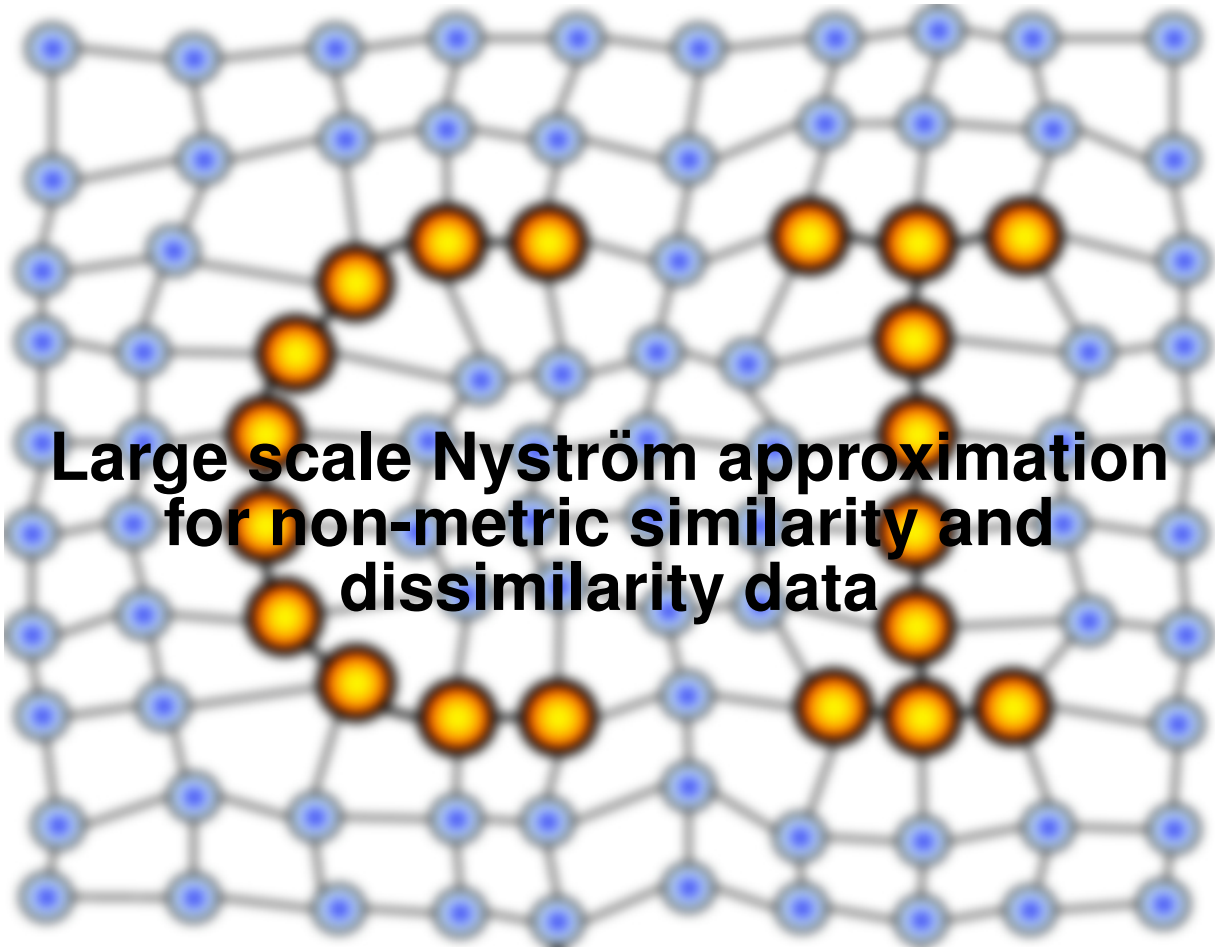
MACHINE LEARNING REPORTS



# Large scale Nyström approximation for non-metric similarity and dissimilarity data

Report 03/2013
Submitted: 01.09.2013
Published: 11.09.2013

F.-M. Schleif*
(1) Theoretical Computer Science, University of Bielefeld, Universitätsstrasse 21-23, 33615 Bielefeld, Germany
corresponding author: *email: fschleif@techfak.uni-bielefeld.de*

**Abstract**

Processing large proximity data such as kernel matrices often includes approximation techniques like the Nyström approximation. Thereby the distance calculations are done on an approximated kernel matrix. This operation is based on the calculation of a pseudo-inverse matrix which itself can become costly for larger data sets. Recent work in this field extended the original Nyström approximation by a randomized subspace technique making large scale problems accessible for positive semi-definite proximity data. Domain specific proximity measures, employed e.g. in alignment algorithms in bio-informatics, are often used to compare complex data objects and to cover domain specific data properties. Lacking an underlying vector space, data are given as pairwise (dis-)similarities and the obtained proximity matrices are typically non-metric. In this contribution we analyse the large scale Nyström approximation for non-psd proximity data including dissimilarities and how it can be used to convert similarities to dissimilarities and vice versa. We provide an approach to explicit control the eigenvalue correction on the approximated matrices.

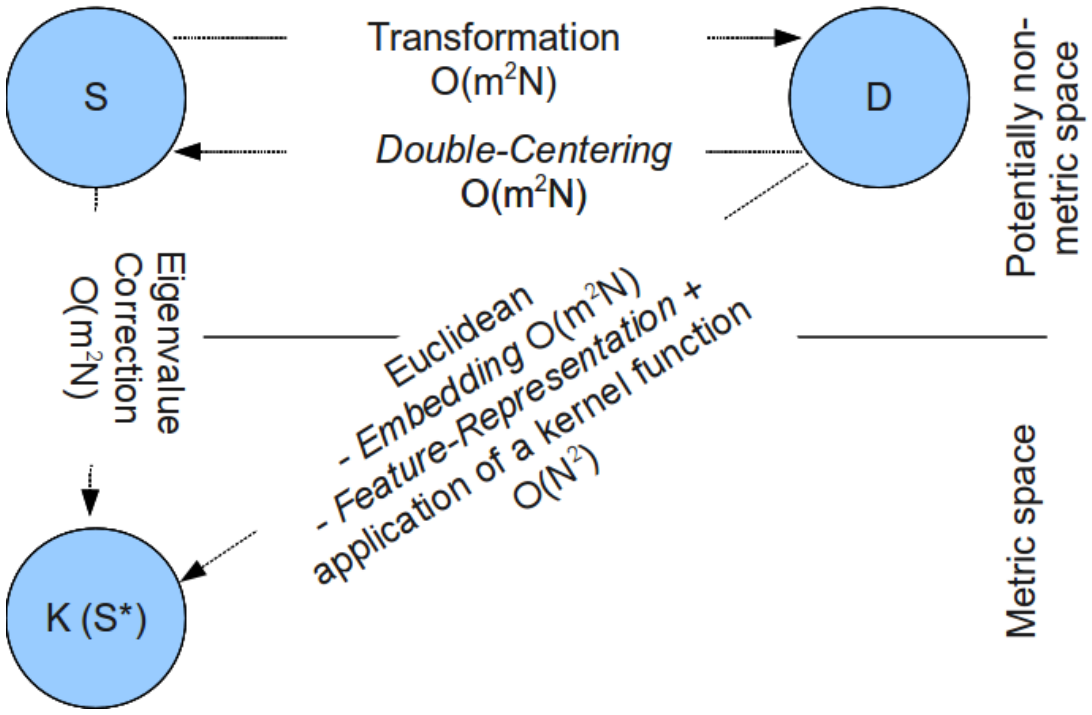



Figure 1: Left: Traveling in similarity and dissimilarity spaces at linear costs using the approach of [18]. The standard approach has in general a of complexity $O(N^2)-O(N^3)$.

# 1 Introduction

In many application areas such as bioinformatics, different technical systems, or the web, electronic data is getting larger and more complex in size and representation, using *domain specific* (dis-)similarity measures as a replacement or complement to Euclidean measures. Many classical machine learning techniques, have been proposed for Euclidean vectorial data. However, modern data are often associated to dedicated structures which make a representation in terms of Euclidean vectors difficult: biological sequence data, text files, XML data, trees, graphs, or time series [15, 12, 1] are of this type. These data are inherently compositional and a feature representation leads to information loss. As an alternative, a dedicated dissimilarity measure such as pairwise alignment, or kernels for structures can be used as the interface to the data. In such cases, machine learning techniques which can deal with pairwise similarities or dissimilarities have to be used [16]. Native methods for the analysis of dissimilarity data have been proposed in [16, 9, 8], but are widely based on non-convex optimization schemes and with quadratic to linear memory and runtime complexity. The analysis of proximity matrices can either be based on a similarity representation (or a kernel) or a dissimilarity representation (distances). It is possible to convert the one into the other representation and vice versa by either using a distance calculation based on similarities or a double centering to convert dissimilarities into its alternative representation of similarities. Since most analysis methods rely on metric input data of the underlying similarities, different preprocessing approaches have been analyzed to correct non-metric or non psd similarity matrices [1], typically based on eigenvalue corrections. The transformation between the different representations as

well as the correction approaches have typically quadratic or cubic costs. In [18] the author proposed a method for the transformation between the different representations including such an eigenvalue correction with linear costs see Figure 1. This approach was based on the Nyström approximation [20]. This transformation represents a given matrix by a small number of so called landmark points and their relation to the remaining data points, as detailed later on. As an inherent step a quadratic matrix of these landmark proximities is used. The number of landmarks and their specific selection from the data has been discussed in [21]. Basically the selection strategy is not so important as long as a sufficiently large number of landmarks can be drawn i.i.d. from the data. This can become very costly for larger datasets, where either the accuracy of the approximation suffers, due to a to small number of landmarks, or the approximation costs raise if a sufficiently large number of landmarks is used. In [14] a new approach for the calculation of the Nyström approximation for *psd* matrices was proposed using a random projection technique. As shown in [14] this strategy is very effective to keep high accuracy of the matrix approximations also for very large data sets. Motivated by these promising results we will derive an extension of our former approach published in [18] using similar strategies. Our objective is to provide a method of linear complexity to tranform similarities into dissimilarities and vice versa including potential eigenvalue corrections to make the problem psd. Especially for metric dissimilarities the approach keeps the known guarantees like generalization bounds (see e.g. [3]) while for non-psd data corresponding proofs are still open, but our experiments are promising. The paper is organized as follows. First we give a short review of previous work of the author proposed in [18]. Then we review the subspace Nyström approximation proposed in [14] and how it can be linked to our former work. Experimental results show the effectiveness of the proposed approach.

# 2 Transformation techniques for dissimilarity data

Let $\mathbf{v}_j \in \mathbb{V}$ be a set of objects defined in some data space, with $|\mathbb{V}| = N$. We assume, there exists a dissimilarity measure such that $\mathbf{D} \in \mathbb{R}^{N \times N}$ is a dissimilarity matrix measuring the pairwise dissimilarities $D_{ij} = d(\mathbf{v}_i, \mathbf{v}_j)$ between all pairs $(\mathbf{v}_i, \mathbf{v}_j) \in \mathbb{V}$. Any reasonable (possibly non-metric) distance measure is sufficient. We assume zero diagonal $d(\mathbf{v}_i, \mathbf{v}_i) = 0$ for all $i$ and symmetry $d(\mathbf{v}_i, \mathbf{v}_j) = d(\mathbf{v}_j, \mathbf{v}_i)$ for all $i, j$.

## 2.1 Analyzing dissimilarities by means of similarities for small $N$

For every dissimilarity matrix $\mathbf{D}$, an associated similarity matrix $\mathbf{S}$ is induced by a process referred to as double centering with costs of $\mathcal{O}(N^2)$[16]:

$$\mathbf{S} = -\mathbf{JDJ}/2 \quad (1)$$

$$\mathbf{J} = (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/N) \quad (2)$$

with identity matrix $\mathbf{I}$ and vector of ones $\mathbf{1}$. $\mathbf{D}$ is Euclidean if and only if $\mathbf{S}$ is positive semi-definite (psd). This means, we do not observe negative eigenvalues in the eigenspectrum of the matrix $\mathbf{S}$ associated to $\mathbf{D}$.

Many classification techniques have been proposed to deal with such psd kernel matrices $\mathbf{S}$ implicitly such as the support vector machine (SVM). In this case, preprocessing is *required to guarantee* psd. In [1] different strategies were analyzed to obtain

valid kernel matrices for a given similarity matrix $\mathbf{S}$, most popular are: *clipping, flipping, shift correction, vector-representation*. The underlying idea is to remove negative eigenvalues in the eigenspectrum of the matrix $\mathbf{S}$ .

Assuming we have a symmetric similarity matrix $\mathbf{S}$, it has an eigenvalue decomposition $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, with orthonormal matrix $\mathbf{U}$ and diagonal matrix $\mathbf{\Lambda}$ collecting the eigenvalues. In general, $p$ eigenvectors of $\mathbf{S}$ have positive eigenvalues and $q$ have negative eigenvalues, $(p, q, N-p-q)$ is referred to as the *signature*.

The *clip*-operation sets all negative eigenvalues to zero, the *flip*-operation takes the absolute values, the *shift*-operation increases all eigenvalues by the absolute value of the minimal eigenvalue.

The corrected matrix $\mathbf{S}^*$ is obtained as $\mathbf{S}^* = \mathbf{U}\mathbf{\Lambda}^*\mathbf{U}^\top$, with $\mathbf{\Lambda}^*$ as the modified eigenvalue matrix using one of the above operations. The obtained matrix $\mathbf{S}^*$ can now be considered as a psd kernel matrix $\mathbf{K}$ suitable e.g. as an input for a kernel clustering or classifier. In [11] an alternative of a classifier directly based on a dissimilarity matrix $D$ was proposed. The main model parameters are so called prototypes $\mathbf{w}$ (similar to cluster centers) which are points constructed as a linar combination of the original data. The basic idea is an implicit computation of distances $d(\cdot,\cdot)$ during the model calculation based on the dissimilarity matrix $\mathbf{D}$ using weights $\alpha$:

$$d(\mathbf{v}_i, \mathbf{w}_j) = [\mathbf{D} \cdot \alpha_j]_i - \frac{1}{2} \cdot \alpha_j^\top \mathbf{D} \alpha_j \tag{3}$$

This approach avoids the need for an accessible vector space, similar like the kernel trick in the context of kernel machines. A more detailed discussion, including alternative approaches is given in [18]. A schematic view of the relations between $\mathbf{S}$ and $\mathbf{D}$ and its transformations [1] using strategies as proposed in [18] and discussed in more detail in the following is shown in Figure 1.

The methods discussed before are suitable for data analysis based on similarity or dissimilarity data where the number of samples $N$ is rather small, e.g. scales by some thousand samples. For larger $N$ only for *metric, similarity data* (valid kernels) efficient approaches have been proposed before, e.g. low-rank linearized SVM [22] or the Core-Vector Machine (CVM) [19].

Now we briefly review concepts already proposed by the author in [8, 18] how potentially non-metric similarities and dissimilarities can be approximated by the Nyström approximation and a coupling with double centering for dissimilarity data.

# 3 Nyström approximation

The aforementioned methods depend on the similarity matrix $\mathbf{S}$ or dissimilarity matrix $\mathbf{D}$, respectively. For kernel methods and more recently for prototype based learning the usage of the Nyström approximation is a well known technique to approximate both types of matrices to obtain effective learning algorithms [20, 8].

## 3.1 Nyström approximation for similarities

The Nyström approximation technique has been proposed in the context of kernel methods in [20] with related proofs and bounds given in [3]. Here, we give a short

[1] Transformation equations are given also in the following sections.

review of this technique. One well known way to approximate a $N \times N$ Gram matrix, is to use a low-rank approximation. This can be done by computing the eigendecomposition of the kernel $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U}$ is a matrix, whose columns are orthonormal eigenvectors, and $\mathbf{\Lambda}$ is a diagonal matrix consisting of eigenvalues $\mathbf{\Lambda}_{11} \geq \mathbf{\Lambda}_{22} \geq ... \geq 0$, and keeping only the $m$ eigenspaces which correspond to the $m$ largest eigenvalues of the matrix. The approximation is $\mathbf{K} \approx \mathbf{U}_{N,m}\mathbf{\Lambda}_{m,m}\mathbf{U}_{m,N}$, where the indices refer to the size of the corresponding submatrix. The Nyström method approximates a kernel in a similar way, without computing the eigendecomposition of the whole matrix, which otherwise is an $O(N^3)$ operation.

By the Mercer theorem kernels $k(\mathbf{x}, \mathbf{y})$ can be expanded by orthonormal eigenfunctions $\psi_i$ and non negative eigenvalues $\lambda_i$ in the form

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y}).$$

The eigenfunctions and eigenvalues of a kernel are defined as the solution of the integral equation

$$\int k(\mathbf{y}, \mathbf{x}) \psi_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \lambda_i \psi_i(\mathbf{y}),$$

where $p(\mathbf{x})$ is the probability density of $\mathbf{x}$. This integral can be approximated based on the Nyström technique by sampling $\mathbf{x}^k$ i.i.d. according to $p(\mathbf{x})$:

$$\frac{1}{m} \sum_{k=1}^{m} k(\mathbf{y}, \mathbf{x}^k) \psi_i(\mathbf{x}^k) \approx \lambda_i \psi_i(\mathbf{y}).$$

Using this approximation and the matrix eigenproblem equation

$$\mathbf{K}^{(m)}\mathbf{U}^{(m)} = \mathbf{U}^{(m)}\mathbf{\Lambda}^{(m)}$$

of the corresponding $m \times m$ Gram sub-matrix $\mathbf{K}^{(m)}$ we can derive the approximations for the eigenfunctions and eigenvalues of the kernel $k$

$$\lambda_i \approx \frac{\lambda_i^{(m)}}{m}, \quad \psi_i(\mathbf{y}) \approx \frac{\sqrt{m}}{\lambda_i^{(m)}} \mathbf{k}_y \mathbf{u}_i^{(m)}, \tag{4}$$

where $\mathbf{u}_i^{(m)}$ is the $i$th column of $\mathbf{U}^{(m)}$. Thus, we can approximate $\psi_i$ at an arbitrary point $\mathbf{y}$ as long as we know the vector $\mathbf{k}_y = (k(\mathbf{x}^1, \mathbf{y}), ..., k(\mathbf{x}^m, \mathbf{y}))^\top$.

For a given $N \times N$ Gram matrix $\mathbf{K}$ we randomly choose $m$ rows and respective columns. The corresponding indices's are also called landmarks, and should be chosen such that the data distribution is sufficiently covered. A specific analysis about selection strategies was recently discussed in [21]. We denote these rows by $\mathbf{K}_{m,N}$. Using the formulas (4) we obtain $\tilde{\mathbf{K}} = \sum_{i=1}^{m} 1/\lambda_i^{(m)} \cdot \mathbf{K}_{m,N}^\top \mathbf{u}_i^{(m)} (\mathbf{u}_i^{(m)})^\top \mathbf{K}_{m,N}$, where $\lambda_i^{(m)}$ and $\mathbf{u}_i^{(m)}$ correspond to the $m \times m$ eigenproblem. Thus we get, $\mathbf{K}_{m,m}^{-1}$ denoting the Moore-Penrose pseudoinverse, an approximation of $\mathbf{K}$ as

$$\tilde{\mathbf{K}} = \mathbf{K}_{m,N}^\top \mathbf{K}_{m,m}^{-1} \mathbf{K}_{m,N}. \tag{5}$$

This approximation is exact, if $\mathbf{K}_{m,m}$ has the same rank as $\mathbf{K}$.

## 3.2 Nyström approximation for dissimilarity data

The subsequent part follows widely prior work given in [7] and [18]. According to the spectral theorem, a symmetric dissimilarity matrix $\mathbf{D}$ can be diagonalized $\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{U}$ being a unitary matrix whose column vectors are the orthonormal eigenvectors of $\mathbf{D}$ and $\mathbf{\Lambda}$ a diagonal matrix with the corresponding eigenvalues of $\mathbf{D}$, Therefore the dissimilarity matrix can be seen as an operator

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y})$$

where $\lambda_i \in \mathbb{R}$ correspond to the diagonal elements of $\mathbf{\Lambda}$ and $\psi_i$ denote the eigenfunctions. The only difference to an expansion of a kernel is that the eigenvalues can be negative. All further mathematical manipulations can be applied in the same way and we can write in an analogy to Equation (5)

$$\hat{\mathbf{D}} = \mathbf{D}_{N,m} \mathbf{D}_{m,m}^{-1} \mathbf{D}_{N,m}^\top.$$

It allows to approximate dissimilarities between a point $\mathbf{w}^k$ represented by a coefficient vector $\alpha_k$ and a data point $\mathbf{x}^i$, as discussed within Eq (3), in the way

$$\begin{aligned} d(\mathbf{x}^i, \mathbf{w}^k) \quad \approx \quad & \left[\mathbf{D}_{m,N}^\top \left(\mathbf{D}_{m,m}^{-1} \left(\mathbf{D}_{m,N} \boldsymbol{\alpha}_k\right)\right)\right]_i \\ & -\frac{1}{2} \cdot \left(\boldsymbol{\alpha}_k^\top \mathbf{D}_{m,N}^\top\right) \cdot \\ & \left(\mathbf{D}_{m,m}^{-1} \left(\mathbf{D}_{m,N} \boldsymbol{\alpha}_k\right)\right) \end{aligned}$$

with a linear submatrix of $m$ rows and a low rank matrix $\mathbf{D}_{m,m}$. Performing these matrix multiplications from right to left, this computation is $\mathcal{O}(m^2 N)$ instead of $\mathcal{O}(N^2)$, i.e. it is linear in the number of data points $N$, assuming fixed approximation $m$.

A benefit of the Nyström technique is that it can be decided priorly which linear parts of the dissimilarity matrix will be used in training. Therefore, it is sufficient to *compute only a linear part of the full dissimilarity matrix* $\mathbf{D}$ to use these methods. A drawback of the Nyström approximation is that a good approximation can only be achieved if the rank of $\mathbf{D}$ is kept as much as possible, i.e. the chosen subset should be representative. The specific selection of the $m$ landmark points has been recently analyzed in [21]. It was found that best results can be obtained by choosing the potential cluster centers of the data distribution as landmarks, rather a random subset, to be able to keep $m$ smallest at lowest representation error. However the determination of these centers can become complicated for large data sets, since it can be obviously not be based on a Nyström approximated set. However the effect is not such severe as long as $m$ is not too small. We will come back to this point in a later section.

# 4 Transformations of (dis-)similarities with linear costs

For *metric* similarity data, kernel methods can be applied directly, or in case of large $N$, the Nyström approximation can be used. Following [18] we will now briefly discuss almost metric *dissimilarity* data $\mathbf{D}$ and consider *non*-metric data later on. Especially

we review a linear cost transformation of $D$ to $S$ using the Nyström approximation, for small $m$ which gives access to efficient kernel methods[2].

## 4.1 Transformation of dissimilarities to similarities

Instead of applying double centering, followed by the Nyström approximation we first approximate the matrix $\mathbf{D}$ and then transform it by double centering, which yields the approximated similarity matrix $\hat{\mathbf{S}}$. As mentioned before double centering of a matrix $\mathbf{D}$ is defined as:

$$\mathbf{S} = -\mathbf{JDJ}/2$$

where $\mathbf{J} = (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/N)$ with identity matrix $\mathbf{I}$ and vector of ones $\mathbf{1}$. $\mathbf{S}$ is positive semi-definite (psd) if and only if D is Euclidean.

Lets start with a dissimilarity matrix D where we apply double centering, subsequently we approximate the obtained $\mathbf{S}$ by integrating the Nyström approximation to the matrix $\mathbf{D}$.

$$\begin{aligned}
\mathbf{S} &= -\frac{1}{2}\mathbf{JDJ} \\
&= -\frac{1}{2}\left(\left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\mathbf{D}\left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\right) \\
&= -\frac{1}{2}\left(\mathbf{IDI} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{DI} - \mathbf{ID}\frac{1}{N}\mathbf{1}\mathbf{1}^\top + \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{D}\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) \\
&= -\frac{1}{2}\left(\mathbf{D} - \frac{1}{N}\mathbf{D}\mathbf{1}\mathbf{1}^\top - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{D} + \frac{1}{N^2}\mathbf{1}\mathbf{1}^\top\mathbf{D}\mathbf{1}\mathbf{1}^\top\right)
\end{aligned}$$

$$\begin{aligned}
\mathbf{S} \overset{Ny}{\approx} \hat{\mathbf{S}} &= -\frac{1}{2}\Big[\mathbf{D}_{N,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,N} - \frac{1}{N}\mathbf{D}_{N,m} \\
&\quad \cdot(\mathbf{D}_{m,m}^{-1}\cdot(\mathbf{D}_{m,N}\mathbf{1}))\mathbf{1}^\top - \frac{1}{N}\mathbf{1}((\mathbf{1}^\top\mathbf{D}_{N,m})\cdot\mathbf{D}_{m,m}^{-1}) \\
&\quad \cdot\mathbf{D}_{m,N} + \frac{1}{N^2}\mathbf{1}((\mathbf{1}^\top\mathbf{D}_{N,m})\cdot\mathbf{D}_{m,m}^{-1}\cdot(\mathbf{D}_{m,N}\mathbf{1}))\mathbf{1}^\top\Big]
\end{aligned} \tag{6}$$

This equation can be rewritten for each entry of the matrix $\hat{\mathbf{S}}$

$$\begin{aligned}
\hat{S}_{ij} &= -\frac{1}{2}\Big[\mathbf{D}_{i,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,j} - \frac{1}{N}\sum_k \mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,j} \\
&\quad -\frac{1}{N}\sum_k \mathbf{D}_{i,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,k} \\
&\quad +\frac{1}{N^2}\sum_{kl}\mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,l}\Big],
\end{aligned}$$

[2]This approach has a weak connection to Landmark MDS, but also substantial differences as discussed in detail in [18]

as well as for the sub-matrices $\hat{\mathbf{S}}_{m,m}$ and $\hat{\mathbf{S}}_{N,m}$, in which we are interested for the Nyström approximation. These two matrices are also interesting later on when we replace the standard Nyström approximation by the proposed extension including the Nyström approximation for large scale problems:

$$\begin{aligned}
\hat{\mathbf{S}}_{m,m} &= -\frac{1}{2}\Bigg[\mathbf{D}_{m,m} - \frac{1}{N}\mathbf{1}\cdot\sum_k \mathbf{D}_{k,m} \\
&\quad -\frac{1}{N}\sum_k \mathbf{D}_{m,k}\cdot\mathbf{1}^\top \\
&\quad +\frac{1}{N^2}\mathbf{1}\cdot\sum_{kl}\mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,l}\cdot\mathbf{1}^\top\Bigg] \\
\hat{\mathbf{S}}_{N,m} &= -\frac{1}{2}\Bigg[\mathbf{D}_{N,m} - \frac{1}{N}\mathbf{1}\cdot\sum_k \mathbf{D}_{k,m} \\
&\quad -\frac{1}{N}\sum_k \mathbf{D}_{N,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,k}\cdot\mathbf{1}^\top \\
&\quad +\frac{1}{N^2}\mathbf{1}\cdot\sum_{kl}\mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,l}\cdot\mathbf{1}^\top\Bigg].
\end{aligned} \tag{7}$$

It should be noted that $\hat{\mathbf{S}}$ is only a valid kernel if $\hat{\mathbf{D}}$ is metric. The information loss obtained by the approximation is $0$ if $m$ corresponds to the rank of $\mathbf{S}$ and increases for smaller $m$.

## 4.2 Non-metric (dis-)similarities

In case of a non-metric $\mathbf{D}$ the transformation shown in equation 6 can still be used, but the obtained matrix $\hat{\mathbf{S}}$ is not a valid kernel. A strategy to obtain a valid kernel matrix $\hat{\mathbf{S}}$ is to apply an eigenvalue correction as discussed above. This however can be prohibitive for large matrices, since to correct the whole eigenvalue spectrum, the whole eigenvalue decomposition is needed, which has $\mathcal{O}(N^3)$ complexity. The Nyström approximation can again decrease computational costs dramatically. Since we now can apply the approximation on an arbitrary symmetric matrix, we can make the correction afterward. To correct an already approximated similarity matrix $\hat{\mathbf{S}}$ it is sufficient to correct the eigenvalues of $\mathbf{S}_{m,m}$. Altogether we get $\mathcal{O}(m^2N)$ complexity.

We can write for the approximated matrix $\hat{\mathbf{S}}$ its eigenvalue decomposition as

$$\hat{\mathbf{S}} = \mathbf{S}_{N,m}\mathbf{S}_{m,m}^{-1}\mathbf{S}_{N,m}^\top = \mathbf{S}_{N,m}\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top\mathbf{S}_{N,m}^\top,$$

where we can correct the eigenvalues $\mathbf{\Lambda}$ by some technique as discussed in section 2.1 to $\mathbf{\Lambda}^*$. The corrected approximated matrix $\hat{\mathbf{S}}^*$ is then simply

$$\hat{\mathbf{S}}^* = \mathbf{S}_{N,m}\mathbf{U}\left(\mathbf{\Lambda}^*\right)^{-1}\mathbf{U}^\top\mathbf{S}_{N,m}^\top. \tag{8}$$

This approach can also be used to correct dissimilarity matrices $\mathbf{D}$ by first approximating them, converting to similarities $\hat{\mathbf{S}}$ using equation 6 and then correcting the

similarities. If it is desirable to work with the corrected dissimilarities, then we should note, that it is possible to transform the similarity matrix $\mathbf{S}$ to a dissimilarity matrix $\mathbf{D}$:

$$D_{ij}^2 = S_{ii} + S_{jj} - 2S_{ij}. \tag{9}$$

This obviously applies as well to the approximated and corrected matrices $\hat{\mathbf{S}}^*$ and $\hat{\mathbf{D}}^*$ and we get by substitution:

$$\hat{\mathbf{D}}^* = \mathbf{D}_{N,m}^* \left(\mathbf{D}_{m,m}^*\right)^{-1} \mathbf{D}_{N,m}^{*\top}. \tag{10}$$

Usually the algorithms are learned on a so called training set and we expect them to perform well on the new unseen data, or the test set. In such cases we need to provide an out of sample extension, i.e. a way to compute the algorithm on the new data. This might be a problem for the techniques dealing with (dis)similarities. If the matrices are corrected, we need to correct the new (dis)similarities as well to get consistent results. Fortunately, it is quite easy in the Nyström framework. By examining Eq. (8) and Eq. (10) we see, that we simply need to extend the matrices $\mathbf{D}_{N,m}$ or $\mathbf{S}_{N,m}$, respectively, by uncorrected (dis)similarities between the new points and the landmarks to obtain the full approximated and *corrected* (dis)similarity matrices, which then can be used by the algorithms to compute the out of sample extension. A more detailed discussion including experiments is available in [18].

# 5 Large scale Nyström approximation for non-metric dissimilarities

In [14] a new approach for the calculation of the Nyström approximation for large *psd* matrices was proposed, we will denote this approach as LSNA and our proposal as extended LSNA (e-LSNA). As one can see in Eq. (5) the Nyström approximation is based on the calculation of a pseudo-inverse of a matrix based on $m$ rows and $m$ columns. The *optimal* landmarks specifying these columns and rows are the cluster centers of the considered data set, which are hard to identify in advance. Accordingly, $m$ is often chosen to be sufficiently large such that the landmarks are likely to cover enough information of the data distribution. For large data sets, containing e.g. million of points the number of landmarks is also getting large e.g. $m = 1000$ and the calculation of the pseudo-inverse may dominate the remaining calculation costs due to the cubic complexity. The idea, presented in [14] is to use a randomized singular value decomposition (SVD) [10] on the landmark matrix to obtain an accurate $m \times m$ matrix in the Nyström approximation at low costs, see Alg. 1.

Thereby the data are represented on a lower dimensional subspace e.g. in $k$ dimensions with $k \ll m$ such that the obtained singular value matrix $L$ can be inverted with low costs of $O(k^3)$. The final Nyström approximation of the original kernel or (as we will see soon) dissimilarity matrix can be obtained by subsequent matrix multiplications leading to a similar formulation as before. The basic algorithm taken from [14] is shown in Alg. 2. Here, $p$ is an over-sampling parameter (typically set to $5$ or $10$) such that the rank of $Q$ is slightly larger than the desired rank($k$), and $q$ is the number of steps of a power iteration (typically set to $1$ or $2$) which is used to speed up the decay of the singular values of $W$ [14].

**Algorithm 1** Randomized SVD [10]

**init:** $m \times m$ matrix $W$, scalars $k$, $p$, $q$
**Output:** $U, \Lambda$
$\Omega \leftarrow m \times (k+p)$ standard gaussian random matrix
$Z \leftarrow W\Omega$, $Y \leftarrow W^{q-1}Z$,
find orthonormal $Q$ such that $Y = QQ^\top Y$
$B(Q^\top Q) = Q^\top Z$
$[V, L] = svd(B)$
$U \leftarrow QV$

**Algorithm 2** The large scale Nyström approximation [14]

**init:** psd matrix $K \in \mathbb{R}^{N \times N}$ , number of landmarks $m$, rank $k$, over-sampling parameter $p$, power parameter $q$
**Output:** $\hat{K}$, an approximation of $K$
$C \leftarrow m$ columns of $K$ sampled uniformly at random without replacement
$W \leftarrow m \times m$ landmark matrix
$[\tilde{U}, \hat{\Lambda}] \leftarrow ranksvd(W, k, p, q)$ using Alg. 1
$U \leftarrow C\tilde{U}\Lambda^{-1}$
$\hat{K} \leftarrow \left(\sqrt{\frac{m}{N}}U\right)\left(\frac{m}{N}\Lambda\right)\left(\sqrt{\frac{m}{N}}U^\top\right)$

If the given proximity data are psd similarities, algorithm 2 can be used directly. For non-psd similarities the SVD used in algorithm 2 implicitly flips negative eigenvalues. Due to the random projection step it may obviously also happen that smaller absolute eigenvalues are removed. While this appears to be a nice feature it is not always clear if flipping is a good strategy as dicussed e.g. in [18, 1]. For example the negativ eigenvalue contributions may account for noise and a clipping may be more desirable. As argued in [17] it may even be desirable to keep also negative eigenvalues, given the subsequent algorithm, used to analyze the data can handle non-psd matrices.

If the data are metric dissimilarities the approach of [14] can be directly applied with the same argumentation as for the standard Nyström approximation on dissimilarities discussed before. For non-metric dissimilarities additional corrections are necessary. In the following we will discuss how Algorithm 1,2 can be used for non-metric similarities and dissimilarities and how this is linked to section 4.

## 5.1 Non-metric similarities

The LSNA approach performs an implicite flipping of negative eigenvalues in the SVD step, to get more control about the handling of negative eigenvalues we will introduce an explicite step to correct the eigenvalue but in the low dimensional projection space avoiding high computational costs. The modified LSNA algorithm is shown in Algorithm 3 and 4.

The new formulation accounts for an explicite eigenvalue correction in Algorithm 3. Note that both unitary matrices $V$ and $V'$ are used. To make the out of sample extension more obvious, it is also convenient to modify the reconstruction of the proximity matrix as shown in Alg. 4, line 6. Now it can be directly seen how to extended the

**Algorithm 3** Randomized SVD with eigenvalue correction

1: **init:** $m \times m$ matrix $W$, scalars $k$, $p$, $q$
2: **Output:** $U, \Lambda, V$
3: $\Omega \leftarrow m \times (k+p)$ standard gaussian random matrix
4: $Z \leftarrow W\Omega$, $Y \leftarrow W^{q-1}Z$,
5: find orthonormal $Q$ such that $Y = QQ^\top Y$
6: $B(Q^\top Q) = Q^\top Z$
7: $[E_b, V_b] = eig(B)$
8: $V_b^* \leftarrow$ flip—clip—shift$(V_b)$
9: $B^* = E_b \cdot V_b^* \cdot E_b^\top$
10: $[V, L, V'] = svd(B^*)$
11: $U \leftarrow QV'$

matrix $K$ by $l$ items. One only needs to calculate the corresponding $m$ similarities to the $m$ landmark points which can become part of an extended matrix $C$.

**Algorithm 4** The large scale Nyström approximation with eigenvalue corrected similarities

1: **init:** psd matrix $K \in \mathbb{R}^{N \times N}$ , number of landmarks $m$, rank $k$, over-sampling parameter $p$, power parameter $q$
2: **Output:** $\hat{K}$, an approximation of $K$
3: $C \leftarrow m$ columns of $K$ sampled uniformly at random without replacement
4: $W \leftarrow m \times m$ landmark matrix
5: $[\tilde{U}, \hat{\Lambda}, V] \leftarrow ranksvd(W, k, p, q)$ using Alg. 3
6: $\hat{K} \leftarrow \left(\sqrt{\frac{m}{N}}C\right)\left(\frac{m}{N}V(\tilde{U}\Lambda^{-1})^\top\right)\left(\sqrt{\frac{m}{N}}C^\top\right)$

Equation (8) can be modified to integrate Algorithm 4 straight forward by replacing $\mathbf{S}_{N,m}$ with $\sqrt{\frac{m}{N}}C$ and $\mathbf{U}\left(\mathbf{\Lambda}^*\right)^{-1}\mathbf{U}^\top$ with $\left(\frac{m}{N}V(\tilde{U}\Lambda^{-1})^\top\right)$.

## 5.2 Non-metric dissimilarities

For non-metric dissimilarities we use Algorithm 4, but now for dissimilarities and without an eigenvalue correction. One samples a landmark matrix $\mathbf{D}_{m,m}$ from $\mathbf{D}$ which is used as input of Algorithm 3 in line 5 of Algorithm 4. Based on this LSNA model for the dissimilarities of $\mathbf{D}$ we calculate the double centered and LSNA approximated similarity matrix $C$ and $W$ using Eq. (1) or Eq. (7). It should be noted that the pre-factors $\sqrt{\frac{m}{N}}$ and $\frac{N}{m}$ cancel out. The approximated $C$ and $W$ are used as input of Algorithm 2 starting at line 5 which may also include eigenvalue corrections. Using this strategy the computational costs are $O(m^2k + k^3)$ for the LSNA approximation of the dissimilarity matrix, costs of $O(mN)$ for the double centering only based on the LSNA approximation and additional costs of $O(m^2k + 2k^3)$ for the randomized SVD on the similarity matrices and a potential eigenvalue correction to obtain a corrected and approximated similarity matrix $\hat{\mathbf{S}}^*$. The matrix $\hat{\mathbf{S}}^*$ can now be transfered back to a corrected dissimilarity matrix $\hat{\mathbf{D}}^*$ using Eq. (9) with costs of $O(Nm)$.

The complete costs of this operations are $O(Nmk + k^3)$ which has still the same complexity as the original LSNA algorithm. We are now able to process potentially non-metric similarities and dissimilarities at large scale.

# 6 Experiments

We apply the priorly derived approach to three non-metric dissimilarity and similarity data and show the effectiveness for a classification task. The considered data are (1) the SwissProt similarity data as described in [12] (DS1, 10988 samples, 30 classes, imbalanced, signature: $[8488, 2500, 0]$) (2) the chromosome dissimilarity data taken from [15] (DS2, 4200 samples, 21 classes, balanced, signature: $[2258, 1899, 43]$) and the prodom dissimilarity data set [4], restricted to classes with at least $10$ entries (DS3, 2518 samples, 33 classes, imbalanced, signature: $[717, 1512, 289]$). All datasets are non-metric, multiclass and contain multiple thousand objects, such that a regular eigenvalue correction with a prior double-centering for dissimilarity data, as discussed before, is already very costly. While the approach presented in [18] is still sufficient for these data, approximations with larger $m$ can become unreliable. Due to the low rank approximation small (negative) eigenvalues are potentially removed such that a low value of $k$ in the extended LSNA approach may already lead to a clipping effect in the data. The same applies for small values of $m$ where only major eigenvalues are kept. Larger $k$ and larger $m$ both will likely lead to an improved approximation, but on the other hand can also increase the influence of negative eigenvalues which may lead to sub-optimal results for the chosen learning algorithms. Note again, that for large dissimilarity matrices the access to kernel methods can only be achieved using the approach in [18] or by using the presented extended LSNA approach and an eigenvalue correction is often necessary.

The data are analyzed in two ways, employing either the clipping or flipping strategy as an eigenvalue correction, or by not-correcting the eigenvalues[3]. To be effective for the large number of object we also apply the Nyström approximation as discussed before using a sample rate of $m = 1\%, 10\%, 30\%$[4], by selecting random landmarks from the data, with $k = 100, p = 5, q = 2$. Other sampling strategies have been discussed in [21, 6], also the impact of the Nyström approximation with respect to kernel methods has been discussed recently in [2], but this is out of the focus of this paper.

To get comparable experiments, the same randomly drawn landmarks are used in each of the corresponding sub-experiments (along a column in the table). New landmarks are only drawn for different Nyström approximations and sample sizes. Classification rates are calculated in a 10-fold crossvalidation using the Core-Vector-Machine (CVM) (see [19]). The crossvalidation does not include a new draw of the landmarks, to cancel out the selection bias of the Nyström approximation, accordingly CVM use the same kernel matrices. However, our objective is not maximum classification performance (which is only one possible application) but to demonstrate the effectiveness of our approach for dissimilarity data of larger scale. The classification results are summarized in Table 1-2 for the different Nyström approximations $1\%, 10\%$ and $30\%$.

First one observes that the eigenvalue correction has a positive effect on the classification performance, it is however less pronounced as in former findings [1, 18]. This

[3] Shift correction was found to have a negative impact on the model as already discussed in [1].
[4] A larger sample size did not lead to further substantial improvements in the results.

Table 1: Average test set accuracy for SwissProt (DS1), Chromosome (DS2), Proteom (DS3) using a Nyström approximation of $1\%$ and $10\%$ and no , clip or flip eigenvalue correction. Kernel matrices have been Nyström approximated either, as proposed during the eigenvalue correction, or later on, like in the standard approach. The signatures are based on the approximated kernel matrices.

| | $DS1_{1\%}$ | $DS2_{1\%}$ | $DS3_{1\%}$ | $DS1_{10\%}$ | $DS2_{10\%}$ | $DS3_{10\%}$ |
|---|---|---|---|---|---|---|
| Signature | [106,0,10882] | [42,1,4157] | [26,1,2491] | [105,0,10883 ] | [98,7,4095] | [105,0,2413] |
| CVM-No | $93.60 \pm 0.73$ | $94.74 \pm 1.19$ | $85.31 \pm 1.95$ | $96.91 \pm 0.50$ | $93.29 \pm 0.73$ | $99.68 \pm 0.31$ |
| Signature | [106,0,10882] | [42,0,4159] | [26,0,2492] | [105,0,10883 ] | [105,0,4095] | [105,0,2413] |
| CVM-Flip | $93.53 \pm 0.80$ | $94.74 \pm 1.19$ | $85.31 \pm 1.95$ | $96.74 \pm 0.51$ | $96.64 \pm 0.78$ | $99.76 \pm 0.28$ |
| Signature | [106,0,10882] | [42,0,4158] | [26,0,2492] | [105,0,10883] | [101,0,4099] | [105,0,2413] |
| CVM-Clip | $93.57 \pm 0.74$ | $94.74 \pm 1.11$ | $85.31 \pm 1.95$ | $97.05 \pm 0.56$ | $96.71 \pm 0.92$ | $99.64 \pm 0.28$ |

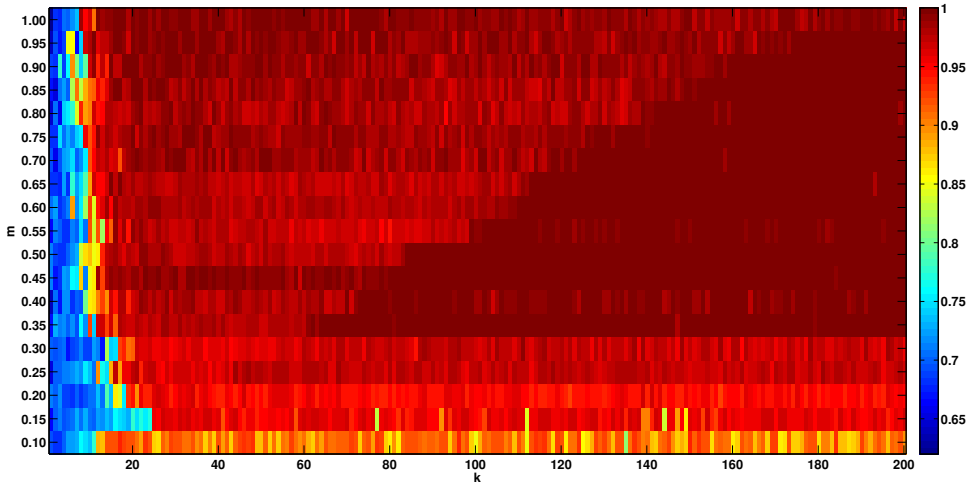
Table 2: Average test set accuracy for SwissProt (DS1), Chromosome (DS2), Proteom (DS3) using a Nyström approximation of $30\%$ and no, clip or flip eigenvalue correction. Kernel matrices have been Nyström approximated (with $L = 30\% \cdot N$) either, as proposed during the eigenvalue correction, or later on, like in the standard approach. The signatures are based on the approximated kernel matrices.

| | DS1 | DS2 | DS3 |
|---|---|---|---|
| Signature | [105,0,10883] | [101,4,4095] | [105,0,2413] |
| CVM-No | $96.92 \pm 0.52$ | $96.64 \pm 0.81$ | $99.92 \pm 0.17$ |
| Signature | [105,0,10883] | [105,0,4095] | [105,0,2413] |
| CVM-Flip | $97.23 \pm 0.45$ | $96.86 \pm 0.98$ | $99.92 \pm 0.16$ |
| Signature | [105,0,10883] | [101,0,4099] | [105,0,2413] |
| CVM-Clip | $97.34 \pm 0.50$ | $96.81 \pm 0.80$ | $99.92 \pm 0.16$ |

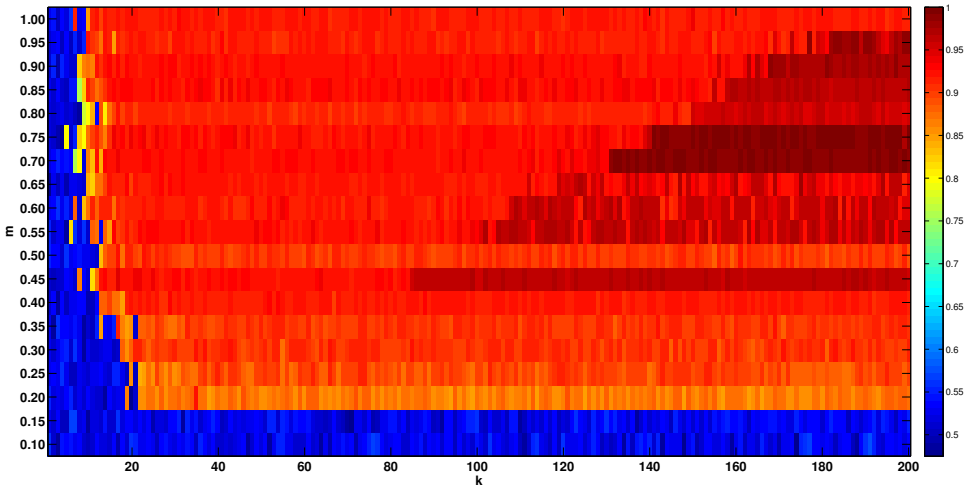
is explained by the low rank approximation used in the SVD which also in case of Algorithm 4 without an eigenvalue correction in line 5 removes smaller absolute eigenvalues in the approximation step. This can be also seen in the signatures for the different data sets without an explicit eigenvalue correction. The majority of the eigenvalues are $0$ in contrast to the original signature of the full, unapproximated data set. This effect is mainly caused due to the approximation. Also without an eigenvalue correction the results are quite good, in contrast to earlier findings in [18] using the classical Nyström approximation. Obviously the smoothing due to the sub-space approximation has a similar effect like a clipping eigenvalue correction. Only when a substantial amount of negative eigenvalues is kept (see e.g for chromosoms with $m = 10\%$) a clear degeneration of the classification accuracy was found.

Regarding the parameter $q$ some small experiments, not shown here, indicate that $2$ is a reasonable value to avoid numerical approximation errors in the $qr$ decomposition. For $q = 1$ numerical instabilties can be observed in parts leading to strong errors in the subsequent Nyström approximation due to an ill-conditioned pseudo inverse of the matrix $\Lambda$ in line 6 of algorithm 4.

In a second experiment we analyze the influence of the parameter $m$ and $k$ for a small artificial ball data set originally proposed in [5]. It is an artificial dataset based on




(a) Flipping




(b) Clipping

Figure 2: Ball data set with clipping and flipping using different parameters of $k$ and $m$. The classification accuracy is indicated by color (blue / dark = low prediction accuracy).

the surface distances of randomly positioned balls of two classes having a slightly different radius. The dataset is non-euclidean with substantial information encoded in the negative part of the eigenspectrum. It is however simple enough to be analyzed also with the extended LSNA. We generated the data with $100$ samples per class leading to a dissimilarity matrix $D = N \times N$, with $N = 200$. Now the data have been processed by the extended LSNA to obtain a valid kernel matrix $S$ with different parameters $m$ and $k$ and by flipping or clipping the negative eigenvalues. The crossvalidation results are shown in Figure 2[5]

As expected flipping performed better than clipping because the dataset contains discriminative information in the negative eigenvalues by construction. For large $m$ and $k$ the accuracy is in general quite high but the optimal values are obtained for slightly smaller $k$ to avoid numerical instabilities in the projection step of the randomized SVD.

[5]For this data set a SVM did not converge on the uncorrected kernel matrix.

Very small $m$ and $k$ are oversimplifying the problem leading to a smooth eigenspectrum such that smaller but relevant eigenvalues are lost. With very few exceptions a perfect discrimination (no error on the test set) could only be obtained by flipping.

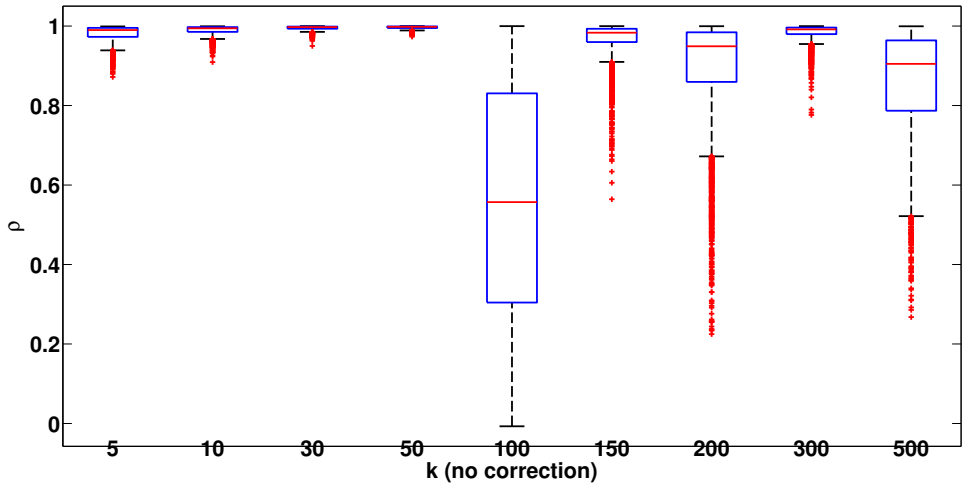
In a third experiment we address the influence of the parameter $k$ controlling the dimensionality of the sub-space projection for a real data set. Smaller $k$ lead to a stronger approximation such that small absolute eigenvalues are neglected while larger values of $k$ (with $k \leq m$) can keep also smaller absolute eigenvalues. Depending on the importance of small eigenvalues for the considered problem a kind of denoising can be expected but for small $k$, it may also happen that relevant information in the data space is removed. We analyze the effect by measuring the classification error and the spearman rank correlation of the original vs the approximated matrix. The results with $m = 0.1 \times N$ and varying $k$ are depicted in Figure 3.

One can clearly identify an optimal $k$ with respect to the rank preservation and the cross validation accuracy. With a Nyström approximation of $m = 10\%$ a reasonable $k$ should be substantial smaller than $420$. As we can see from Figure 3 an increase of $k$ close to $m$ leads to a degeneration of the rank accuracy. With respect to the different eigenvalue normalization (no - Fig. 3(a), clip - Fig. 3(b), flip - Fig.3(c)) it is obvious that an eigenvalue correction is beneficial for this data set. Without a correction and for larger $k$ see Fig. 3(a) and Fig. 3(d) the matrix is less accurate reconstructed, with respect to the proximity relations as well as with respect to the accuracy on test data. Also the flipping procedure is less effective for larger $k$ because the negative eigenvalues are kept in the data representation. Only the clipping operation was found to be widely stable and effective also for larger $k$. For small $k$ the eigenvalue spectrum is implicitely clipped with respect to smaller absolute eigenvalues.

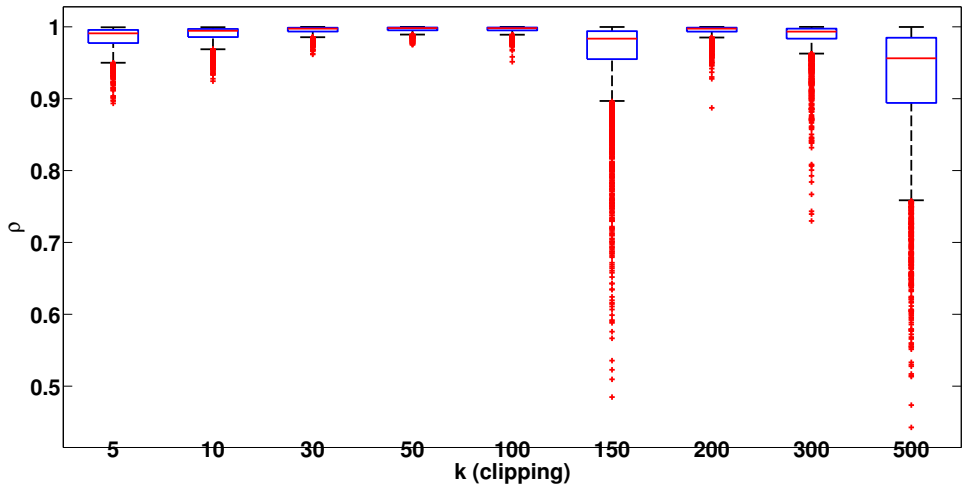
# 7 Outlook and Conclusions

In this paper we discussed the relation between similarity and dissimilarity data and effective ways to move across the different representations in a systematic way also for very large data sets where a standard Nyström approximation gets to its limits or can not any longer be applied. We extended LSNA such that non-metric similarities *and* dissimilarities can be processed. Using the presented approach, effective and *accurate* transformations are possible as summarized in Figure 1. Kernel approaches but also dissimilarity learners are now accessible for both types of data at large scale. While the parametrization of the Nyström approximation is already studied in [13, 21] there are still different open issues. In future work we will analyze our approach in the context of unsupervised problems. A further point is the overall question whether eigenvalues correction should be done. As argued in [17] for some dedicated data formats such a correction should be avoided and dedicated methods for non-metric data are of interest. In contrast to LSNA the extended approach permits more control on this process, although due to the inherent approximation by LSNA smaller eigenvalues are typically removed by purpose. If a dataset contains larger negative eigenvalues, which are not relevant for the problem, our approach can be used to remove these eigenvalues in a systematic ways also for extended LSNA with larger values on $m$ and $k$ which is not possible using the original LSNA.

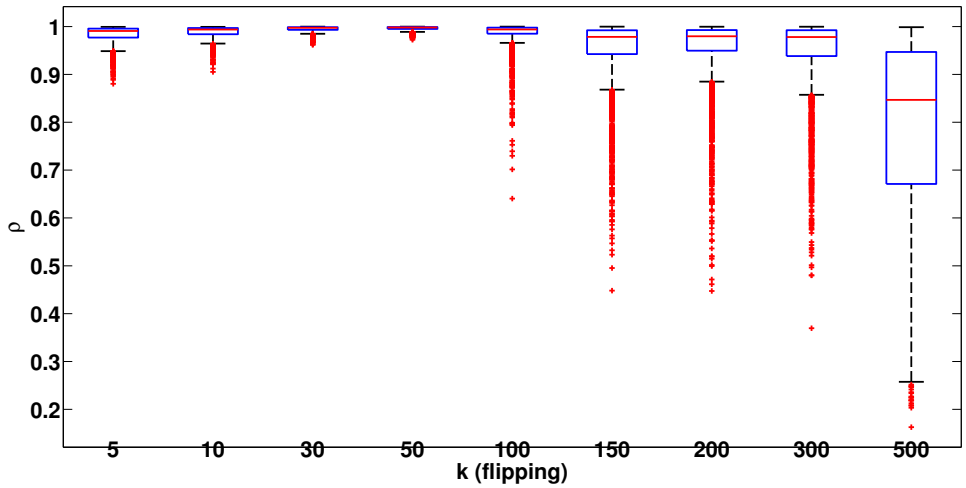
For non-psd data the error introduced by the Nyström approximation is not yet fully understood and bounds similar as proposed in [3] are still an open issue. In our ex-



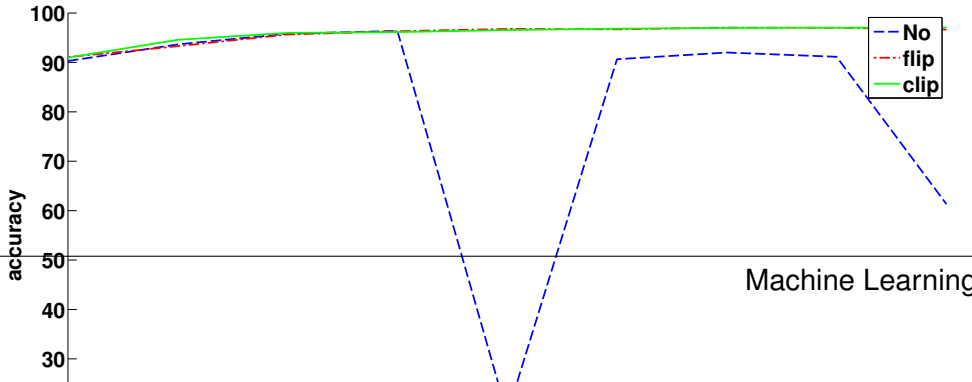

(a) Rank preservation for chromosom data uncorrected




(b) Rank preservation for Chromosom with clipping




(c) Rank preservation for Chromosom with flipping



periments we observed that flipping was an effective approach to keep the relevant structure of the data but these are only heuristic findings and not yet completely understood, we will address this in future work.

# References


[1] Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A., Cazzanti, L.: Similarity-based classification: Concepts and algorithms. JMLR 10, 747–776 (2009)

[2] Cortes, C., Mohri, M., Talwalkar, A.: On the impact of kernel approximation on learning accuracy. JMLR - Proceedings Track 9, 113–120 (2010)

[3] Drineas, P., Mahoney, M.W.: On the nyström method for approximating a gram matrix for improved kernel-based learning. Journal of Machine Learning Research 6, 2153–2175 (2005)

[4] Duin, R.P.: PRTools (march 2012), http://www.prtools.org

[5] Duin, R.P.W., Pekalska, E.: Non-euclidean dissimilarities: Causes and informativeness. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SSPR/SPR. Lecture Notes in Computer Science, vol. 6218, pp. 324–333. Springer (2010)

[6] Farahat, A.K., Ghodsi, A., Kamel, M.S.: A novel greedy algorithm for nyström approximation. JMLR - Proceedings Track 15, 269–277 (2011)

[7] Gisbrecht, A., Mokbel, B., Hammer, B.: The Nystrom approximation for relational generative topographic mappings. In: NIPS Workshop (2010)

[8] Gisbrecht, A., Mokbel, B., Schleif, F.M., Zhu, X., Hammer, B.: Linear time relational prototype based learning. Journal of Neural Systems p. in press (2012), pdf/ijns_2012.pdf

[9] Graepel, T., Obermayer, K.: A stochastic self-organizing map for proximity data. Neural Computation 11(1), 139–155 (1999)

[10] Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review 53(2), 217–288 (2011)

[11] Hammer, B., Mokbel, B., Schleif, F.M., Zhu, X.: Prototype-based classification of dissimilarity data. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA. Lecture Notes in Computer Science, vol. 7014, pp. 185–197. Springer (2011), pdf/ida_2011.pdf

[12] Kohonen, T., Somervuo, P.: How to make large self-organizing maps for nonvectorial data. Neural Networks 15(8-9), 945–952 (2002)

[13] Kumar, S., Mohri, M., Talwalkar, A.: On sampling-based approximate spectral decomposition. In: ICML. ACM International Conference Proceeding Series, vol. 382, p. 70. ACM (2009)

[14] Li, M., Kwok, J.T., Lu, B.L.: Making large-scale nyström approximation possible. In: Fürnkranz, J., Joachims, T. (eds.) ICML. pp. 631–638. Omnipress (2010)

[15] Neuhaus, M., Bunke, H.: Edit distance based kernel functions for structural pattern classification. Pattern Recognition 39(10), 1852–1863 (2006)

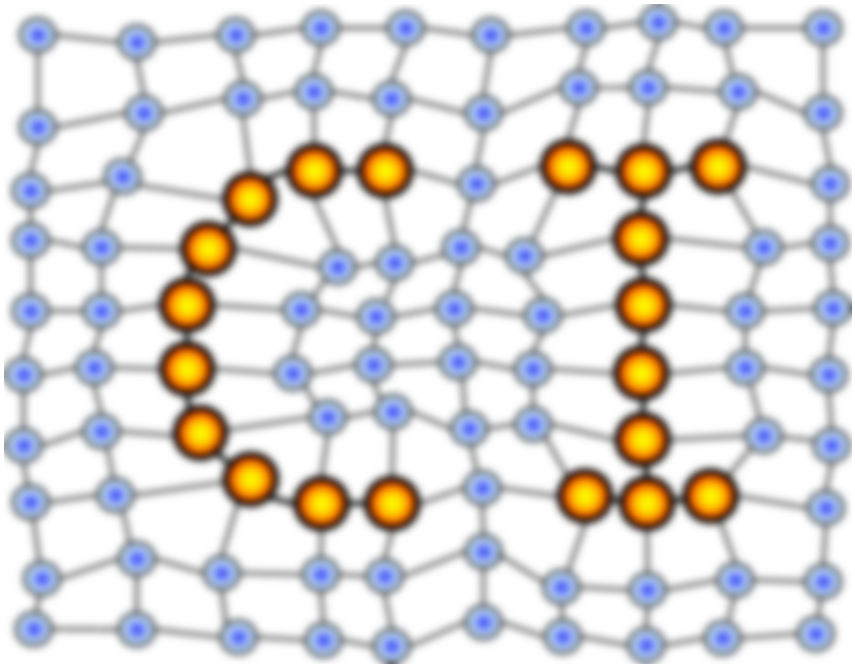
[16] Pekalska, E., Duin, R.: The dissimilarity representation for pattern recognition. World Scientific (2005)

[17] Pekalska, E., Duin, R.P.W., Günter, S., Bunke, H.: On not making dissimilarities

euclidean. In: SSPR/SPR. Lecture Notes in Computer Science, vol. 3138, pp. 1145–1154. Springer (2004)
[18] Schleif, F.M., Gisbrecht, A.: Data analysis of (non-)metric proximities at linear costs. In: Proceedings of SIMBAD 2013. pp. 59–74 (2013), pdf/simbad_2013.pdf
[19] Tsang, I.W., Kocsor, A., Kwok, J.T.: Simpler core vector machines with enclosing balls. In: ICML. ACM International Conference Proceeding Series, vol. 227, pp. 911–918. ACM (2007)
[20] Williams, C.K.I., Seeger, M.: Using the nyström method to speed up kernel machines. In: NIPS. pp. 682–688. MIT Press (2000)
[21] Zhang, K., Kwok, J.T.: Clustered nyström method for large scale manifold learning and dimension reduction. IEEE Transactions on Neural Networks 21(10), 1576–1587 (2010)
[22] Zhang, K., Lan, L., Wang, Z., Moerchen, F.: Scaling up kernel svm on limited resources: A low-rank linearization approach. JMLR - Proceedings Track 22, 1425–1434 (2012)

# MACHINE LEARNING REPORTS

Report 03/2013




**Impressum**

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about





▽ Acknowledgments

We would like to thank the reviewers for their time and patience.