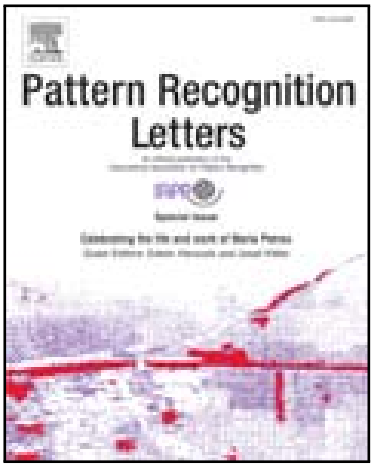
Journal Pre-proof



Sparsification of core set models in non-metric supervised learning

Frank-Michael Schleif, Christoph Raab, Peter Tino

PII: S0167-8655(19)30303-4
DOI: https://doi.org/10.1016/j.patrec.2019.10.024
Reference: PATREC 7672

To appear in: *Pattern Recognition Letters*

Received date: 4 February 2019
Revised date: 27 August 2019
Accepted date: 21 October 2019

Please cite this article as: Frank-Michael Schleif, Christoph Raab, Peter Tino, Sparsification of core set models in non-metric supervised learning, *Pattern Recognition Letters* (2019), doi: https://doi.org/10.1016/j.patrec.2019.10.024

Highlights

- two sparsification methods for indefinite learning models are proposed
- a formulation of an indefinite core vector regression is derived
- sparsification is evaluated on supervised classification and regression problems

Pattern Recognition Letters
journal homepage: www.elsevier.com

Sparsification of core set models in non-metric supervised learning

Frank-Michael Schleif[a,b,**], Christoph Raab[a], Peter Tino[b]

[a] University of Appl. Sc. Würzburg-Schweinfurt, Dept. of Computer Science, Sanderheinrichsleitenweg 20, 97074 Würzburg, Germany
[b] University of Birmingham School of Computer Science, Edgbaston, B15 2TT, Birmingham, UK

ABSTRACT

Supervised learning employing positive semi definite kernels has gained wide attraction and lead to a variety of successful machine learning approaches. The restriction to positive semi definite kernels and a hilbert space is common to simplify the mathematical derivations of the respective learning methods, but is also limiting because more recent research indicates that non-metric, and therefore non positive semi definite, data representations are often more effective. This challenge is addressed by multiple approaches and recently dedicated algorithms for so called indefinite learning have been proposed. Along this line, the Krĕin space Support Vector Machine (KSVM) and variants are very efficient classifiers for indefinite learning problems, but with a non-sparse decision function. This very dense decision function prevents practical applications due to a costly out of sample extension. We focus on this problem and provide two post processing techniques to sparsify models as obtained by a Krĕin space SVM approach. In particular we consider the indefinite Core Vector Machine and indefinite Core Vector Regression Machine which are both efficient for psd kernels, but suffer from the same dense decision function, if the Krĕin space approach is used. We evaluate the influence of different levels of sparsity and employ a Nyström approach to address large scale problems. Experiments show that our algorithm is similar efficient as the non-sparse Krĕin space Support Vector Machine but with substantially lower costs, such that also problems of larger scale can be processed.

Learning of classification models for indefinite kernels received substantial interest with the advent of domain specific similarity measures. Indefinite kernels are a severe problem for most kernel based learning algorithms because classical mathematical assumptions such as positive definiteness, used in the underlying optimization frameworks are violated. As a consequence e.g. the classical Support Vector Machine (SVM) (Vapnik, 2000) has no longer a convex solution - in fact, most standard solvers will not even converge for this problem (Loosli et al., 2016). Researchers in the field of e.g. psychology (Hodgetts and Hahn, 2012), vision (Scheirer et al., 2014; Xu et al., 2011)and machine learning (Duin and Pekalska, 2010) have criticized the typical restriction to metric similarity measures. In (Duin and Pekalska, 2010) it is shown that many real life problems are better addressed by e.g. kernel functions which are not restricted to be based on a metric. Non-metric measures (leading to kernels which are not positive semi-definite (non-psd)) are common in many disciplines. The use of divergence measures (Schnitzer et al., 2012; Zhang et al., 2009) is very popular for spectral data analysis in chemistry, geo- and medical sciences (Mwebaze et al., 2010; van der Meer, 2006), and are in general not metric. Also the popular Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978) algorithm provides a non-metric alignment score which is often used as a proximity measure between two one-dimensional functions of different length. In image processing and shape retrieval indefinite proximities are often obtained by means of the inner distance (Ling and Jacobs, 2007) - another non-metric measure. Further prominent examples for genuine non-metric proximity measures can be found in the field of bioinformatics where classical sequence alignment algorithms (e.g. smith-waterman score (Gusfield, 1997)) produce non-metric proximity values. Multiple authors argue that the non-metric part of the data contains valuable information and should not be removed (Scheirer et al., 2014; Pekalska and Duin, 2005). Furthermore, it has been shown (Loosli et al., 2016; Schleif and Tiňo, 2015) that work-arounds such as eigenspectrum modifications are often inappropriate or undesirable, due to a loss of information and problems with the out-of sample extension. Nevertheless they are still often used and can serve as a baseline approach. Due

**Corresponding author: Tel.: +49(0) 931 351 18127
e-mail: frank-michael.schleif@fhws.de (Frank-Michael Schleif)

Table 1: Overview of the different datasets. We provide the dataset size (N) and the origin of the indefiniteness. For vectorial data the indefiniteness is caused artificial by using the tanh kernel.

Dataset	#samples	proximity measure and data source
Sonatas	1068	normalized compression distance on midi files (Schleif and Tiňo, 2015)
Delft	1500	dynamic time warping (Schleif and Tiňo, 2015)
a1a	1605	tanh kernel (Luss and d'Aspremont, 2009)
zongker	2000	template matching on handwritten digits (Pekalska and Haasdonk, 2009)
prodom	2604	pairwise structural alignment on proteins (Pekalska and Haasdonk, 2009)
PolydistH57	4000	Hausdorff distance (Pekalska and Haasdonk, 2009)
chromo	4200	edit distance on chromosomes (Pekalska and Haasdonk, 2009)
Mushrooms	8124	tanh kernel (Srisuphab and Mitrpanont, 2009)
swiss-10k	$\approx 10k$	smith waterman alignment on protein sequences (Schleif and Tiňo, 2015)
checker-100k	100.000	tanh kernel (indefinite)
skin	245.057	tanh kernel (indefinite)(UCI, 2016)
checker	1 Mill	tanh kernel (indefinite)

to its strong theoretical foundations, Support Vector Machine (SVM) has been extended for indefinite kernels in a number of ways (Haasdonk, 2005; Luss and d'Aspremont, 2009; Gu and Guo, 2012). A recent survey on indefinite learning is given in (Schleif and Tiňo, 2015). In (Loosli et al., 2016) a stabilization approach was proposed to calculate a valid SVM model in the Krĕin space which can be directly applied on indefinite kernel matrices. This approach has shown great promise in a number of learning problems, but has intrinsically quadratic to cubic complexity and provides a dense decision model. The approach can also be used for the recently proposed indefinite Core Vector Machine (iCVM) (Schleif and Tiňo, 2017) which has better scalability but still suffers from the dense model. The initial sparsification approach of the iCVM proposed in (Schleif and Tiňo, 2017) is not always applicable and we will provide an alternative in this paper.

Another indefinite SVM formulation was provided in (Alabdulmohsin et al., 2016), but it is based on an empirical feature space technique, which changes the feature space representation. Additionally, the imposed input dimensionality scales with the number of input samples, which is unattractive in out of sample extensions.

The present paper improves the work of (Schleif and Tiňo, 2017) by providing a sparsification approach such that the otherwise very dense decision model becomes sparse again. The new decision function approximates the original one with high accuracy and makes the application of the model practical.

We now review the main parts of the Krĕin space SVM provided in (Loosli et al., 2016) showing why the obtained α-vector is dense. The effect is the same for to the Core Vector Machine as shown in (Schleif and Tiňo, 2017). For details on the iCVM derivation we refer the reader to (Schleif and Tiňo, 2017).

1. Learning with non-psd kernels

Learning with non-psd kernels can be a challenging problem and may occur very quickly as discussed before, if domain specific measure are used or simply due to noise. The metric violations cause negative eigenvalues in the eigenspectrum of the kernel matrix K, leading to non-psd similarity matrices or indefinite kernels. Many learning algorithms are based on kernel formulations which have to be symmetric and psd. The mathematical meaning of a kernel is the inner product in some Hilbert space (Shawe-Taylor and Cristianini, 2004). However, it is often loosely considered simply as a pairwise "similarity" measure between data items, leading to a similarity matrix S

If a particular learning algorithm requires the use of Mercer kernels and the similarity measure does not fulfill the kernel conditions, steps must be taken to ensure a valid model.

1.1. Eigenspectrum approaches

A natural way to address the indefiniteness problem and to obtain a psd similarity matrix is to correct the eigenspectrum of the original similarity matrix S. Popular strategies include *flipping, clipping and shift correction*. The non-psd similarity matrix S is decomposed by an eigen decomposition: $S = U\Lambda U^\top$, where U contains the eigenvectors of S and Λ contains the corresponding eigenvalues. One can now adapt the eigenvalues to get rid of the negative eigenvalues and to end up with a psd kernel.

Clip eigenvalue correction:. All negative eigenvalues in Λ are set to 0. Spectrum clip leads to the nearest psd matrix S in terms of the Frobenius norm (Higham, 1988).

Flip eigenvalue correction:. All negative eigenvalues in Λ are set to $\lambda_i := |\lambda_i| \ \forall i$ which at least keeps the absolute values of the negative eigenvalues and can be relevant if these eigenvalue contain important information (Pekalska and Duin, 2005).

Shift eigenvalue correction:. The shift operation was already discussed earlier by different researchers (Filippone, 2009) and modifies Λ such that $\Lambda := \Lambda - \min_{ij} \Lambda$. Spectrum shift enhances all the self-similarities by the amount of ν and does not change the similarity between any two different data points, but it may also increase the intrinsic dimensionality of the data space and amplify noise contributions.

In the experiments we will only compare with the clip and flip approach. The latter one is also an algorithmic part of the Krĕin space SVM model. If one of the former corrections is

applied to the input kernel any standard kernel based learning method like SVM can be used. One major drawback of these approaches is the rather complicated out of sample extension to new test points but also that the data representation may have changed completely, leading to inferior results.

1.2. Krĕin space SVM

The Krĕin Space SVM (KSVM) (Loosli et al., 2016), replaced the classical SVM minimization problem by a stabilization problem in the Krĕin space. The respective equivalence between the stabilization problem and a standard convex optimization problem was shown in (Loosli et al., 2016). Let $x_i \in X, i \in \{1, \ldots, N\}$ be training points in the input space X, with labels $y_i \in \{-1, 1\}$, representing the class of each point. The input space X is often considered to be $\mathbb{R}^d$, but can be any suitable space due to the kernel trick. For a given positive C, SVM is the minimum of the following regularized empirical risk functional

$$\begin{aligned} J_C(f,b) &= \min_{f\in\mathcal{H},b\in\mathbb{R}} \frac{1}{2}\|f\|^2_{\mathcal{H}} + C\cdot H(f,b) \qquad (1)\\ H(f,b) &= \sum_{i=1}^{N} \max(0, 1 - y_i(f(x_i)+b)) \end{aligned}$$

Using the solution of Equation (1) as $(f_C^*, b_c^*) := \arg\min J_C(f,b)$ one can introduce $\tau = H(f_C^*, b_C^*)$ and the respective convex quadratic program (QP)

$$\min_{f\in\mathcal{H},b\in\mathbb{R}} \frac{1}{2}\|f\|^2_{\mathcal{H}} \quad s.t. \sum_{i=1}^{N} \max(0, 1 - y_i(f(x_i)+b)) \le \tau \qquad (2)$$

where we detail the notation in the following. This QP can be also seen as the problem of retrieving the orthogonal projection of the null function in a Hilbert space $\mathcal{H}$ onto the convex feasible set. The view as a projection will help to link the original SVM formulation in the Hilbert space to a KSVM formulation in the Krein space. First we need a few definitions, widely following (Loosli et al., 2016). A Krĕin space is an *indefinite* inner product space endowed with a Hilbertian topology.

Definition 1 (Inner products and inner product space)**.** *Let $\mathcal{K}$ be a real vector space. An inner product space with an indefinite inner product $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ on $\mathcal{K}$ is a bi-linear form where all $f, g, h \in \mathcal{K}$ and $\alpha \in \mathbb{R}$ obey the following conditions:*

- *Symmetry: $\langle f, g\rangle_{\mathcal{K}} = \langle g, f\rangle_{\mathcal{K}}$,*
- *linearity: $\langle \alpha f + g, h\rangle_{\mathcal{K}} = \alpha\langle f, h\rangle_{\mathcal{K}} + \langle g, h\rangle_{\mathcal{K}}$*
- *and $\langle f, g\rangle_{\mathcal{K}} = 0\ \forall g \in \mathcal{K}$ implies $f = 0$.*

An inner product is positive definite if $\forall f \in \mathcal{K}$, $\langle f, f\rangle_{\mathcal{K}} \ge 0$, negative definite if $\forall f \in \mathcal{K}$, $\langle f, f\rangle_{\mathcal{K}} \le 0$, otherwise it is indefinite. A vector space $\mathcal{K}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ is called inner product space.

Definition 2 (Krĕin space and pseudo Euclidean space)**.** *An inner product space $(\mathcal{K}, \langle\cdot,\cdot\rangle_{\mathcal{K}})$ is a Krĕin space if there exist two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$ spanning $\mathcal{K}$ such that $\forall f \in \mathcal{K}$, $f = f_+ + f_-$ with $f_+ \in \mathcal{H}_+$, $f_- \in \mathcal{H}_-$ and $\forall f, g \in \mathcal{K}$, $\langle f, g\rangle_{\mathcal{K}} = \langle f_+, g_+\rangle_{\mathcal{H}_+} - \langle f_-, g_-\rangle_{\mathcal{H}_-}$. A finite-dimensional Krĕin-space is a so called pseudo Euclidean space (pE).*

If $\mathcal{H}_+$ and $\mathcal{H}_-$ are reproducing kernel hilbert spaces (RKHS), $\mathcal{K}$ is a reproducing kernel Krĕin space (RKKS). For details on RKHS and RKKS see e.g. (Pekalska and Duin, 2005). In this case the uniqueness of the functional decomposition (the nature of the RKHSs $\mathcal{H}_+$ and $\mathcal{H}_-$) is not guaranteed. In (Ong et al., 2004) the reproducing property is shown for a RKKS $\mathcal{K}$. There is a unique symmetric kernel $k(x, x)$ with $k(x,\cdot) \in \mathcal{K}$ such that the reproducing property holds (for all $f \in \mathcal{K}, f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{K}}$) and $k = k_+ - k_-$ where k_+ and k_- are the reproducing kernels of the RKHSs $\mathcal{H}_+$ and $\mathcal{H}_-$.

As shown in (Ong et al., 2004) for any symmetric non-positive kernel k that can be decomposed as the difference of two positive kernels k_+ and k_-, a RKKS can be associated to it. In (Loosli et al., 2016) it was shown how the classical SVM problem can be reformulated by means of a stabilization problem. This is necessary because a classical norm as used in Eq. (2) does not exist in the RKKS but instead the norm is reinterpreted as a projection which still holds in RKKS and is used as a regularization technique (Loosli et al., 2016). This allows to define SVM in RKKS (viewed as Hilbert space) as the orthogonal projection of the null element onto the set (Loosli et al., 2016):

$$S = \{f \in \mathcal{K}, b \in \mathbb{R} | H(f,b) \le \tau\} \text{ and } 0 \in \partial_b H(f,b)$$

where ∂_b denotes the sub differential with respect to b. The set S leads to a unique solution for SVM in a Krĕin space (Loosli et al., 2016). As detailed in (Loosli et al., 2016) one finally obtains a stabilization problem which allows one to formulate a SVM in a Krĕin space.

$$\mathrm{stab}_{f\in\mathcal{K},b\in\mathbb{R}} \frac{1}{2}\langle f, f\rangle_{\mathcal{K}} \quad \text{s.t.} \sum_{i=1}^{l} \max(0, 1 - y_i(f(x_i)+b)) \le \tau \quad (3)$$

where stab means stabilize as detailed in the following: In a classical SVM in RKHS the solution is regularized by minimizing the norm of the function f. In Krĕin spaces however minimizing such a norm is meaningless since the dot-product contains both the positive and negative components. That's why the regularization in the original SVM through minimizing the norm f has to be transformed in the case of Krĕin spaces into a min-max formulation, where we jointly minimize the positive part and maximize the negative part of the norm. The authors of (Ong et al., 2004) termed this operation the stabilization projection, or stabilization. Further mathematical details can also be found in (Hassibi, 1996). An example illustrating the relations between minimum, maximum and the projection/stabilization problem in the Krĕin space is illustrated in (Loosli et al., 2016).

In (Loosli et al., 2016) it is further shown that the stabilization problem Eq. (3) can be written as a minimization problem using a semi-definite kernel matrix. By defining a projection operator with transition matrices it is also shown how the dual RKKS problem for the SVM can be related to the dual in the RKHS. We refer the interested reader to (Loosli et al., 2016). One - finally - ends up with a flipping operator applied to the eigenvalues of the indefinite kernel matrix[1] K as well as

[1] Obtained by evaluating $k(x, y)$ for training points x, y.

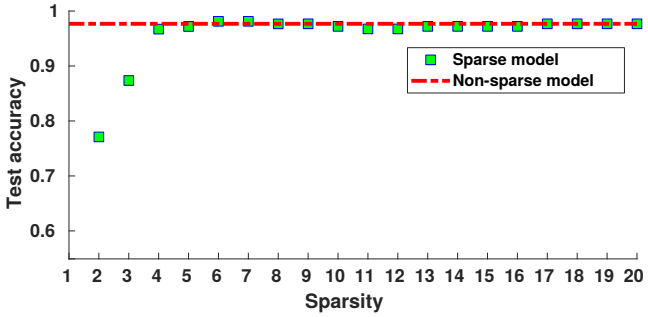


Fig. 1: Prediction results for the protein dataset using a varying level of sparsity and the OMP sparsity methods. For comparison the prediction accuracy of the non-sparse model is shown by a straight line.

to the α parameters obtained from the stabilization problem in the Krěin space, which can be solved using classical optimization tools on the flipped kernel matrix. This permits to apply the obtained model from the Krěin space directly on the non-positive input kernel without any further modifications. The algorithm is shown in Alg. 1. There are four major steps: 1) an eigen-decomposition of the full kernel matrix, with cubic costs (which can be potentially restricted to a few dominating eigenvalues - referred to as KSVM-L); 2) a flipping operation; 3) the solution of an SVM solver on the modified input matrix; 4) the application of the projection operator obtained from the eigen-decomposition on the α vector of the SVM model. U in Alg. 1

Algorithm 1 Krěin Space SVM (KSVM) - adapted from (Loosli et al., 2016).

Krěin SVM:
$[U, D]$:= EigenDecomposition(K)
$\hat{K} := USDU^\top$ with $S := sign(D)$
$[\alpha, b]$:= SVMSolver($\hat{K}$, Y, C)
$\tilde{\alpha} := USU^\top\alpha$ (now $\tilde{\alpha}$ is **dense**)
return $\tilde{\alpha}, b$;

contains the eigenvectors, D is a diagonal matrix of the eigenvalues and S is a matrix containing only $\{1, -1\}$ on the diagonal as obtained from the respective function sign.

As pointed out in (Loosli et al., 2016), this solver produces an exact solution for the stabilization problem. The main weakness of this Algorithm is, that it requires the user to pre-compute the whole kernel matrix and to decompose it into eigenvectors/eigenvalues. Further today's SVM solvers have a theoretical, worst case complexity of $\approx O(N^2)$. The other point to mention is that the final solution $\tilde{\alpha}$ is not sparse. The iCVM from (Schleif and Tiňo, 2017) has a similar derivation and leads to a related decision function, again with a dense $\tilde{\alpha}$, but the model fitting costs are $\approx O(N)$.

2. Sparsification of iCVM

2.1. Sparsification of iCVM by OMP

We can formalize the objective to approximate the decision function, which is defined by the $\tilde{\alpha}$ vector, obtained by KSVM or iCVM (both are structural identical), by a sparse alternative with the following mathematical problem:

$$\min \quad |\tilde{\alpha}|_0 \quad \text{s.t. } \textstyle\sum_m \tilde{\alpha}_m \Phi(x_m)^\top \Phi(x) \approx f(x)$$

It is well-known that this problem is NP hard in general, and a variety of approximate solution strategies exist in the literature. Here, we rely on a popular and very efficient approximation offered by **orthogonal matching pursuit** (OMP) (Geoffrey M. Davis, 1994; Pati et al., 1993). Given an acceptable error $\epsilon > 0$ or a maximum number n of non-vanishing components of the approximation, a greedy approach is taken: the algorithm iteratively determines the most relevant direction and the optimum coefficient for this axes to minimize the remaining residual error.

Algorithm 2 OMP to approximate the α vector.

1: **OMP:**
2: $I := \emptyset$; $r := y := K\tilde{\alpha}$; % initial residuum
3: **while** $|I| < n$ **do**
4: $l_0 := \text{argmax}_l |[Kr]_l|$; % find relevant direction + index
5: $I := I \cup \{l_0\}$ % track relevant indices
6: $\tilde{\gamma} := (K_{\cdot I})^+ \cdot y$ % restricted (inverse) projection
7: $r := y - (K_{\cdot I}) \cdot \tilde{\gamma}$ % residuum of the approximated decision function
8: **end while**
9: **return** $\tilde{\gamma}$ (as the new **sparse** $\tilde{\alpha}$)

In line 2 of Alg. 2 we define the initial residuum to be the vector $K\tilde{\alpha}$ as part of the decision function. In line 4 we identify the most contributing dimension (assuming an empirical feature space representation of our kernel - it becomes the dictionary). Then in line 6 we find the current approximation of the sparse $\tilde{\alpha}$-vector - called $\tilde{\gamma}$ to avoid confusion, where $^+$ indicates the pseudo inverse. In line 7 we update the residuum by removing the approximated $K\tilde{\alpha}$ from the original unapproximated one. A Nyström based approximation of the Algorithm 2 is straight forward using the concepts provided in (Gisbrecht and Schleif, 2015; Schleif and Gisbrecht, 2013). There it is also shown that the Nyström approximation holds for non-psd kernels, with a simplified proof given in (Oglic and Gärtner, 2019). With the Nyström technique a symmetric matrix psd (Williams and Seeger, 2000) or non-pdf (Gisbrecht and Schleif, 2015; Schleif and Gisbrecht, 2013) is approximated by a low-rank approach using a subset of the original datapoints, called landmarks. As shown in (Williams and Seeger, 2000; Gisbrecht and Schleif, 2015) this approximation is exact if the rank of original data is smaller or equal to the number of landmarks. The landmarks are often chosen randomly, with more advanced strategies proposed e.g. in (Musco and Musco, 2017)

2.2. Sparsification of iCVM by late subsampling

The parameters $\tilde{\alpha}$ are dense as already noticed in (Loosli et al., 2016). A naive sparsification by using only $\tilde{\alpha}_i$ with large absolute magnitude is not possible as can be easily checked by counter examples. One may now approximate $\tilde{\alpha}$ by using the (for this scenario slightly modified) OMP algorithm from the former section or by the following strategy, both compared in the experiments.

As a second sparsification strategy we can use the approach suggested by (Schleif and Tiňo, 2017), to restrict the projection operator and hence the transformation matrix of iCVM to

a subset of the original training data. We refer to this approach as iCVM-sparse-sub.

To get a consistent solution we have to recalculate parts of the eigen-decomposition as shown in Alg. 3. To obtain the respective subset of the training data we use the samples which are core vectors[2]. The number of core vectors is guaranteed to be very small (Tsang et al., 2006) and hence even for a larger number of classes the solution remains widely sparse. The suggested approach is given in Alg. 3. We assume that the original

Algorithm 3 Sparsification of iCVM by late subsampling

1: **Sparse iCVM:**
2: Apply iCVM - see (Schleif and Tiño, 2017)
3: ζ - vector of projection points by using the core set points
4: construct a reduced K' using indices ζ as $\bar{K}$
5: [U,D] := EigenDecomposition($\bar{K}$); $S := sign(D)$
6: $\bar{\alpha} := USU^\top\alpha$ %U restricted to core set indices
7: $\tilde{\alpha} := 0 \quad \tilde{\alpha}_\zeta := \bar{\alpha}$ % assign $\bar{\alpha}$ to $\tilde{\alpha}$ using indices of ζ
8: $b := Y\tilde{\alpha}^\top$ % recalculate bias using the sparse $\tilde{\alpha}$
9: **return** $\tilde{\alpha}, b$;

projection function (line 6 of Algorithm 3, detailed in (Loosli et al., 2016)), is smooth and can be potentially restricted to a small number of construction points with low error. We observed that in general few construction points are sufficient to keep high accuracy, as seen in the experiments.

3. Indefinite Core-Vector-Regression - iCVR

As already indicated in (Schleif and Tiño, 2017) the Krěin space approach considered before can also be used in similar minimum enclosing ball (MEB) based optimization problems. In particular we will consider the sparsification in the context of core vector regression for indefinite kernels, subsequently referred to as iCVR.

Assume points $x_i \in \mathbb{R}^d, i \in \{1, \ldots, N\}$ and real-valued outputs $y_i \in \mathbb{R}$ are given. Further, we assume a kernel function k (for the moment it is assumed this kernel is a psd kernel) is given with a feature map Φ. A kernel regression trains a function of the following form: $x \mapsto w^\top\Phi(x) + b$, where w is a normal vector of a decision plane and b a bias term. The training objective is to get as many points as possible approximately right while preserving a large margin. In the classical core vector regression (CVR) (Tsang et al., 2006) an ϵ-tube formalization is used to achieve this objective. For an ϵ-tube a data point is correctly predicted iff its image is within ϵ of the desired value. The corresponding dual core vector regression is described by (for details see (Tsang et al., 2006)):

$$\max_{\alpha_i,\alpha_i^*\geq 0,\sum\alpha_i+\alpha_i^*=1} -\frac{1}{2}\cdot\sum_{i,j}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(k_{ij}+1)$$
$$-\frac{1}{2}\sum_{i=1}^{N}\alpha_i^2/C-\frac{1}{2}\cdot\sum_{i=1}^{m}(\alpha_i^*)^2/C+\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)y_i$$

[2]A similar strategy for KSVM may be possible but is much more complicated because typically quite many points are support vectors and special sparse SVM solvers would be necessary.

This problem is of the form:

$$\max_{\alpha_i,\alpha_i^*\geq 0,\sum\alpha_i+\alpha_i^*=1} -\frac{1}{2}\begin{pmatrix}\alpha\\ \alpha^*\end{pmatrix}^\top\tilde{K}\begin{pmatrix}\alpha\\ \alpha^*\end{pmatrix}+\begin{pmatrix}\alpha\\ \alpha^*\end{pmatrix}^\top\begin{pmatrix}y\\ -y\end{pmatrix} \tag{4}$$

where

$$\tilde{K}=\begin{pmatrix}K+\mathbf{1}\mathbf{1}^\top+1/C\cdot\mathbf{I} & -(K+\mathbf{1}\mathbf{1}^\top)\\ -(K+\mathbf{1}\mathbf{1}^\top) & K+\mathbf{1}\mathbf{1}^\top+1/C\cdot\mathbf{I}\end{pmatrix}$$

$\tilde{K}$ is a valid kernel and as shown in (Tsang et al., 2006) a core / MEB algorithm can be used to solve Eq. (4). If the underlying kernel function k is indefinite also $\tilde{K}$ becomes indefinite. We can now use the same argumentation as for iCVM (Schleif and Tiño, 2017) and following the work in (Loosli et al., 2016) to modify the kernel $\tilde{K}$ by a flipping operation, to calculate a valid CVR model. Using the projection approach of (Loosli et al., 2016) the obtained solution vector can again be mapped into the Krěin space to obtain a model for iCVR. This final solution does not need any kernel modification for new test points to be applied. The whole algorithmic workflow to derive a iCVR model is described in Algorithm 4. Once more a Nyström

Algorithm 4 Calculating a iCVR model

1: **Indefinite CVR (iCVR):**
2: $[U, D]$ = EigenDecomposition(K)
3: $\hat{K} = USDU^\top$ with $S = sign(D)$
4: $[\alpha]$ = CoreVectorRegressionSolver($\hat{K}$, Y, C)
5: $\tilde{\alpha} = USU^\top\alpha \qquad b = Y\tilde{\alpha}^\top$
6: **return** $\tilde{\alpha}, b$;

approximation can be employed in Algorithm 4 line 2, for an indefinite kernel using the concepts proposed in (Gisbrecht and Schleif, 2015; Schleif and Gisbrecht, 2013) and also in line 4 following (Williams and Seeger, 2000) for psd kernel matrices.

One can easily see that the solution vector obtained in Algo. 4 is non-sparse. We will therefore apply again the two post-processing approaches suggested before to sparsify the iCVR model. This can be done in the same way as for iCVM with results given in the experimental section.

4. Experiments - iCVM

This part contains a series of experiments that show that our approach leads to a substantially lower complexity, while keeping similar prediction accuracy compared to the non-sparse approach. To allow for large datasets with two much hassle we provide sparse results only for the MEB approaches, namely iCVM and iCVR. The modified OMP approach will work also for sparse KSVM or KSVR but the late sampling sparsification is not well suited if many support vectors are given in the original model, asking for a sparse SVM implementation. We follow the experimental design given in (Loosli et al., 2016). Methods that require to modify test data are excluded as also done in (Loosli et al., 2016). Finally we compare the experimental complexity of the different solvers. The used data are explained in Table 1. Additional larger data sets have been added to motivate our approach in the line of learning with large scale indefinite kernels.

4.1. Experimental setting

For each dataset, we have run 20 times the following procedure: a random split to produce a training and a testing set, a 5-fold cross validation to tune each parameter (the number of parameters depending on the method) on the training set, and the evaluation on the testing set. If $N > 1000$ we use $m = 200$ randomly chosen landmarks from the given classes to approximate the kernel matrix using the Nyström technique. If the input data are vectorial data we used a tanh kernel with parameters $[a = 1, r = 1]$ to obtain an indefinite kernel. Where tanh is given as: $k(x, y) = \tanh(a < x, y > +r)$.

4.2. Results

In Table 2 we show the results for large scale data (having at least 1000 points) using iCVM with sparsification. We observe much smaller models, especially for larger datasets with often comparable prediction accuracy with respect to the non-sparse model. The runtimes are similar to the non-sparse case but in general slightly higher due to the extra eigen-decompositions on a reduce set of the data as shown in Algorithm 3.But the focus is not on a faster runtime (which is linear for iCVM and iCVR), but on a simple, sparse model and hence an easy out of sample extension. A typical result for the protein data set using the OMP-sparsity technique and various values for sparsity is shown in Figure 1.

5. Experiments - iCVR

We show the effectiveness of iCVR on a number of simulated and real life benchmark regression problems and compare with solutions as obtained by using standard CVR but for flipped (all signs of the eigenspectrum become positive) and clipped eigenspectra (negative einvalues are set to 0) of the respective kernel matrices. Data are given as $X \in \mathbb{R}^D$. Target function values $y_i \in \mathbb{R}^1$. The following one-dimensional simulated datasets have been used:

- (SIM1) basic sinc sample, with 200 samples, $f(x) = sinc(x/\pi) + 0.05 \cdot \sigma$ where σ is gaussian noise and x is linearly spread in $[-30, 30]$
- (SIM2) Friedman function, with 200 samples, $f(x) = 10 \cdot \sin(\pi \cdot \sigma_1 \cdot \sigma_2) + 20 \cdot (\sigma_3 - 0.5)^2 + 10 \cdot \sigma_4 + 5 \cdot \sigma_5 + \sigma$; and uniform noise $\sigma_1, \ldots, \sigma_5$, σ is gaussian noise
- (SIM3) The Mackey glass data,with 12000 samples, in 1 dimension as detailed in (Mackey and Glass, 1977)

Further we used the following real life regression datasets.

- (DS1) Abalone - age prediction, with 4177 samples, $D = 8$ taken from (Lichman, 2013)
- (DS2) Forest fires, with 517 samples, $D = 13$, dimension 13 was used as output variable, taken from (Cortez and Morais, 2007)
- (DS3) Breast cancer (radius) prediction, with 569 samples, $D = 32$, dimension 3 was used as output variable, taken from (Lichman, 2013) (wdbc)
- (DS4) White wine quality (scored 0-10), with 4898 samples, $D = 12$, dimension 12 was used as output variable, taken from (Cortez et al., 2009) [3]

The indefiniteness was caused using a Manhattan kernel $K_m = -\|X - X^\top\|$. The regression profiles for SIM1-SIM3 are depicted in Figure 2. In the experiments we apply the iCVR approach on the given datasets and compare it with the standard CVR algorithm were the indefinite input kernel was corrected by applying a flip or clip eigenspectrum transformation.

In Figure 3 a plot of the output function for SIM2 and its prediction using iCVR and CVR on a clipped kernel is shown. The plot shows substantial prediction errors on the clipped kernel in contrast to the prediction of iCVR with the indefinite mahalanobis kernel. Considering the results shown in Table 3 we observe that the clipping is in general worse than flipping or the iCVR. The sparse models of iCVR are in general only slightly worse than the non-sparse model. In parts we can even see a better performance of the sparse iCVR model (see last column) compared to the iCVR. This may be due to a denoising effect, caused by the implicit low rank approach used in OMP.

It should be noted that an application of the standard CVR on the indefinite kernels is not possible, which was also experimentally verified, because the obtained problem becomes non-convex and the solver is unable to provide a solution to the optimization problem.

6. Complexity analysis

The original KSVM has runtime costs (with full eigendecomposition) of $O(N^3)$ and memory storage $O(N^2)$, where N is the number of points. The iCVM or respectively iCVR involves an extra Nyström approximation of the kernel matrix to obtain $K_{(N,m)}$ and $K^{-1}_{(m,m)}$, if not already given. If we have m landmarks, $m \ll N$, this gives memory costs of $O(mN)$ for the first matrix and $O(m^3)$ for the second, due to the matrix inversion. Further a Nyström approximated eigendecomposition has to be done to apply the eigenspectrum flipping operator. This leads to runtime costs of $O(N \times m^2)$. The runtime costs for the sparse iCVM/iCVR are $O(N \times m^2)$ and the memory complexity is the same as for iCVM/iCVR. Due to the used Nyström approximation the prior costs only hold if $m \ll N$, which is the case for many datasets as shown in the experiments.

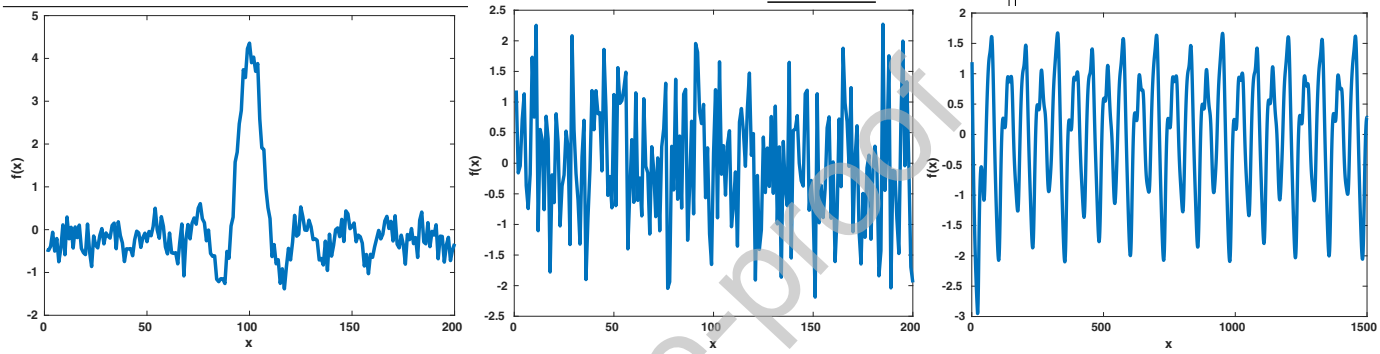
The application of a new point to a KSVM, iCVM or iCVR model requires the calculation of kernel similarities to all N training points, for the sparse iCVM/iCVR this holds only in the worst case. In general the sparse iCVM/iCVR provides a simpler out of sample extension as shown in Table 2, but is data dependent.

The (i)CVM/(i)CVR model generation has not more than N iterations or even a constant number of 59 points, if the probabilistic sampling trick is used (Tsang et al., 2006; Smola and Schölkopf, 2000). As show in (Tsang et al., 2006) the classical CVM has runtime costs of $O(1/\epsilon^2)$. The evaluation of a kernel function using the Nyström approximated kernel can be done

[3] Available at: `http://www3.dsi.uminho.pt/pcortez/wine/`

Table 2: Prediction errors (mean ± std.-dev.) on the test sets. The percentage of projection points (pts) is calculated using the unique set over core vectors over all classes in comparison to all training points. All sparse-OMP models use only 10 points in the final models. Best results are shown in bold. Best sparse results are underlined. Datasets with substantially reduced prediction accuracy are marked by ⊙ (anova $p < 5\%$).

Dataset	iCVM (sparse-sub)	pts	iCVM (sparse-OMP)	iCVM (non-sparse)
Sonatas	**12.64 ± 1.71**	76.84%	22.56 ± 4.16⊙	13.01 ± 3.82
Delft	16.53 ± 2.79⊙	52.48%	3.27 ± 0.6	**3.20 ± 0.84**
a1a	39.50 ± 2.88⊙	1.25%	27.85 ± 2.8	**20.56 ± 1.34**
zongker	29.20 ± 2.48⊙	52.81%	7.50 ± 1.7	**6.40 ± 2.11**
prodom	2.89 ± 1.17	26.31%	3.12 ± 0.11	**0.87 ± 0.64**
PolydistH57	6.12 ± 1.38	12.92%	29.35 ± 8⊙	**0.70 ± 0.19**
chromo	11.50 ± 1.17	33.76%	**3.74 ± 0.58**	6.10 ± 0.63
Mushrooms	7.84 ± 2.21	6.46%	18.39 ± 5.7⊙	**2.54 ± 0.56**
swiss-10k	35.90 ± 2.52⊙	17.03%	**6.73 ± 0.72**	12.08 ± 3.47
checker-100k	**8.54 ± 2.35**	2.26%	19.54 ± 2.1⊙	9.66 ± 2.32
skin	9.38 ± 3.30	0.06%	9, 43 ± 2.41	**4.22 ± 1.11**
checker	8.94 ± 0.84	0.24%	**1.44 ± 0.3**	9.38 ± 2.73



(a) SIM1 (b) SIM2 (c) SIM3

Fig. 2: Plots of the simulated data.

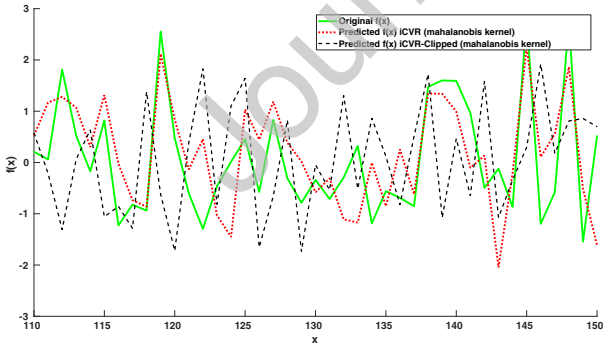


Fig. 3: Zoom in a plot of the Friedman output function (green, line). We also show the predicted output using CVR on a clipped mahalanobis kernel (black dashed+dotted) and a prediction of the output function using iCVR on the indefinite mahalanobis kernel (red, dashed).

with cost of $O(m^2)$ in contrast to constant costs if the full kernel is available. Accordingly, If we assume $m \ll N$ the overall runtime and memory complexity of iCVM/iCVR is linear in N, this is two magnitudes less as for KSVM for reasonable large N and for low rank input kernels.

7. Discussions and Conclusions

As discussed in (Loosli et al., 2016), there is no good reason to enforce positive-definiteness in kernel methods. A very detailed discussion on reasons for using KSVM or iCVM is given in (Loosli et al., 2016), explaining why a number of alternatives or pre-processing techniques are in general inappropriate. Our experimental results show that an appropriate Krěin space model provides very good prediction results and using one of the proposed sparsification strategies this can also be achieved for a sparse model in most cases. The proposed iCVM-sparse-OMP is only slightly better than the former iCVM-sparse-sub model with respect to the prediction accuracy but has very few final modeling vectors, with an at least competitive prediction accuracy in the vast majority of data sets. Similar observations are found for the iCVR in comparison to CVR with flipping or clipping. As is the case for KSVM, the presented approach can be applied without the need for transformation of test points, which is a desirable property for practical applications.

Table 3: Mean square error (mean ± std-dev.) in the 10-fold crossvalidation. The percentage of projection points (pts) is calculated using the unique set over core vectors over all classes in comparison to all training points. All sparse-OMP models use only 50 points in the final models. Best results are shown in bold.

Dataset	iCVR (non-sparse)	iCVR-flip	iCVR-clip	iCVR (sparse-sub)	pts	iCVR (sparse-OMP)
SIM1	**0.25 ± 0.12**	0.44 ± 0.43	0.46 ± 0.50	0.33 ± 0.13	17.25%	**0.25 ± 0.12**
SIM2	**0.14 ± 0.16**	0.15 ± 0.18	0.15 ± 0.16	0.15 ± 0.16	50%	0.15 ± 0.18
SIM3	**0.01 ± 0.00**	**0.01 ± 0.00**	**0.01 ± 0.00**	**0.01 ± 0.00**	18.68%	0.06 ± 0.01
DS1	0.83 ± 0.09	0.81 ± 0.07	2.46 ± 4.64	0.85 ± 0.06	7.83%	**0.77 ± 0.08**
DS2	1.34 ± 0.57	1.19 ± 0.38	2.15 ± 0.61	1.57 ± 1.08	5.03%	**1.12 ± 0.15**
DS3	**0.0 ± 0.0**	**0.01 ± 0.00**	**0.01 ± 0.00**	**0.02 ± 0.00**	23.04%	0.05 ± 0.00
DS4	1.17 ± 0.28	1.24 ± 0.23	1.20 ± 0.16	1.29 ± 0.19	5.22%	**0.75 ± 0.05**

Acknowledgments

We thank Gaelle Bonnet-Loosli for providing support with the Krěin Space SVM and R. Duin, Delft University for various support with distools and prtools. PT was supported by the EC Horizon 2020 ITN SUNDIAL (SUrvey Network for Deep Imaging Analysis and Learning), Project ID: 721463. FMS,CR are supported by the FuE program of the StMWi, project OBerA, grant number IUK-1709-0011// IUK530/010

References

Alabdulmohsin, I.M., Cissé, M., Gao, X., Zhang, X., 2016. Large margin classification with indefinite similarities. Machine Learning 103, 215–237.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems 47, 547–553.

Cortez, P., Morais, A., 2007. A Data Mining Approach to Predict Forest Fires using Meteorological Data, in: Neves, J., Santos, M.F., Machado, J. (Eds.), Proc. EPIA 2007, pp. 512–523.

Duin, R.P.W., Pekalska, E., 2010. Non-euclidean dissimilarities: Causes and informativeness, in: SSPR&SPR 2010, pp. 324–333.

Filippone, M., 2009. Dealing with non-metric dissimilarities in fuzzy central clustering algorithms. Int. J. of Approx. Reasoning 50, 363–384.

Geoffrey M. Davis, Stephane G. Mallat, Z.Z., 1994. Adaptive time-frequency decompositions. SPIE Journal of Optical Engineering 33, 21832191.

Gisbrecht, A., Schleif, F., 2015. Metric and non-metric proximity transformations at linear costs. Neurocomputing 167, 643–657.

Gu, S., Guo, Y., 2012. Learning SVM classifiers with indefinite kernels, in: Proc. of the 26th AAAI Conf. on AI, July 22-26, 2012.

Gusfield, D., 1997. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press.

Haasdonk, B., 2005. Feature space interpretation of SVMs with indefinite kernels. IEEE TPAMI 27, 482–492.

Hassibi, B., 1996. Indefinite metric spaces in estimation, control and adaptive filtering. Ph.D. thesis. Stanford Univ., Dept. of Elec. Eng.. Stanford.

Higham, N., 1988. Computing a nearest symmetric positive semidefinite matrix. Linear Algebra and Its Applications 103, 103–118.

Hodgetts, C., Hahn, U., 2012. Similarity-based asymmetries in perceptual matching. Acta Psychologica 139, 291–299.

Lichman, M., 2013. UCI machine learning repository. URL: `http://archive.ics.uci.edu/ml`.

Ling, H., Jacobs, D.W., 2007. Shape classification using the inner-distance. IEEE Trans. Pattern Anal. Mach. Intell. 29, 286–299.

Loosli, G., Canu, S., Ong, C.S., 2016. Learning svm in krein spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 1204–1216.

Luss, R., d'Aspremont, A., 2009. Support vector machine classification with indefinite kernels. Mathematical Programming Computation 1, 97–118.

Mackey, M., Glass, L., 1977. Oscillation and chaos in physiological control systems. Science 197, 287–289.

van der Meer, F., 2006. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. International Journal of Applied Earth Observation and Geoinformation 8, 3–17.

Musco, C., Musco, C., 2017. Recursive sampling for the nystrom method, in: Proc. of NIPS 2017, pp. 3836–3848.

Mwebaze, E., Schneider, P., Schleif, F.M., 2010. Divergence based classification in learning vector quantization. Neurocomputing 74, 1429–1435.

Oglic, D., Gärtner, T., 2019. Scalable learning in reproducing kernel krein spaces, in: Proc. of the 36th Int. Conf. on ML, ICML 2019, Long Beach, California, USA, pp. 4912–4921.

Ong, C.S., Mary, X., Canu, S., Smola, A.J., 2004. Learning with non-positive kernels, in: (ICML 2004).

Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S., 1993. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: Proc. 27th Conf. on Signals, Sys. & Comp., pp. 40–44.

Pekalska, E., Duin, R., 2005. The dissimilarity representation for pattern recognition. World Scientific.

Pekalska, E., Haasdonk, B., 2009. Kernel discriminant analysis for positive definite and indefinite kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 1017–1031.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on 26, 43–49.

Scheirer, W.J., Wilber, M.J., Eckmann, M., Boult, T.E., 2014. Good recognition is non-metric. Pattern Recognition 47, 2721–2731.

Schleif, F., Tiňo, P., 2015. Indefinite proximity learning: A review. Neural Computation 27, 2039–2096.

Schleif, F., Tiňo, P., 2017. Indefinite core vector machine. Pattern Recognition 71, 187–195.

Schleif, F.M., Gisbrecht, A., 2013. Data analysis of (non-)metric proximities at linear costs, in: Proceedings of SIMBAD 2013, pp. 59–74.

Schnitzer, D., Flexer, A., Widmer, G., 2012. A fast audio similarity retrieval method for millions of music tracks. Multimedia Tools and Appl. 58, 23–40.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis and Discovery. Cambridge University Press.

Smola, A.J., Schölkopf, B., 2000. Sparse greedy matrix approximation for machine learning, in: Langley, P. (Ed.), Proceedings of the 17th Int. Conf. on Machine Learning (ICML 2000), Morgan Kaufmann. pp. 911–918.

Srisuphab, A., Mitrpanont, J., 2009. Gaussian kernel approx. algorithm for feedforward nn design. Appl. Math. and Comp. 215, 2686–2693.

Tsang, I.W.H., Kwok, J.T.Y., Zurada, J.M., 2006. Generalized core vector machines. IEEE TNN 17, 1126–1140.

UCI, 2016. Skin segmentation database. URL: `http://archive.ics.uci.edu/ml/datasets/Skin+Segmentation`.

Vapnik, V., 2000. The nature of statistical learning theory. Statistics for engineering and information science, Springer.

Williams, C.K.I., Seeger, M., 2000. Using the Nyström Method to Speed Up Kernel Machines, in: Advances in Neural Information Processing Systems 13: NIPS'2000, pp. 682–688.

Xu, W., Wilson, R., Hancock, E., 2011. Determining the cause of negative dissimilarity eigenvalues. LNCS 6854 LNCS, 589–597.

Zhang, Z., Ooi, B.C., Parthasarathy, S., Tung, A.K.H., 2009. Similarity search on bregman divergence: Towards non-metric indexing. Proc. VLDB Endow. 2, 13–24.

Conflicts of interest

There is no conflict of interest.