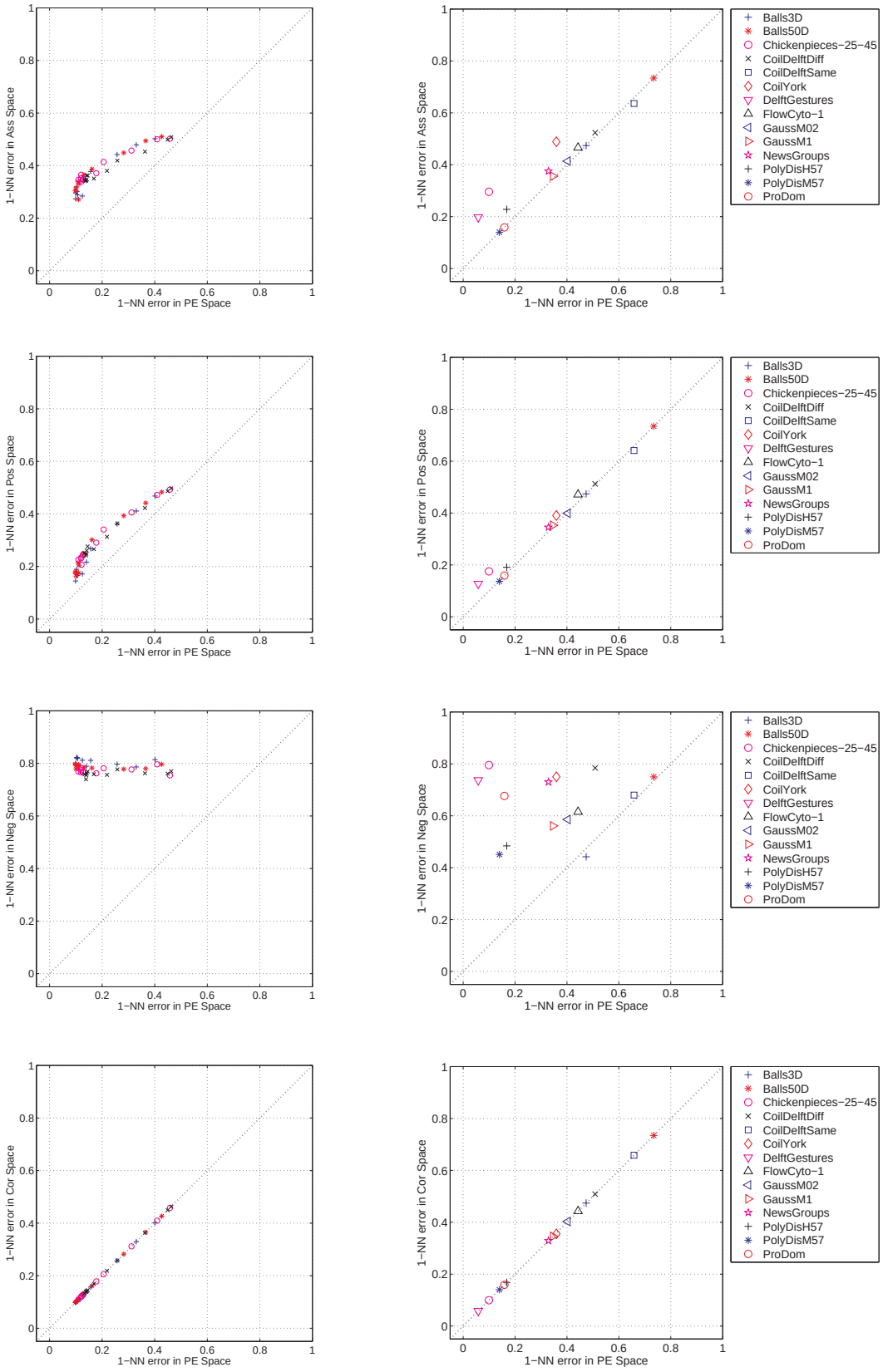


Figure 66: Scatterplots of the 1-NN errors in various spaces versus the original PE Space (defined by the given dissimilarities) based on 10 runs of 2-fold crossvalidation for 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.2.2 Parzen errors in the PE Space and its modified spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) Neg Space is not useful for the Parzen classifier. The results in the Neg Space are especially bad for all Chickenpieces data.

(b) Obviously (as it is evident from the construction of the spaces), the performance of the Parzen classifier is the same in the Cor and PE Spaces.

(c) The Parzen classifier performs nearly the same in the three spaces: PE Space, Ass Space and Pos Space for all datasets.

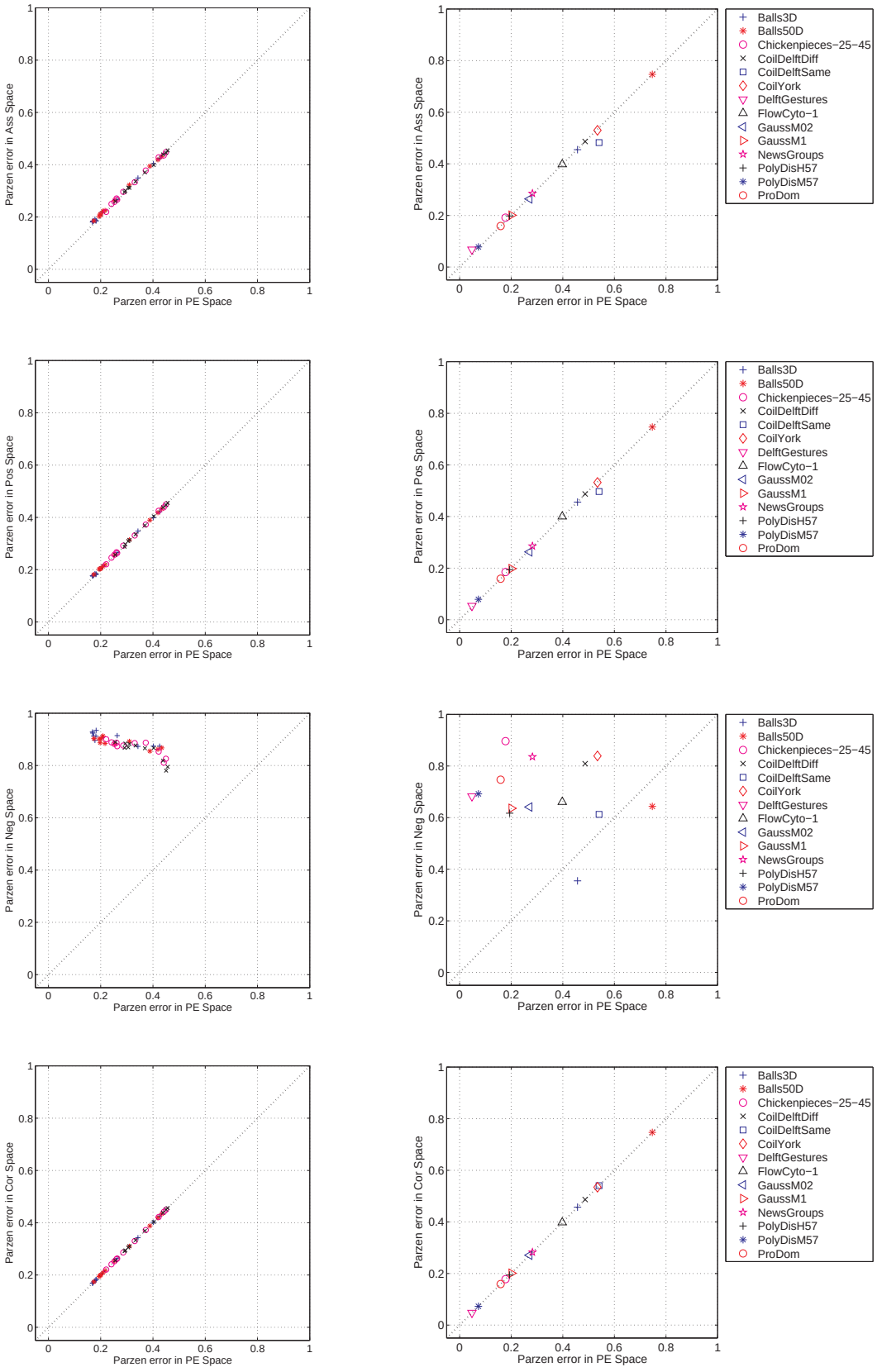


Figure 67: Scatterplots of the Parzen errors in various spaces versus the original PE Space (defined by the given dissimilarities) based on 10 runs of 2-fold crossvalidation for 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.2.3 Nearest Mean errors in the PE Space and its modified spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) Neg Space is not useful for the NM classifier. The NM errors are nearly identical for all Chickenpieces sets.

(b) Although the Neg Space shows bad to random performances, removing or inverting its contribution is counterproductive as follows from the deteriorating performances of the Pos Space and Ass Space.

(c) In general (except for two Chickenpieces sets), we have:
PE Space $\approx_{\mathsf{NM}}$ Cor Space $\succeq_{\mathsf{NM}}$ Pos Space $\succeq_{\mathsf{NM}}$ Ass Space $\succ_{\mathsf{NM}}$ Neg Space.

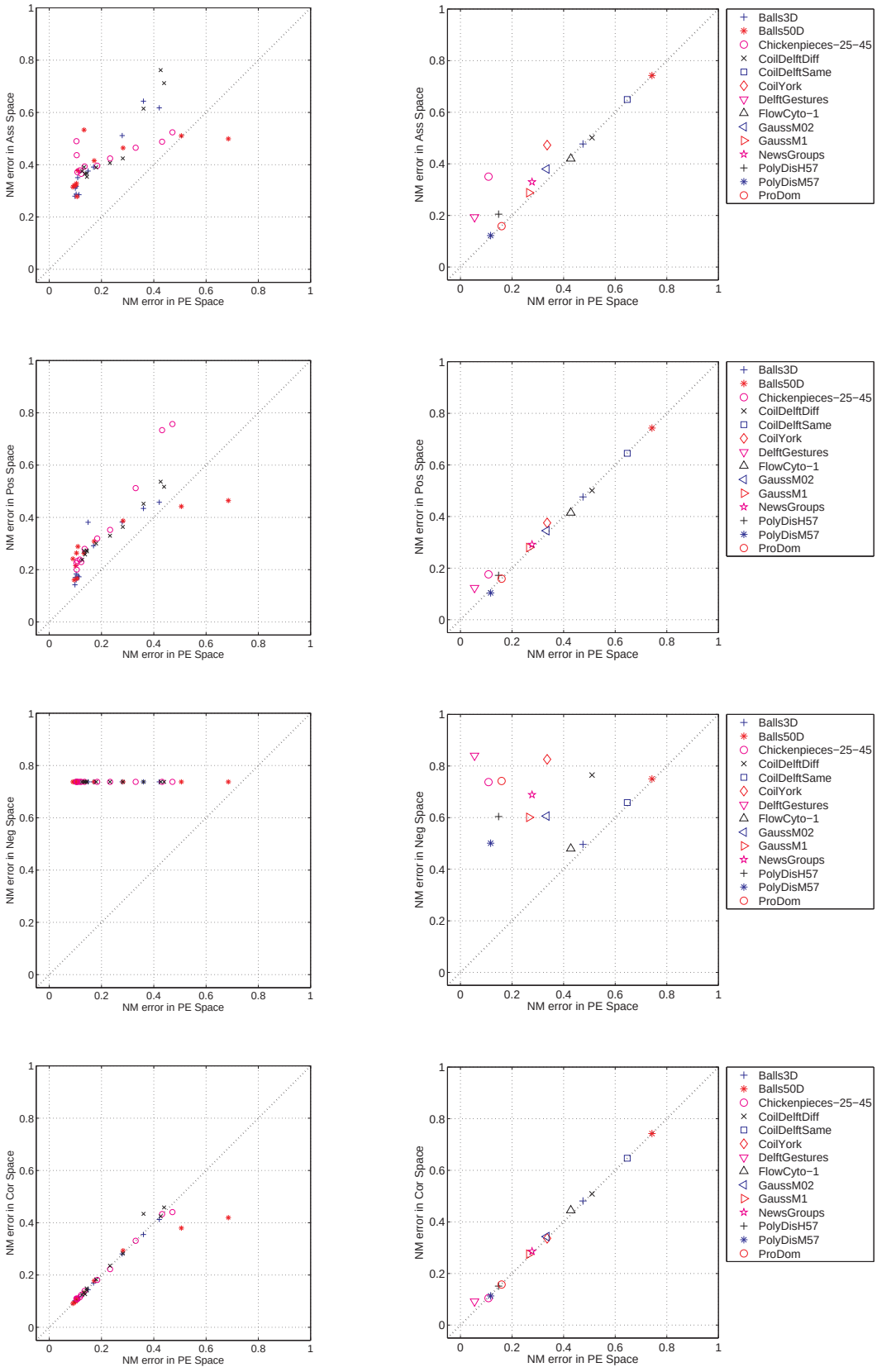


Figure 68: Scatterplots of the NM errors in various spaces versus the original PE Space (defined by the given dissimilarities) based on 10 runs of crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.2.4 Linear SVM errors in the PE Space and its modified spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) Neg Space is not useful for the linear SVM classifier.

(b) The SVM-1 errors are larger in the PE Space than in the other three spaces: Ass Space, Pos Space and Cor Space. The results of SVM-1 are similar in all these spaces for datasets that are nearly Euclidean as judged by a small NER (negative eigenratio) of ≤ 0.03 (on the whole set). These are the Balls3D, Balls50D, ProDom and DelftGestures datasets.

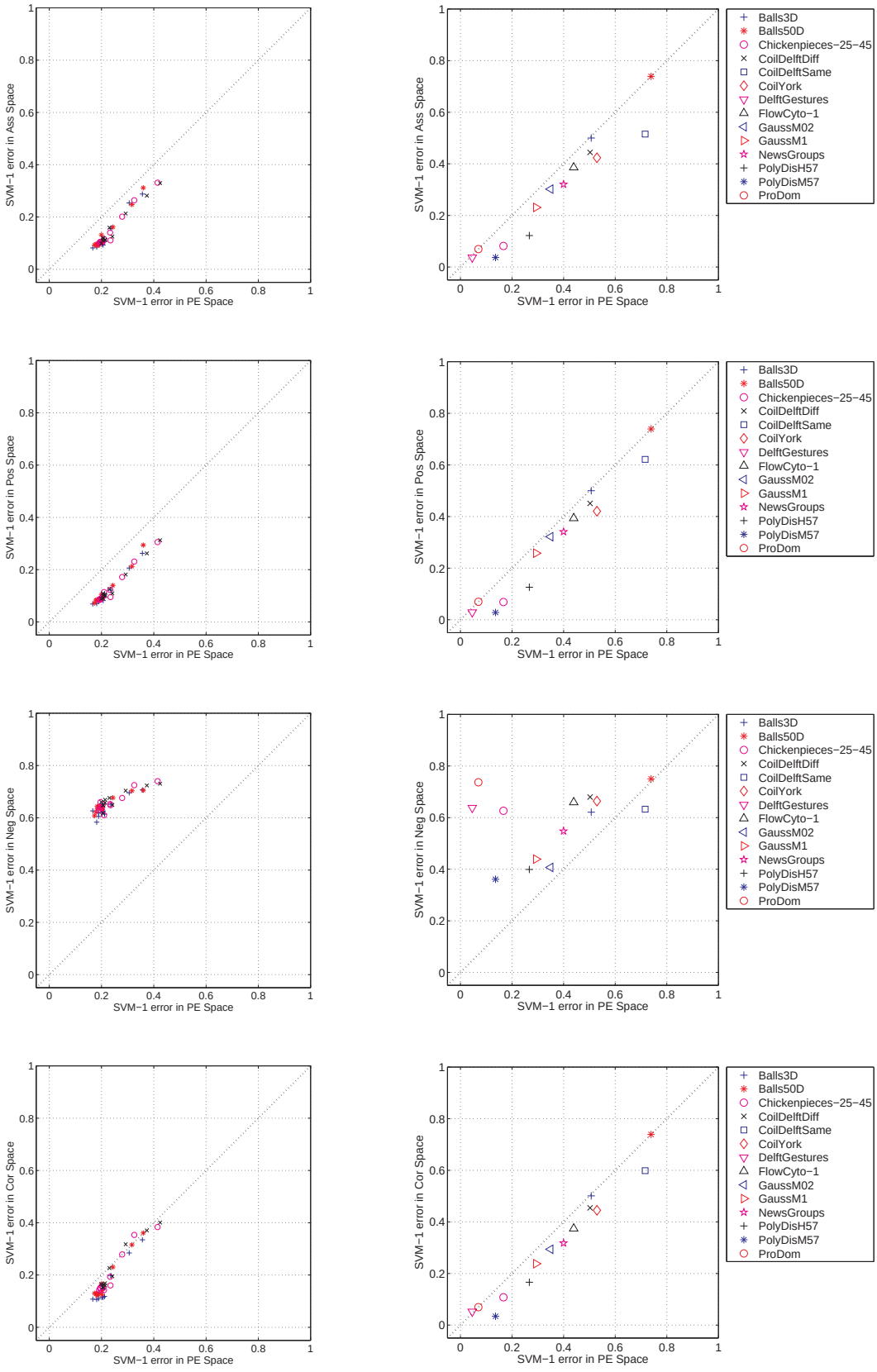


Figure 69: Scatterplots of the SVM-1 errors in various spaces versus the original PE Space (defined by the given dissimilarities) based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.3 Classification errors in the dissimilarity spaces

Here, we consider the performance of various classifiers in the original dissimilarity space (PE Dis Space) and its modified variants. The scatterplots below provide a comparison of the PE Dis Space and other modified dissimilarity spaces for the given classifier. Each sub-subsection is devoted to an individual classifier.

3.3.1 1-NN errors in the dissimilarity spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) In general, Neg Dis Space is not useful for the 1-NN rule. But, surprisingly, the Neg Dis Space is the only meaningful dissimilarity space for the Balls3D and Balls50D sets.

(b) In general, we have:
Cor Dis Space $\succeq_{1-\mathsf{NN}}$ Ass Dis Space $\approx_{1-\mathsf{NN}}$ Pos Dis Space $\approx_{1-\mathsf{NN}}$ PE Dis Space.

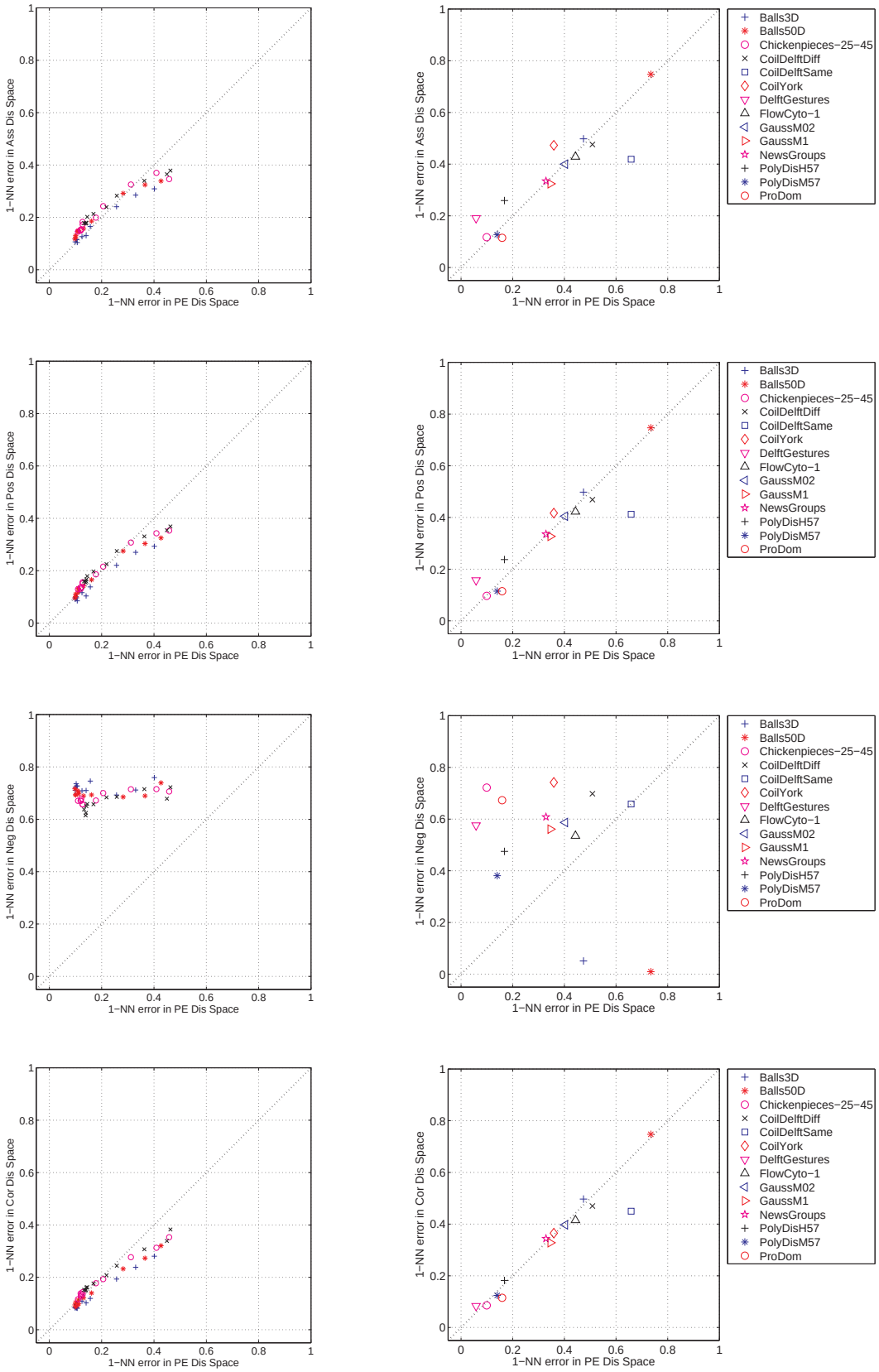


Figure 70: Scatterplots of the 1-NN errors in various spaces against the original dissimilarity space (PE Dis Space) based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.3.2 Parzen errors in the dissimilarity spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) In general, Neg Dis Space is not useful for the Parzen classifier. But, surprisingly, the Neg Dis Space is the only meaningful dissimilarity space for the Balls3D and Balls50D sets.

(b) The Parzen classifier in the PE Dis Space nearly always outperforms the Parzen classifier in the other dissimilarity spaces (but the Neg Dis Space). An exception holds for the metric CoilDelftSame data with a very small NEF (negative eigenfraction) of 0.027 on the whole set and ≈ 0.01 for a subset (of 50 objects per class) and a relatively large NEF (negative eigenratio) of 0.18 on the whole set.

(c) In general, we have:
PE Dis Space $\succeq_{\mathsf{Parzen}}$ Cor Dis Space $\succeq_{\mathsf{Parzen}}$ Pos Dis Space $\approx_{\mathsf{Parzen}}$ Ass Dis Space.

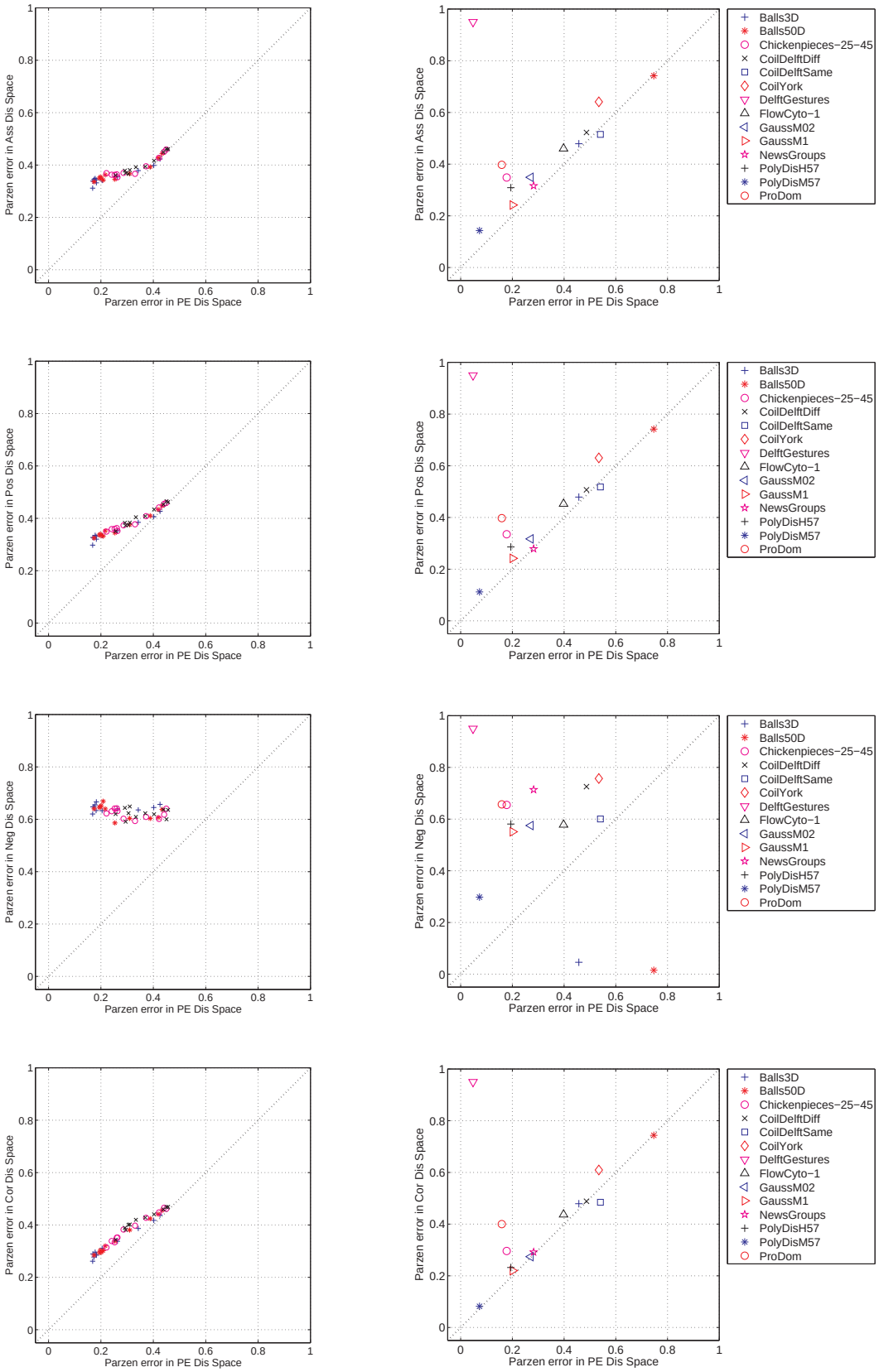


Figure 71: Scatterplots of the Parzen errors in various spaces against the original dissimilarity space (PE Dis Space) based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.3.3 Nearest Mean errors in the dissimilarity spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) In general, Neg Dis Space is not useful for the NM classifier. But, surprisingly, the Neg Dis Space is the only meaningful dissimilarity space for the Balls3D and Balls50D sets.

(b) The NM results for the 20-class DelftGestures data are really bad in all dissimilarity spaces, except for the PE Dis Space, where good performance is achieved.

(c) For the Chickenpieces data, the NM performance in the PE Dis Space is often similar or worse than in the three dissimilarity spaces: Pos, Ass and Cor Spaces.

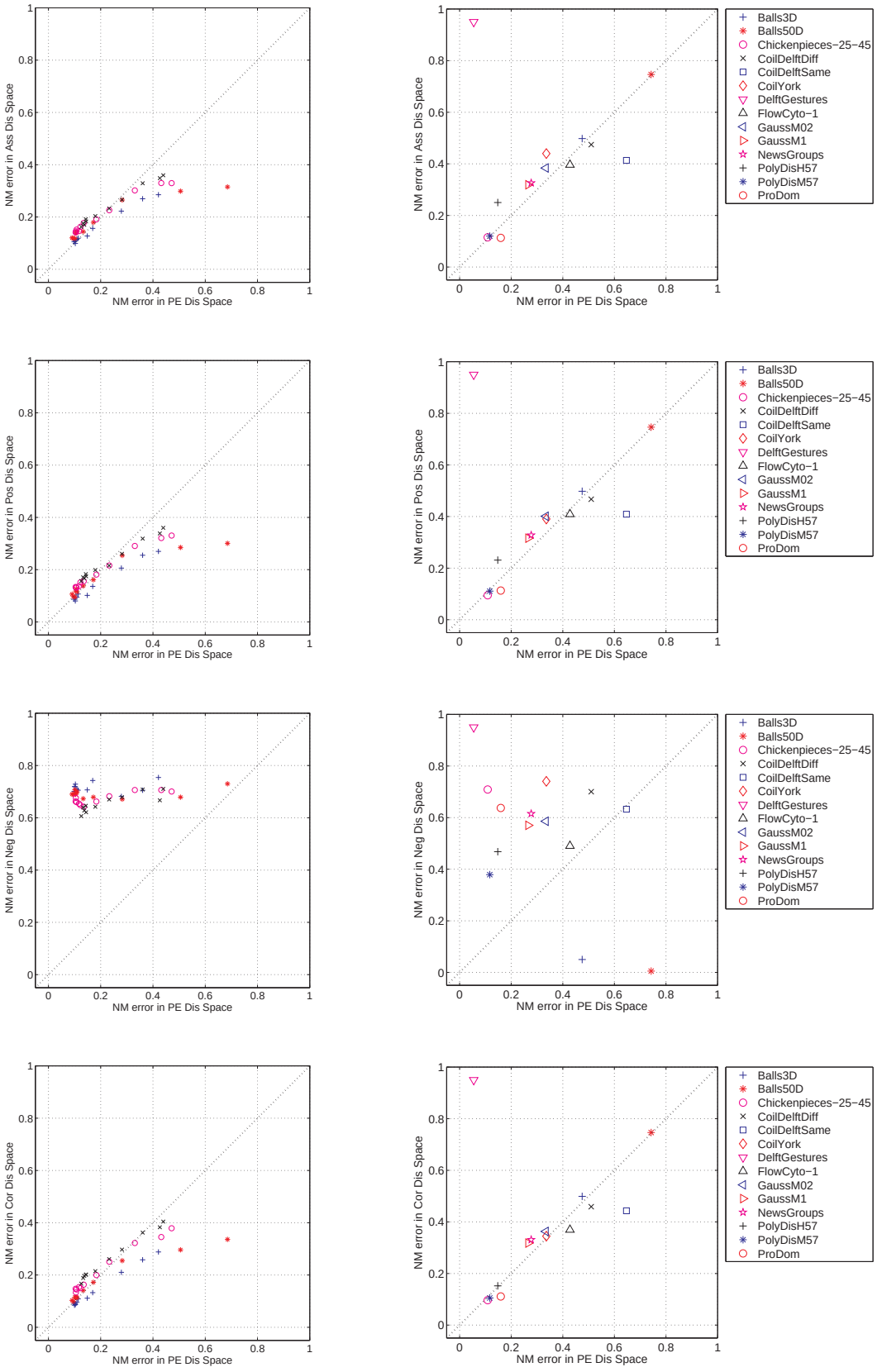


Figure 72: Scatterplot of the NM errors in various spaces against the original dissimilarity space (PE Dis Space) based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.3.4 Linear SVM errors in the dissimilarity spaces

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) In general, Neg Dis Space is not useful for the SVM-1 classifier. But, the Neg Dis Space is the only meaningful dissimilarity space for the Balls3D and Balls50D sets.

(b) For all Chickenpieces sets, SVM-1 in the PE Dis Space is outperformed by SVM-1 in the Ass, Pos and Cor Dis Spaces.

(c) The SVM-1 errors are usually larger in the PE Space than in the other three spaces: Ass Space, Pos Space and Cor Space. The results of SVM-1 are similar in all these spaces for datasets that are nearly Euclidean as judged by a small NER (negative eigenratio) of ≤ 0.03 on the whole set. These are the Balls3D, Balls50D, ProDom and DelftGestures datasets.

(d) In general, we have:
Pos Dis Space $\succeq_{\mathsf{SVM}-1}$ Ass Dis Space $\succeq_{\mathsf{SVM}-1}$ Cor Dis Space $\approx_{\mathsf{SVM}-1}$ PE Dis Space.

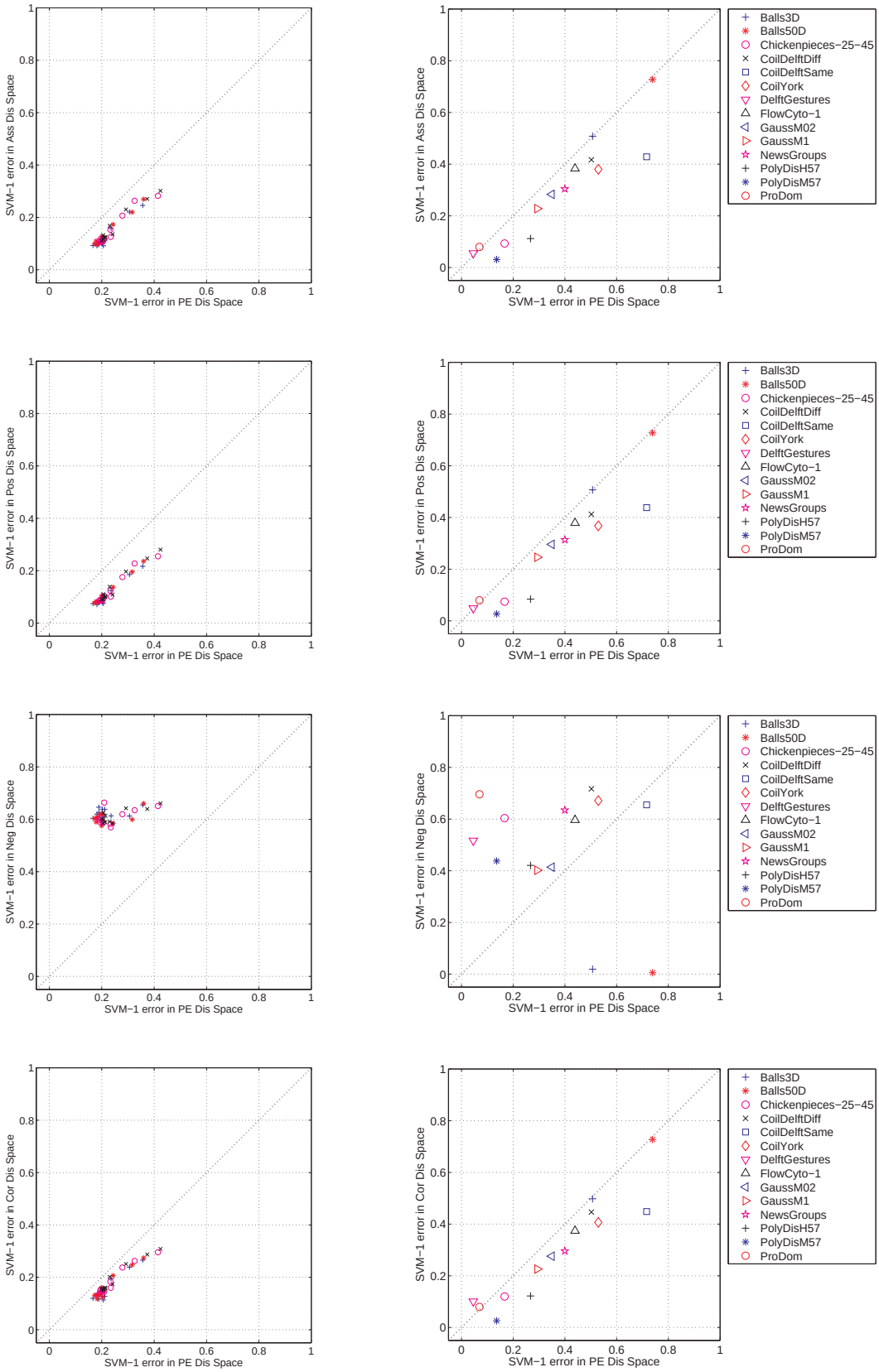


Figure 73: Scatterplots of the SVM-1 errors in various spaces against the original dissimilarity space (PE Dis Space) based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.4 Embedded PE-related spaces versus dissimilarity spaces

Here, we focus on the comparison between the PE-related spaces and the corresponding dissimilarity spaces. The following sub-subsections are devoted to specify classifiers.

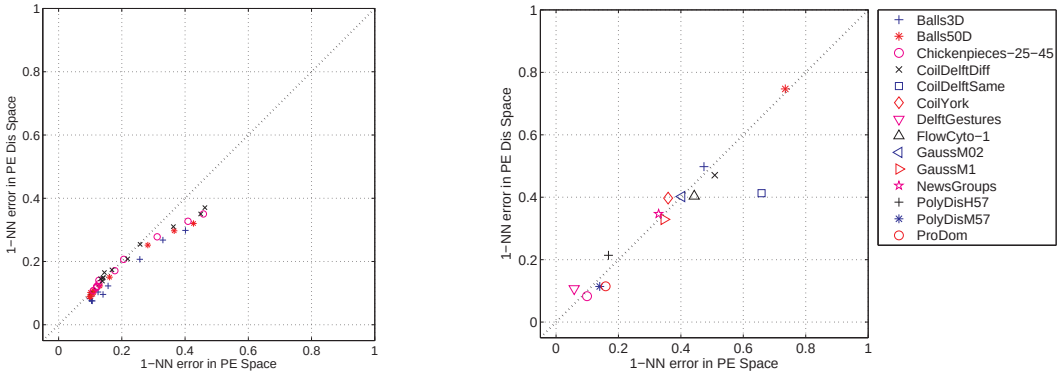
3.4.1 Embedded spaces and dissimilarity spaces compared by 1-NN errors

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) For the Chickenpieces sets the 1-NN performance in dissimilarity spaces is better than the 1-NN performance in the corresponding embedded space. Consequently we have:

PE Dis Space	$\succeq_{1-NN}$	PE Space
Ass Dis Space	$\succ_{1-NN}$	Ass Space
Pos Dis Space	$\succ_{1-NN}$	Pos Space
Neg Dis Space	$\succ_{1-NN}$	Neg Space
Cor Dis Space	$\succeq_{1-NN}$	Cor Space

(b) For all remaining datasets, the 1-NN performance in dissimilarity spaces is often similar or better than the 1-NN performance in the corresponding embedded spaces (PE space and its variants).



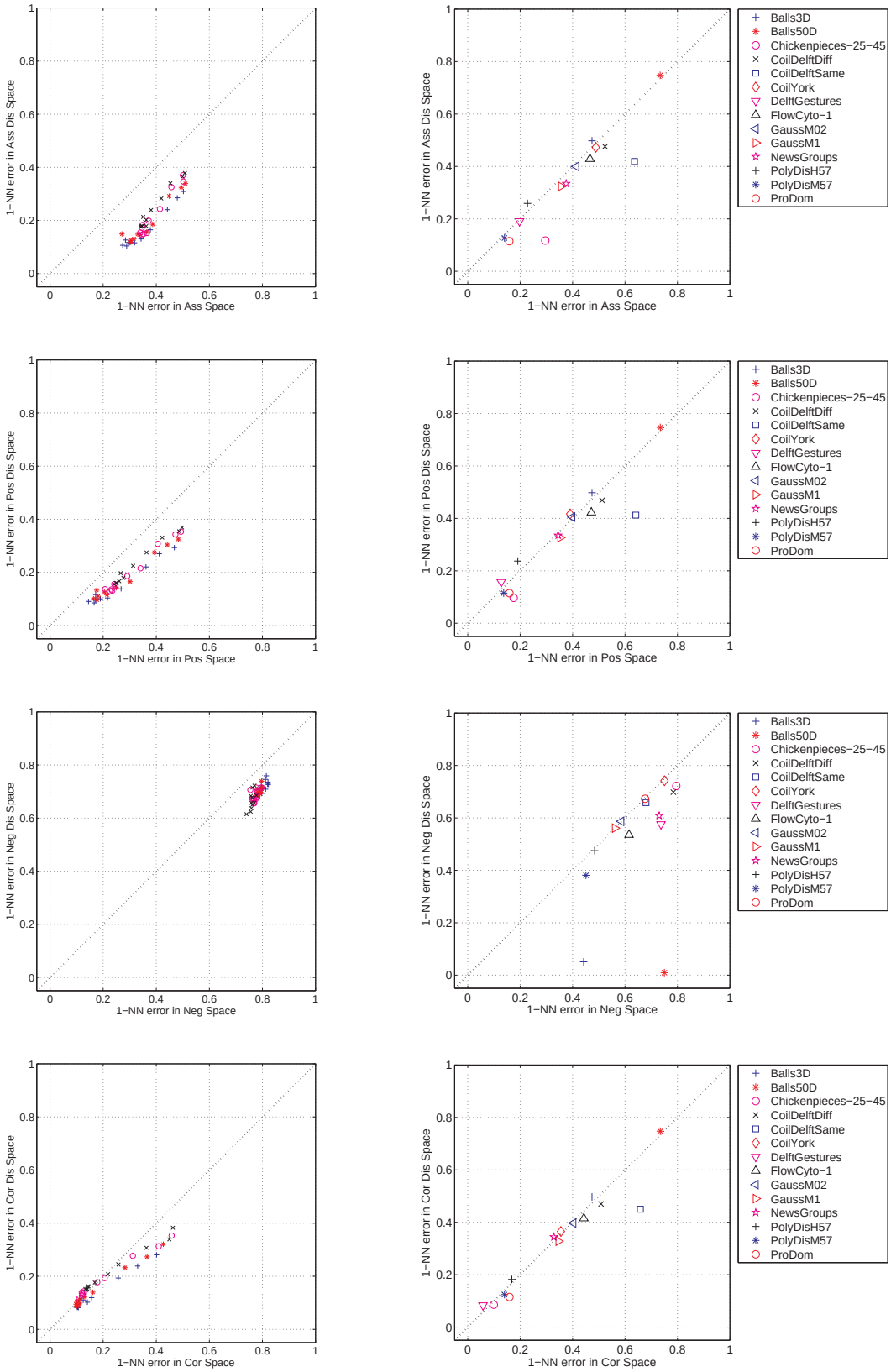


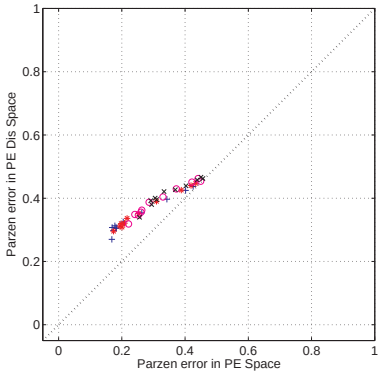
Figure 74: Scatterplots of the 1-NN errors comparing the results in the embedded spaces versus the results in the corresponding dissimilarity spaces. This is based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

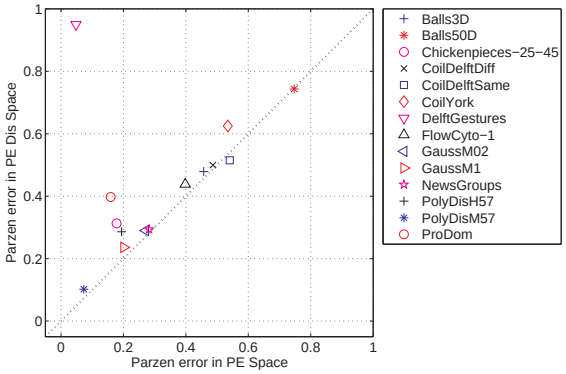
3.4.2 Embedded spaces and dissimilarity spaces compared by Parzen errors

The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

(a) In general (with three exceptions), the Parzen performance in dissimilarity spaces is better than the Parzen performance in the corresponding embedded space, except for the Neg Dis Space. Consequently we have we have:

$$\begin{aligned}
\text{PE Space} &\succ_{\text{Parzen}} \text{PE Dis Space} \\
\text{Ass Space} &\succeq_{\text{Parzen}} \text{Ass Dis Space} \\
\text{Pos Space} &\succeq_{\text{Parzen}} \text{Pos Dis Space} \\
\text{Cor Space} &\succ_{\text{Parzen}} \text{Cor Dis Space} \\
\text{Neg Dis Space} &\succ_{\text{Parzen}} \text{Neg Space}
\end{aligned}$$





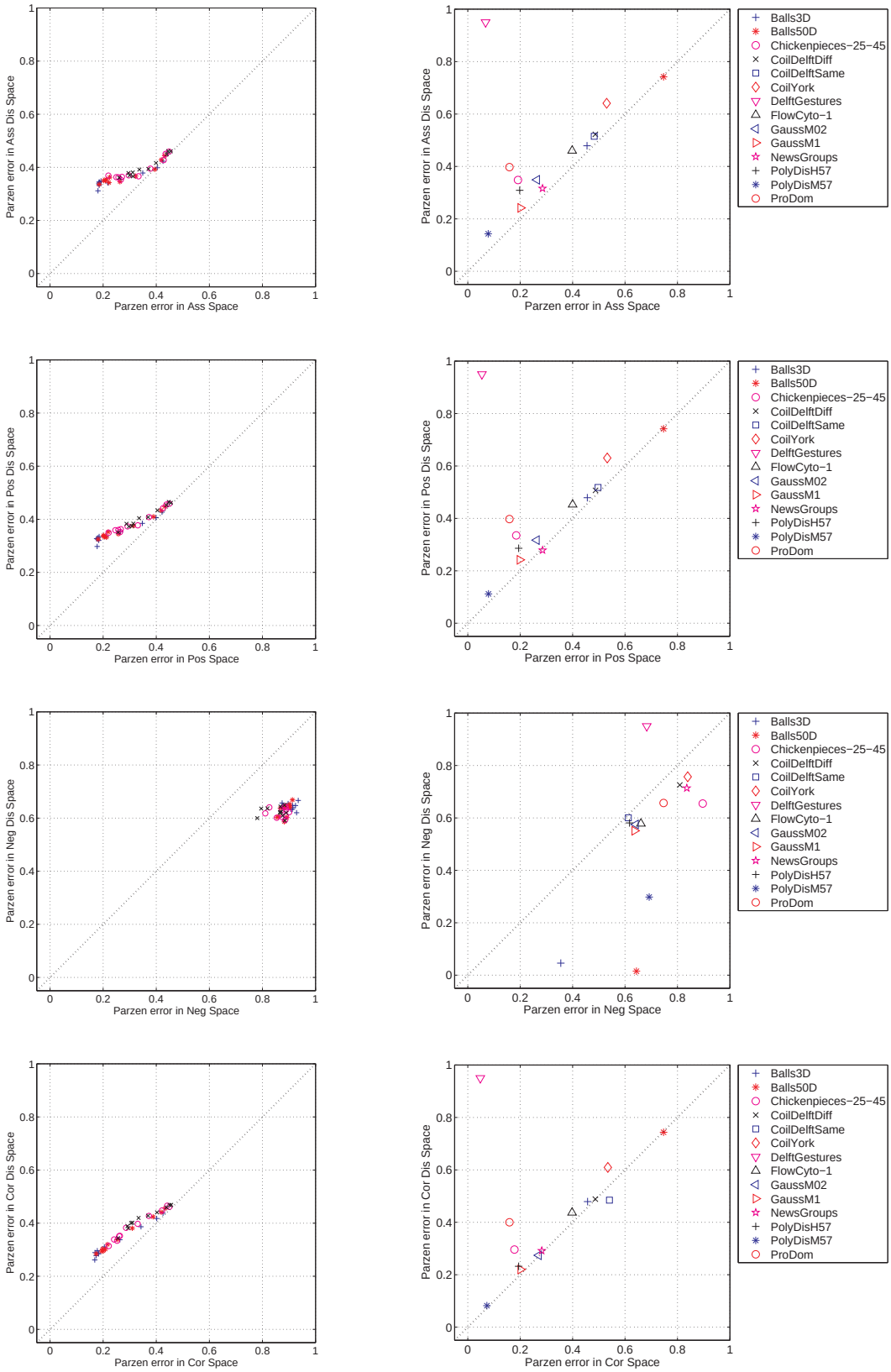


Figure 75: Scatterplots of the Parzen errors comparing the results in the embedded spaces versus the results in the corresponding dissimilarity spaces. This is based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.4.3 Embedded spaces and dissimilarity spaces compared by NM errors

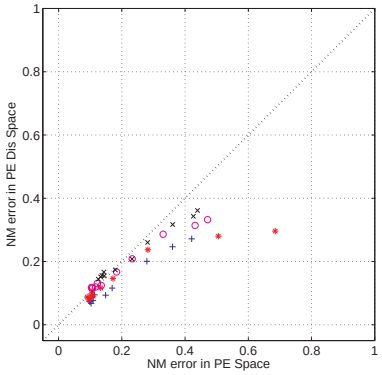
The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

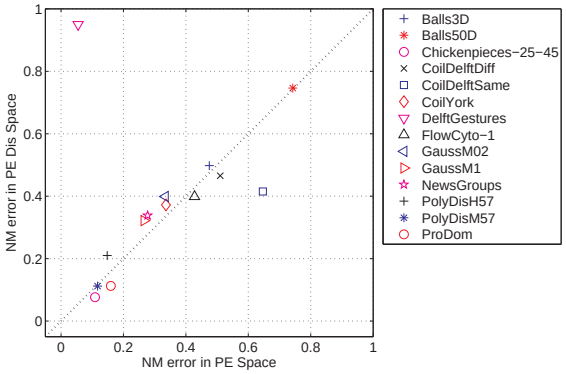
(a) For the Chickenpieces sets, we have:

PE Dis Space $\succeq_{NM}$ PE Space
Ass Dis Space $\succ_{NM}$ Ass Space
Pos Dis Space $\succ_{NM}$ Pos Space
Neg Dis Space $\succ_{NM}$ Neg Space

In addition, in half of the case, the NM classifier performs better in the Cor Dis Space than in the Cor Space.

(b) For all remaining datasets, the NM performance in the dissimilarity spaces tends to be similar or better than in the corresponding embedded spaces (PE space and its related spaces).





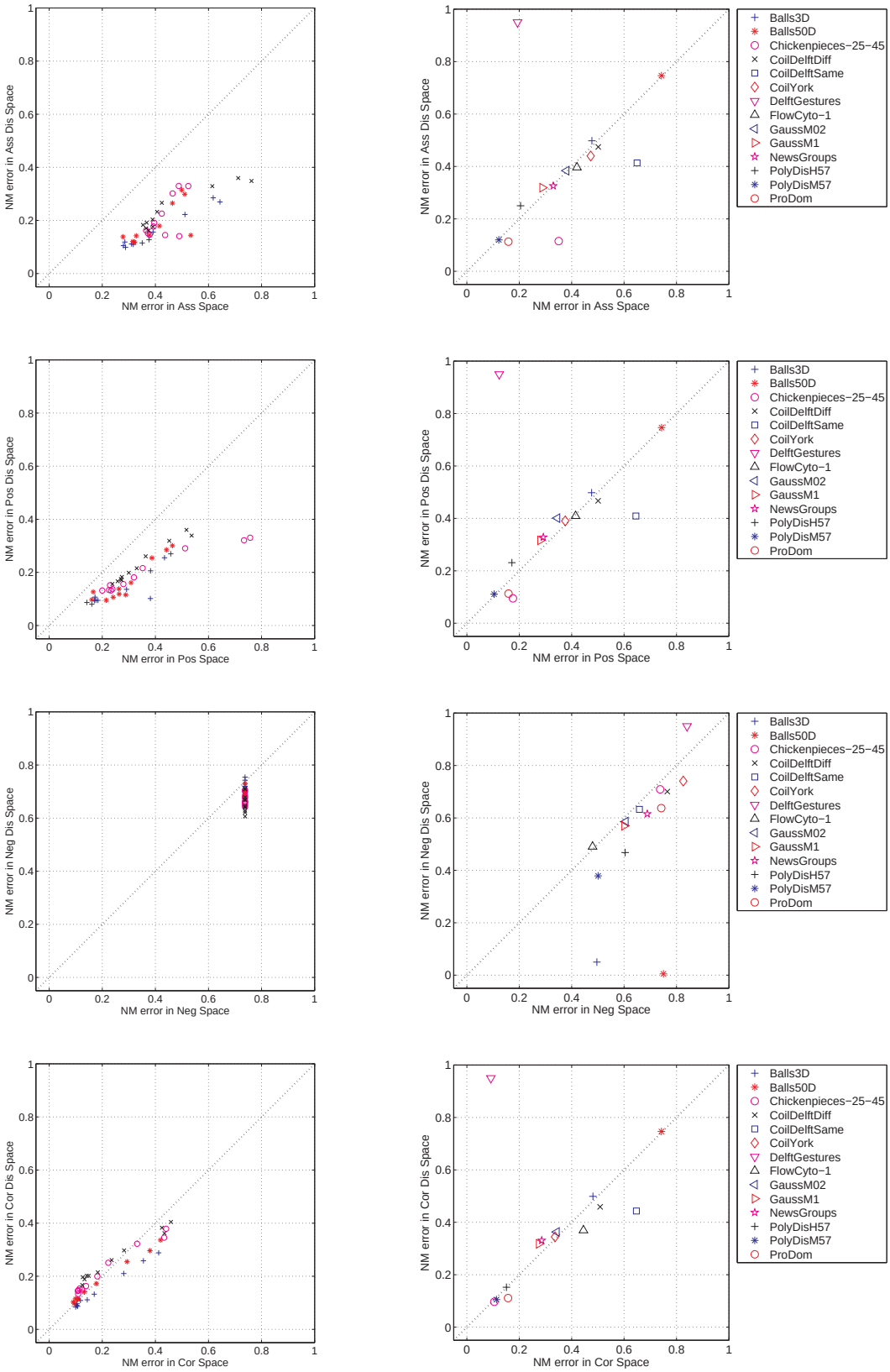


Figure 76: Scatterplots of the NM errors comparing the results in the embedded spaces versus the results in the corresponding dissimilarity spaces. This is based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

3.4.4 Embedded spaces and dissimilarity spaces compared by SVM-1 errors

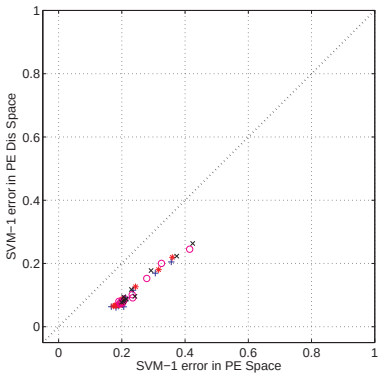
The results presented here are based on the averaged 2-fold crossvalidation errors. The following observations can be made:

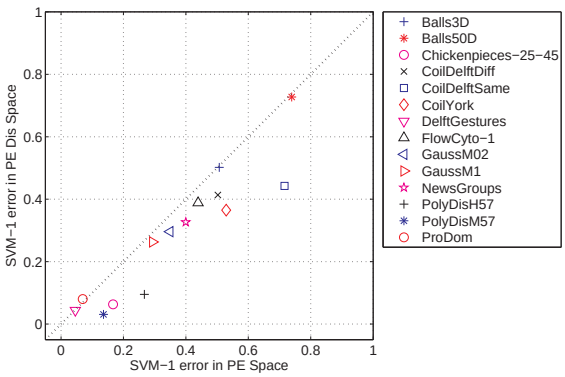
(a) In general, we have:

PE Dis Space $\succ_{\mathsf{SVM}-1}$ PE Space
Ass Dis Space $\succeq_{\mathsf{SVM}-1}$ Ass Space
Pos Dis Space $\succeq_{\mathsf{SVM}-1}$ Pos Space
Cor Dis Space $\succeq_{\mathsf{SVM}-1}$ Cor Space

For Chickenpieces datasets, the SVM-1 classifier performs similarly or only somewhat worse in the Pos and Ass Spaces than in the Pos and Ass Dis Spaces.

(b) The Neg Dis Space usually leads to better results than the Neg Space.





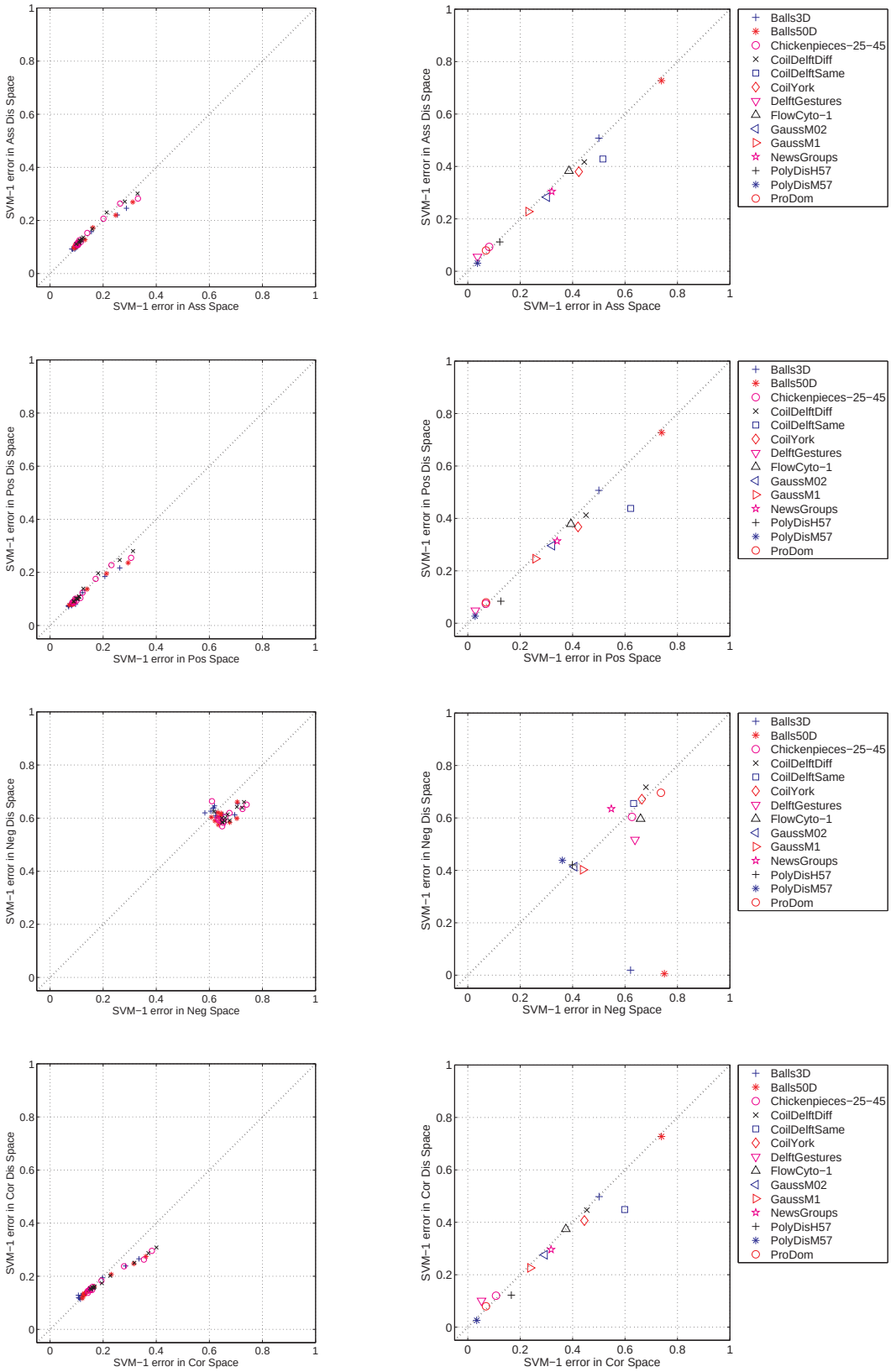


Figure 77: Scatterplots of the SVM-1 errors comparing the results in the embedded spaces versus the results in the corresponding dissimilarity spaces. This is based on 10 runs of 2-fold crossvalidation of 50 objects per class. The left plot shows all Chickenpieces datasets, while the right plot shows all remaining ones.

4 Final discussion

This report introduces a collection of proximity datasets to be used within the SIMBAD project. We always study dissimilarity data, so similarity matrices are appropriately transformed for this purpose. Each dissimilarity set is first made symmetric and scaled so that the average dissimilarity is 1 and then characterized by a number of specific indices as well as classification performance in ten vector spaces. These are five embedded spaces based on the Pseudo Euclidean (PE) space and its modified Euclidean versions, and five dissimilarity spaces corresponding to the five PE-related spaces. All vector spaces are constructed by using both (suitably scaled) training and test sets and no feature/prototype reduction is applied. The following classifiers are used: 1-NN (Nearest Neighbor), NM (Nearest Mean), Parzen and linear SVM. In order to compare over various problems and spaces, all experiments are performed on the subsets that consist of 50 examples per class taken from the original data. The results are presented as the averaged 2-fold classification errors. Hence, the performances are comparable by the way the experiments are set up.

There is much more to be studied in relation to the intrinsic dimension of the considered PE-related and dissimilarity spaces. Various feature selection methods can be applied in the PE-related spaces and various prototype reduction techniques can be used in the dissimilarity spaces. This may lead to improved classification performance, especially in the dissimilarity spaces. Such a study is left for future research as the main goal of this report is the presentation of the collected set of datasets.

Now, we will briefly summerize our main observations and findings. In relation to the datasets, our observations are:

(a) Concerning the Euclidean behavior of the dissimilarity matrices, a broad collection of dissimilarity datasets is available:

- Highly non-Euclidean examples: CatCortex, FlowCyto-data, GaussM02, PolyDisH57 or Zongker data,
- Moderately non-Euclidean examples: CoilDelftDiff, Chickenpieces-{30,35}-{45,60} data,
- Somewhat non-Euclidean examples: Chickenpieces-{5,7}-{90,120} data
- Nearly Euclidean or Euclidean examples: Balls3D, Balls50D, Protein or ProDom data.

(b) Concerning the metric properties, both metric and non-metric dissimilarity measures are present:

- Metric examples: Balls50D, GaussM1, PolyDisH57 and Chickenpieces-5,7-data.
- Slightly non-metric examples: Balls3D, DelftGestures, NewsGroups and ProDom data.
- Highly non-metric examples: FlowCyto-data, GaussM02, PolyDisM57 and Zongker data.

(c) The most non-metric and non-Euclidean data (as judged by both NER and NEF) is the Zongker dataset. The Euclidean (hence metric) data is the Balls50D set.

(d) If we focus on a set of dissimilarity datasets defined for the same problem but based on different measures, we will observe similar trends in the behavior of classifiers over various spaces. Examples include: Coil-data, Balls-data, Polygon-data, FlowCyto-data or Chickenpieces-data.

(e) The artificial Balls3D and Balls50D data are very peculiar. The good classification results are obtained only in negative spaces: Neg Space and Neg Dis Space. All other results are very bad.

In relation to the classification performance in various spaces, our main observations are:

(a) In general, the classifiers in the dissimilarity spaces usually perform similarly or better than the same classifiers in the corresponding embedded spaces.

(b) In general, the Pos Space and Pos Dis Space lead to similar or somewhat better classification results than in the Ass Space and Ass Dis Space.

(c) The Neg Space and Neg Dis Space are non-informative, expect for the Balls-data. It is the only space for these sets which leads to good results of all classifiers.

(d) There are examples when the original PE space is informative, i.e. when the 1-NN and NM classifiers perform similarly or better in the PE space than in the Pos Space. This holds for the CoilYork, PolyDisH57, PolyDisM57, FlowCyto-3 and many Chickenpieces datasets.

(e) Classifiers in the original PE space or original dissimilarity space perform sometimes very well. However, there are also situations when variants of these spaces allow to reach better results.

In relation to specific classifiers, our main observations are:

(a) Linear SVM in the PE space performs similar or worse than in other PE-related spaces (except for the Neg Space).

(b) Linear SVM often gives the best results, either in the PE-related spaces or in the dissimilarity spaces. Examples of the best SVM results are:

- SVM-1 in the PE Dis Space: CoilDelftDiff, all Chickenpieces-data
- SVM-1 in the PE Space and other variants: ProDom, DelftGestures (both nearly Euclidean)
- SVM1 in the Ass Space: PolyDisM57
- SVM-1 in the Pos Dis Space: PolyDisH57, FlowCyto-1
- SVM-1 in the Cor Dis Space: FlowCyto-3
- SVM-1 in the Neg Dis Space: Balls3D, Balls50D

(c) The 1-NN rule in dissimilarity spaces may lead to better results than in the embedded spaces, e.g. Zongker or Chickenpieces data.

(d) The Parzen classifier in dissimilarity spaces may lead to worse results than in the embedded spaces, e.g. DelftGestures, CoilYork and Chickenpieces sets.

(e) There are examples when the 1-NN, Parzen and NM classifiers perform better in the Cor Dis Space than in the PE Dis Space (original dissimilarity space). This holds for the GaussM02, CoilYork, Zongker and PolyDisH57 data.

(f) There are examples where all classifiers perform better in the Pos Dis Space than in the PE Dis Space (original dissimilarity space). This holds for the NewsGroups and PolyDisM57 data.

In conclusion, we state that we have collected a broad set of datasets that differ in many characteristics. They may thereby constitute a good material for further study of the analysis and application of dissimilarity data.

A Experiments and Software

All experiments in this report are performed by the Matlab packages DisTools and PRTools. PRTools is a general toolbox for pattern recognition which can be obtained through http://prtools.org/. DisTools is a collection of tools written on top of PRTools especially designed for dealing with (dis)similarity data. It contains many routines used by the authors of this report and their students and colleagues for analyzing such data.

During the preparation of this report DisTools has been extended and upgraded. The most important subset (which we gave version number 1.0) is made public by this report. It includes all routines used in the creation of this report.

On the next page the Contents file is listed. The pages thereafter show the two top routines that are used for computing the dataset characteristics, the curves and the crossvalidation experiments listed for each dataset. They make use of the LIBSVM package that is not distributed with PRTools or DisTools. In order to run our scripts users should first obtain the LIBSVM package.

```
  DisTools Table of Contents
  Version 1.0 30-Sep-2009

 This Matlab toolbox for the analysis of dissimilarity data works only
 if also the pattern recognition toolbox PRTools is available.
 See http://prtools.org

   E. Pekalska,  ela.pekalska@googlemail.com, University of Manchester
   R.P.W. Duin,  r.duin@ieee.org, Delft University of Technology

Characterization of dissimilarity matrices
------------------------------------------
CHECKEUCL    Check whether a square dissimilarity matrix has a Euclidean behavior
CHECKTR      Check whether a square dissimilarity matrix obeys triangle inequality
CHARDMAT     Fiand several characteristic of (dis)similarity data
CORRTR       Correct a square dissimilarity matrix to obey the triangle inequality
DISCHECK     Dissimilarity matrix check
DISNORM      Normalization of a dissimilarity matrix
DISSTAT      Basic statistics of the dissimilarity matrix
ISSQUARE     Check whether a matrix is square
ISSYM        Check whether a matrix is symmetric
ASYMMETRY    Compute asymmetry of dissimilarity matrix
NNE          Leave-one-out Nearest Neighbor error on a dissimilarity matrix
NNERR        Exact expected NN error from a dissimilarity matrix

Dissimilarity Measures
----------------------------------------------
COSDISTM     Distance matrix based on inner products
EUDISTM      Euclidean distance matrix
HAMDISTM     Hamming distance matrix between binary vectors
HAUSDM       Hausdorff and modified Hausdorff distance between datasets of image blobs

Transformations
DISSIMT      Fixed DISsimilarity-SIMilarity transformation
MAKESYM      Make a matrix symmetric
PE_EM        Pseudo-Euclidean embedding (includes Classical Scaling as a special case)

Classification in Pseudo-Euclidean Space and indefinite kernels
-----------------------------------------------
SETSIG       Set PE signature for mappings or datasets
GETSIG       Set PE signature for mappings or datasets
ISPE_DATASET Test dataset for PE signature setting
ISPE_EM      Test mapping for PE signature setting
PE_DISTM     Square pseudo-Euclidean distance between two datasets
PE_KERNELM   Compute kernel in PE space
PE_MTIMES    Matrix multiplication (inner product) in PE space
PE_PARZENC   Parzen classifier in PE space
PE_KNNC      KNN classifier in PE space
PE_NMC       Nearest mean classifier in PE space
PE_EM        Pseudo-Euclidean linear embedding
PLOTSPECTRUM Plot spectrum of eigenvalues
```

```
Routines supporting in learning from dissimilarity matrices
------------------------------------------------------------------------
CROSSVALD    Cross-validation error for dissimilarity representations
DISSPACES    Compute various spaces out of a dissimilarity matrix
GENDDAT      Generate random training and test sets for dissimilarity data
GENREP       Generate a representation set
GENREPI      Generate indices for representation, learning and testing sets
SELCDAT      Select Class Subset from a Square Dissimilarity Dataset
PROTSELFD    Forward prototype selection
DLPC         LP-classifier on dissimilarity (proximity) data
KNNDC        K-Nearest Neighbor classifier for dissimilarity matrices
PARZENDDC    Parzen classifier for dissimilarity matrices
TESTKD       Test k-NN classifier for dissimilarity data
TESTPD       Test Parzen classifier for dissimilarity data

EXAMPLES
--------
CROSSVALD_EX Crossvalidation of several classifiers
```

The routine `chardmat` is used to compute the dataset characteristics as listed in section 2.

```
%CHARDMAT Characterization of a square, labeled dissimilarity matrix
%
%  [C,D_OUT] = CHARDMAT(D)
%
% Characterizes a square (dis)similarity dataset D.
% D_OUT is the symmetric, normalized dissimilarity dataset D. If D is
% a similarity dataset it is converted to dissimilarities first.
% The following fields are returned in the structure C.
% name       - dataset name as stored by PRTools or read command
% desc       - dataset description as stored by read command
% link       - web links as stored by read command
% ref        - references as stored by read command
% asym       - asymmetry, 2*|D-D'|./(|D|+|D'|)
% size       - number of objects
% classes    - number of classes
% clsizes    - vector with class sizes
% type       - 'dis' for dissimilarities, 'sim' for similarities
%
%         all following items are computed for a transformation
%         of D by MAKESYM and DISNORM (make average distance 1),
%         similarities are first transformed into dissimilarities
%         by d(i,j) = sqrt(d(i,j) + d(j,j) - d(i,j) - d(j,i))
%
% within_mean- average within class dissimilarity
% between_mean- average between class dissimilarity
% pe_mapping - Pseudo-Euclidean mapping as computed by PE_EM
% signature  - 2 component vector with # of positive and negative
%              eigenvalues obtained during the PE embedding
% eigenvalues- the eigenvalues obtained during the PE embedding,
%              see PE_EM for their ranking
% nef         - Negative Eigen Fraction (sum of absolute negative
%               eigenvalues divided by sum of all absolute eigenvalues)
% ner         - Negative Eigen Ratio (- largest negative eigenvalue
%               divided by largest positive eigenvalue)
% trineq      - number of triangle inequality violations
%
%         the following characteristics refer to a set of five spaces:
%         - Pseudo-Euclidean space based on a full embedding. Distances
%           in this space are identical to D.
%         - Associated space, the same vector spaces, but now treated as
%           an Euclidean space
%         - Positive space based on the positive eigenvalues only
%         - Negative space based on the negative eigenvalues only
%         - Corrected space based on an embedding of sqrt(D.^2+2*Lmin)
%           in which Lmin is the absolute values of the largest negative
%           eigenvalue. The result is a proper Euclidean space.
```

```
%
% loo_a       - leave-one-out nearest neighbor errors for all five
%               embedded spaces
% loo_d       - leave-one-out nearest neighbor errors for the dissimilarity
%               spaces related to the above five embedded spaces
% lcurve_a    - nearest neighbor learning curves for the five embedded
%               spaces
% lcurve_a    - nearest neighbor learning curves for the five dissimilarity
%               spaces
% anames      - names of the five embedded spaces, useful for annotation
% dnames      - names of the five dissimilarity spaces

function [c,d] = chardmat(d)

  isdataset(d);
  datname = getname(d);
  discheck(d,[],1);
  m = size(d,1);
  nclass = getsize(d,3);

  c.name = datname;
  c.desc = getuser(d,'desc');
  c.link = getuser(d,'link');
  c.ref = getuser(d,'ref');
  c.asym = asymmetry(d);
  c.size = m;
  c.classes  = nclass;
  c.clsizes  = classsizes(d);

  if discheck(d);
    c.type = 'dis';
  else
    c.type = 'sim';
    d = dissimt(d,'sim2dis');
  end

  % we now have a dissimilarity matrix with positive distances

  d = makesym(d);  % make it symmetric now
  d = d*disnorm(d);
  uc = zeros(1,nclass);
  for j=1:c.classes
    nj = c.clsizes(j);
    dj = +selcdat(d,j);
    uc(j) = sum(dj(:))/(nj*(nj-1));
  end
  c.within_mean = uc*(c.clsizes'.^2-c.clsizes')/(sum(c.clsizes.^2) - m);

  ud = (m*(m-1) - uc*(c.clsizes'.^2-c.clsizes')) / (m*(m-1) - sum(c.clsizes.^2) + m);
  c.between_mean = ud;
```

```
  [nef,ner,w] = checkeucl(d);
  c.pe_mapping  = w;
  c.signature   = getsig(w);
  c.eigenvalues = getdata(w,'eval');
  c.nef      = nef;
  c.ner      = ner;
  c.trineq   = checktr(d);

  [A D] = disspaces(d,w);
  nspaces = length(A);
  c.loo_a = zeros(1,nspaces); % LOO NN errors embedding spaces
  c.loo_d = zeros(1,nspaces); % LOO NN errors dis spaces
  c.lcurve_a = cell(1,nspaces);  % NN Learning curves embedding spaces
  c.lcurve_d = cell(1,nspaces);  % NN Learning curves dis spaces
    c.anames = cell(1,nspaces);    % names embedded spaces
    c.dnames = cell(1,nspaces);    % names dis spaces
  for j=1:nspaces
    c.loo_a(j) = nne(Dj);
    c.loo_d(j) = nne(distm(Dj));
    c.lcurve_aj = nnerr(Dj);
    c.anamesj = getname(Aj);
    c.lcurve_dj = nnerr(distm(Dj));
    c.dnamesj = getname(Dj);
  end

return
```

The routine crossvald_ex is used to perform the crossvalidation experiments for the 4 classifiers in the 10 spaces as presented in the tables of section 2.

```
%CROSSVALD_EX Crossvalidation of several classifiers for some spaces
%
%  [EA,ED] = CROSSVALD_EX(D,TRAINSIZE,NFOLD,RUNS)
%
% Selects at random training sets of size TRAINSIZE per class from the
% square dissimilarity dataset D and runs NFOLD crossvalidation on them.
% The total procedure is run RUNS times and results are stores in
% EA for PE spaces and ED for dissimilarity spaces. See DISSPACES.
% The following classifiers are used:
% Nearest Neighbor, PARZEN, Nearest Mean and SVM (by LIBSVM)

function [erra,errd] = crossvald_ex(d,m,n,runs)
    if nargin < 4, runs = 1; end
    if nargin < 3, n = 2; end
    if nargin < 2, m = 50; end
    discheck(d);
    d = d*disnorm(d);
    csizes = classsizes(d);
    nclass = length(csizes);
    if any(csizes <= m*1.2)
        error('Some classes are too small for desired training set sizes');
    end

    nspaces = 5; % number of spaces we get fro disspaces
    we = knnc([],1),nmc,parzenc,libsvc([],[],100);  % Eucl classifiers
    wp = pe_knnc([],1),pe_nmc,pe_parzenc,libsvc([],pe_kernelm,100); % PE classf
    erra = zeros(length(we),nspaces,runs);
    errd = zeros(length(we),nspaces,runs);
    it = 0;
    for j=1:runs
        randreset(j);
        [A D] = disspaces(genddat(d,m*ones(1,nclass)));
        for i=1:nspaces
            it = it+1;
            randreset(j);
            ba = Ai;
            if ispe_dataset(ba)
                era = crossval(ba,wp,n);
            else
                era = crossval(ba,we,n);
            end
            randreset(j);
            bd = Di;
            erd = crossval(bd,we,2);
            erra(:,i,j) = era(:);
            errd(:,i,j) = erd(:);
        end
    end
return
```