Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Case Studies

Indefinite kernels in least squares support vector machines and principal component analysis [☆]

Xiaolin Huang [a,c,*], Andreas Maier [c], Joachim Hornegger [c], Johan A.K. Suykens [b]

[a] *Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 200240 Shanghai, PR China*
[b] *KU Leuven, ESAT-STADIUS, B-3001 Leuven, Belgium*
[c] *Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen–Nürnberg, 91058 Erlangen, Germany*

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 29 August 2016
Accepted 2 September 2016
Available online xxxx
Communicated by Charles K. Chui

Keywords:
Least squares support vector machine
Indefinite kernel
Classification
Kernel principal component analysis

ABSTRACT

Because of several successful applications, indefinite kernels have attracted many research interests in recent years. This paper addresses indefinite learning in the framework of least squares support vector machines (LS-SVM). Unlike existing indefinite kernel learning methods, which usually involve non-convex problems, the indefinite LS-SVM is still easy to solve, but the kernel trick and primal-dual relationship for LS-SVM with a Mercer kernel is no longer valid. In this paper, we give a feature space interpretation for indefinite LS-SVM. In the same framework, kernel principal component analysis with an infinite kernel is discussed as well. In numerical experiments, LS-SVM with indefinite kernels for classification and kernel principal component analysis is evaluated. Its good performance together with the feature space interpretation given in this paper imply the potential use of indefinite LS-SVM in real applications.

1. Introduction

Mercer's condition is the traditional requirement on the kernel applied in classical kernel learning methods, such as support vector machine with the hinge loss (C-SVM, [1]), least squares support vector machines (LS-SVM, [2,3]), and kernel principal components analysis (kPCA, [4]). However, in practice, one may meet sophisticated similarity or dissimilarity measures which lead to kernels violating Mercer's condition.

[☆] This work was partially supported by the Alexander von Humboldt Foundation and National Natural Science Foundation of China (61603248). Johan Suykens acknowledges support of ERC AdG A-DATADRIVE-B (290923), KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; FWO: G.0377.12, G.088114N, SBO POM (100031); IUAP P7/19 DYSCO.

[*] Corresponding author at: Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 200240 Shanghai, PR China.

E-mail addresses: xiaolinhuang@sjtu.edu.cn (X. Huang), anreas.maier@fau.de (A. Maier), joachim.hornegger@fau.de (J. Hornegger), johan.suykens@esat.kuleuven.be (J.A.K. Suykens).

http://dx.doi.org/10.1016/j.acha.2016.09.001
1063-5203/© 2016 Published by Elsevier Inc.

Since the kernel matrices induced by such kernels are real, symmetric, but not positive semi-definite, they are called *indefinite kernels* and the corresponding learning methodology is called *indefinite learning* [5–15].

Two important problems arise for indefinite learning. First, it lacks the classical feature space interpretation for a Mercer kernel, i.e., we cannot find a nonlinear feature mapping such that its inner-dot gives the value of an indefinite kernel function. Second, lack of positive definitiveness makes many learning models become non-convex if an indefinite kernel is used. In the last decades, there has been continuous progress aiming at these issues. In theory, indefinite learning in C-SVM has been discussed in the Reproducing Kernel Kreĭn Spaces (RKKS), cf. [5–8]. The kernel Fisher discriminant analysis with an indefinite kernel can be found in [9–11], which is also discussed on RKKS. In algorithm design, the current mainstream is to find an approximate positive semi-definite (PSD) kernel and then apply classical kernel learning algorithm based on that PSD kernel. These methods can be found in [12–14] and they have been reviewed and compared in [15]. That review also discusses directly applying indefinite kernels in some classical kernel learning methods which are not sensitive to metric violations. As suggested by [16], one can use an indefinite kernel to replace the PSD kernel in the dual formulation of C-SVM and solve it by sequential minimization optimization [17,18]. This kind of methods enjoy a similar computational efficiency as the classical learning methods and hence is more attractive in practice.

Following the way of introducing indefinite kernels to C-SVM, we consider indefinite learning based on LS-SVM. Notice that using an indefinite kernel in C-SVM results in a non-convex problem but indefinite learning based on LS-SVM is still easy to solve. However, Mercer's theorem is no longer valid. We have to find a new feature space interpretation and to give a characterization in terms of primal and dual problems, which are the theoretical targets of this paper. Since kPCA can be conducted in the framework of LS-SVM [19], we will discuss kPCA with an indefinite kernel as well.

This paper is organized as follows. Section 2 briefly reviews indefinite learning. Section 3 addresses LS-SVM with indefinite kernels and provides its feature space interpretation. Similar discussion on kPCA is given in Section 4. Then the performance of indefinite learning based on LS-SVM is evaluated by numerical experiments in Section 5. Finally, Section 6 gives a short conclusion.

2. Indefinite kernels

We start the discussion from C-SVM with a Mercer kernel. Given a set of training data $\{\mathbf{x}_i, y_i\}_{i=1}^m$ with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$, we are trying to construct a discriminant function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ and use its sign for classification. Except of linearly separable problems, a nonlinear feature mapping $\phi(\mathbf{x})$ is needed and the discriminant function is usually formulated as $f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b$. C-SVM trains $\mathbf{w}$ and b by the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum\nolimits_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ \forall i \in \{1, \dots, m\} \\ & \xi_i \geq 0,\ \forall i \in \{1, \dots, m\}. \end{aligned} \tag{1}$$

It is well known that the dual problem of (1) takes the formulation as below,

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m y_i \alpha_i \mathbf{K}_{ij} \alpha_j y_j \\ \text{s.t.} \quad & \sum\nolimits_{i=1}^m y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad \forall i \in \{1, \dots, m\}, \end{aligned} \tag{2}$$

where $\mathbf{K}_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$ is the kernel matrix. For any kernel $\mathcal{K}$ which satisfies Mercer's condition, there is always a feature map ϕ such that $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$. This allows us to construct a classifier to maximize the margin in the feature space without explicitly knowing ϕ.

Traditionally, in (2), we require the positiveness on $\mathbf{K}$. But in some applications, especially in computer version, there are many distances or dissimilarities, for which the corresponding matrices are not PSD [20–22]. It is also possible that though a kernel is PSD but is very hard to verify [5]. Even for a PSD kernel, noise may make the dissimilarity matrix non-PSD [23,24]. All these facts motivated the researchers to think about indefinite kernels in C-SVM. Notice that "indefinite kernels" literally cover many kernels, including asymmetric ones induced by asymmetric distances. But as all indefinite learning literature, we in this paper restrict "indefinite kernel" to the kernels that correspond to real symmetric indefinite matrices.

In theory, using indefinite kernels in C-SVM makes Mercer's theorem not applicable, which means that (1) and (2) are not a pair of primal-dual problems and then the solution of (2) cannot be explained as margin maximization in a feature space. Moreover, the learning theory and approximation theory about C-SVM with PSD is not valid, since the functional space spanned by indefinite kernels does not belong to any Reproducing Kernel Hilbert Space (RKHS). To link the indefinite kernel to RKHS, we need a *positive decomposition*. Its definition is given by [5] as follows: an indefinite kernel $\mathcal{K}$ has a positive decomposition if there are two PSD kernels $\mathcal{K}_+$, $\mathcal{K}_-$ such that

$$\mathcal{K}(\mathbf{u}, \mathbf{v}) = \mathcal{K}_+(\mathbf{u}, \mathbf{v}) - \mathcal{K}_-(\mathbf{u}, \mathbf{v}), \quad \forall \mathbf{u}, \mathbf{v}. \tag{3}$$

For an indefinite kernel $\mathcal{K}$ that has a positive decomposition, there exist Reproducing Kernel Kreĭn Spaces (RKKS). Conditions for the existence of positive decomposition are given by [5]. However, for a specific kernel, those conditions are usually hard to verify in practice. But at least, when the training data are given, the kernel matrix $\mathbf{K}$ has a decomposition which is the difference of two PSD matrices. Whether any indefinite kernel has a positive decomposition is still an open question. Fortunately, (3) is always valid for $\mathbf{u}, \mathbf{v} \in \{\mathbf{x}_i\}_{i=1}^m$. Thus, indefinite learning can be theoretically analyzed in RKKS and be implemented based on matrix decomposition in practice.

The feature space interpretation for indefinite learning is given by [24] for a finite-dimensional Kreĭn space, which is also called a pseudo-Euclidean (pE) space. A pE space is denoted as $\mathbb{R}^{(p,q)}$ with non-negative integers p and q. This space is a product of two Euclidean vector spaces $\mathbb{R}^p \times i\mathbb{R}^q$. An element in $\mathbb{R}^{(p,q)}$ can be represented by its coordinate vector and the coordinate vector gives the inner product: $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathrm{pE}} = \mathbf{u}^\top \mathbf{M} \mathbf{v}$, where $\mathbf{M}$ is a diagonal matrix with the first p components equal 1 and others equal to -1. If we link the components of $\mathbf{M}$ with the signs of eigenvalues of the indefinite kernel matrix $\mathbf{K}$, solving (2) for $\mathbf{K}$ is interpreted in [24] as distance minimization in $\mathbb{R}^{(p,q)}$. For learning behavior in RKKS, one can find the discussion on the space size in [5], error bound in [25], and asymptotic convergence in [26,27].

When an indefinite kernel is used in C-SVM, (2) becomes a non-convex quadratic problem, since $\mathbf{K}$ is not positive semi-definite. For a non-convex problem, many algorithms based on global optimality are invalid. An alternative way is to find an approximate PSD matrix $\tilde{\mathbf{K}}$ for an indefinite one $\mathbf{K}$, and then solve (2) for $\tilde{\mathbf{K}}$. To obtain $\tilde{\mathbf{K}}$, one can adjust the eigenvalues of $\mathbf{K}$ by: i) setting all negative values as zero [12]; ii) flipping signs of negative values [13]; iii) squaring the eigenvalues [26,27]. It also can be implemented by minimizing the Frobenius distance between $\mathbf{K}$ and $\tilde{\mathbf{K}}$, as introduced by [14]. Since training and classification are based on two different kernels, the above methods are efficient only when $\mathbf{K}$ and $\tilde{\mathbf{K}}$ are similar. Also those methods are time-consuming since they additionally involve eigenvalue problems. To pursue computational effectiveness, we can use descent algorithms, e.g., sequential minimization optimization (SMO) developed by [17,18], to directly solve (2) for an indefinite kernel matrix. Though only local optima are guaranteed, the performance is still promising, as reported by [15] and [16].

3. LS-SVM with real symmetric indefinite kernels

The current indefinite learning discussions are mainly for C-SVM. In this paper, we propose to use indefinite kernels in the framework of least squares support vector machines. In the dual space, LS-SVM is to solve the following linear system [2]:

$$\begin{bmatrix} 0 & \mathbf{y}^\top \\ \mathbf{y} & \mathbf{H} + \frac{1}{\gamma}\mathbf{I} \end{bmatrix} [b, \alpha_1, \ldots, \alpha_m]^\top = \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix}, \tag{4}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{1}$ is an all ones vector with the proper dimension, and $\mathbf{H}$ is given by

$$\mathbf{H}_{ij} = y_i y_j \mathbf{K}_{ij} = y_i y_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j).$$

We assume that the matrix in (4) is full rank. Then its solution can be effectively obtained and the corresponding discriminant function is

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i) + b.$$

The existing discussion about LS-SVM usually requires $\mathbf{K}$ to be positive semi-definite such that Mercer's theorem is applicable and the solution of (4) is related to Fisher discriminant analysis in feature space [28].

Now let us investigate indefinite kernels in LS-SVM (4). One good property is that even when $\mathbf{K}$ is indefinite, (4) is still easy to solve, which differs from C-SVM, where an indefinite kernel makes (2) non-convex. Though (4) with an indefinite kernel is easy to solve, the solution looses many properties of PSD kernels and its feature space interpretations have to be analyzed also in a pE space. This is based on the following proposition:

Proposition 1. *Let α^*, b^* be the solution of (4) for a symmetric but indefinite kernel matrix $\mathbf{K}$.*

i) There exist two feature maps ϕ_+ and ϕ_- such that

$$\mathbf{w}_+^* = \sum_{i=1}^{m} \alpha_i^* \phi_+(\mathbf{x}_i), \qquad \mathbf{w}_-^* = \sum_{i=1}^{m} \alpha_i^* \phi_-(\mathbf{x}_i),$$

which is a stationary point of the following primal problem:

$$\begin{aligned} \min_{\mathbf{w}_+, \mathbf{w}_-, b, \xi} \quad & \frac{1}{2}\left(\mathbf{w}_+^\top \mathbf{w}_+ - \mathbf{w}_-^\top \mathbf{w}_-\right) + \frac{\gamma}{2} \sum_{i=1}^{m} \xi_i^2 \\ \text{s.t.} \quad & y_i \left(\mathbf{w}_+^\top \phi_+(\mathbf{x}_i) + \mathbf{w}_-^\top \phi_-(\mathbf{x}_i) + b\right) = 1 - \xi_i, \forall i \in \{1, 2, \ldots, m\}. \end{aligned} \tag{5}$$

ii) The dual problem of (5) is given by (4), where

$$\mathbf{K}_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{K}_+(\mathbf{x}_i, \mathbf{x}_j) - \mathcal{K}_-(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

with two PSD kernels $\mathcal{K}_+$ and $\mathcal{K}_-$:

$$\mathcal{K}_+(\mathbf{x}_i, \mathbf{x}_j) = \phi_+(\mathbf{x}_i)^\top \phi_+(\mathbf{x}_j), \tag{7}$$

and

$$\mathcal{K}_-(\mathbf{x}_i, \mathbf{x}_j) = \phi_-(\mathbf{x}_i)^\top \phi_-(\mathbf{x}_j). \tag{8}$$

Please cite this article in press as: X. Huang et al., Indefinite kernels in least squares support vector machines and principal component analysis, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.09.001

Proof. For an indefinite kernel $\mathcal{K}$, we can always find two PSD kernels $\mathcal{K}_+$, $\mathcal{K}_-$ and the corresponding feature maps ϕ_+, ϕ_- to satisfy (6)–(8). Using ϕ_+ and ϕ_- in (5), its Lagrangian of (5) can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{w}_+, \mathbf{w}_-, b, \xi; \alpha) = \frac{1}{2}\left(\mathbf{w}_+^\top \mathbf{w}_+ - \mathbf{w}_-^\top \mathbf{w}_-\right) + \frac{\gamma}{2}\sum_{i=1}^m \xi_i^2 \\ - \sum_{i=1}^m \alpha_i \left(y_i(\mathbf{w}_+^\top \phi_+(\mathbf{x}_i) + \mathbf{w}_-^\top \phi_-(\mathbf{x}_i) + b) - 1 + \xi_i\right).\end{aligned}$$

Then the condition of a stationary point yields

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{w}_+} &= \mathbf{w}_+ - \sum\nolimits_{i=1}^m \alpha_i y_i \phi_+(\mathbf{x}_i) = 0,\\ \frac{\partial \mathcal{L}}{\partial \mathbf{w}_-} &= -\mathbf{w}_- - \sum\nolimits_{i=1}^m \alpha_i y_i \phi_-(\mathbf{x}_i) = 0,\\ \frac{\partial \mathcal{L}}{\partial b} &= \sum\nolimits_{i=1}^m \alpha_i = 0,\\ \frac{\partial \mathcal{L}}{\partial \xi_i} &= \gamma \xi_i - \alpha_i = 0,\\ \frac{\partial \mathcal{L}}{\partial \alpha_i} &= y_i(\mathbf{w}_+^\top \phi_+(\mathbf{x}_i) + \mathbf{w}_-^\top \phi_-(\mathbf{x}_i) + b) - 1 + \xi_i = 0.\end{aligned}$$

Eliminating the primal variables $\mathbf{w}_+$, $\mathbf{w}_-$, ξ, we get the optimality conditions: $\sum\limits_{i=1}^m \alpha_i = 0$ and

$$y_i\left(\sum\nolimits_{j=1}^m \alpha_j \phi_+(\mathbf{x}_i)^\top \phi_+(\mathbf{x}_j) - \sum\nolimits_{j=1}^m \alpha_j \phi_-(\mathbf{x}_i)^\top \phi_-(\mathbf{x}_j)\right) - b - \frac{\alpha_i}{\gamma} = 0.$$

Substituting (6)–(8) into the above condition leads to (4). Therefore, (4) is the dual problem of (5). If α^*, b^* is the solution of (4), then b^* and

$$\mathbf{w}_+^* = \sum\nolimits_{i=1}^m \alpha_i^* \phi_+(\mathbf{x}_i), \qquad \mathbf{w}_-^* = \sum\nolimits_{i=1}^m \alpha_i^* \phi_-(\mathbf{x}_i)$$

satisfy the first-order optimality condition of (5), i.e., $\mathbf{w}_+^*$, $\mathbf{w}_-^*$, b^* is a stationary point of (5).

Proposition 1 gives the primal problem and a feature space interpretation for LS-SVM with an indefinite kernel. Its proof relies on the positive decomposition (6) on $\mathbf{K}$, which exists for all real symmetric kernel matrices. But it does not mean that we can find a positive decomposition for $\mathcal{K}$, i.e., (3) is not necessarily valid. The verification is usually hard for a specific kernel. If such kernel decomposition exists, Proposition 1 further shows that (4) is pursuing a small within-class scatter in a pE space $\mathbb{R}^{(p,q)}$. If not, the within-class scatter is minimized in a space associated with an approximate kernel $\tilde{\mathcal{K}} = \mathcal{K}_+ - \mathcal{K}_-$ which is equal to $\mathcal{K}$ on all the training data. In (7) and (8), the dimension of the feature map could be indefinite and then the conclusion is extended to the corresponding RKKS.

4. Real symmetric indefinite kernel in PCA

In the last section, we considered LS-SVM with an indefinite kernel for binary classification. The analysis is applicable to other tasks which can be solved in the framework of LS-SVM. In [19], the link between kernel principal component analysis and LS-SVM has been investigated. Accordingly, we can give the feature space interpretation for kernel PCA with an indefinite kernel.

For given data $\{\mathbf{x}_i\}_{i=1}^m$, the kernel PCA is to solve an eigenvalue problem:

$$\Omega\alpha = \lambda\alpha, \tag{9}$$

where the centered kernel matrix Ω is induced from a kernel $\mathcal{K}$ as follows,

$$\Omega_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^m \mathcal{K}(x_i, x_r) - \frac{1}{m}\sum_{r=1}^m \mathcal{K}(x_j, x_r) + \frac{1}{m^2}\sum_{r=1}^m\sum_{s=1}^m \mathcal{K}(x_r, x_s).$$

Traditionally, $\mathcal{K}$ is limited to be a PSD kernel. Then a Mercer kernel is employed and (9) maximizes the variance in the related feature space.

Following the same way of introducing an indefinite kernel in C-SVM or LS-SVM, we can directly use an indefinite kernel for kPCA (9). Notice that for an indefinite kernel, the eigenvalues will be positive and negative. All these eigenvalues will be still real for the use of a symmetric kernel. There is no difference on the problem itself and the projected variables can be calculated as the same. However, the feature space interpretation fundamentally changes, which is discussed in the following proposition.

Proposition 2. *Let α^* be the solution of (9) for an indefinite kernel $\mathcal{K}$.*

i) There are two feature maps ϕ_+ and ϕ_- such that

$$\mathbf{w}_+^* = \sum_{i=1}^m \alpha_i^*(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}),$$
$$\mathbf{w}_-^* = \sum_{i=1}^m \alpha_i^*(\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}),$$

which is a stationary point of the following primal problem:

$$\begin{aligned} \max_{\mathbf{w}_+,\mathbf{w}_-,\xi} \quad & \frac{\gamma}{2}\sum_{i=1}^m \xi_i^2 - \frac{1}{2}\left(\mathbf{w}_+^\top\mathbf{w}_+ - \mathbf{w}_-^\top\mathbf{w}_-\right) \\ \text{s.t.} \quad & \xi_i = \mathbf{w}_+^\top(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) + \mathbf{w}_-^\top(\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}), \quad \forall i \in \{1,\dots,m\}. \end{aligned} \tag{10}$$

Here, $\hat{\mu}_{\phi_+}$, $\hat{\mu}_{\phi_-}$ are the centering terms, i.e.,

$$\hat{\mu}_{\phi_+} = \frac{1}{m}\sum_{i=1}^m \phi_+(\mathbf{x}_i) \quad \text{and} \quad \hat{\mu}_{\phi_-} = \frac{1}{m}\sum_{i=1}^m \phi_-(\mathbf{x}_i).$$

ii) If we choose γ as $\gamma = \frac{1}{\lambda}$ and decompose $\mathcal{K}$ as in (6)–(8), then the dual problem of (10) is given by (9).

Proof. Again, for an indefinite kernel $\mathcal{K}$, we can find two PSD kernels $\mathcal{K}_+$, $\mathcal{K}_-$, and the corresponding nonlinear feature maps ϕ_+, ϕ_- to satisfy (6)–(8).

The Lagrangian of (10) can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{w}_+, \mathbf{w}_-, \xi; \alpha) = {} & \frac{1}{2}\left(\mathbf{w}_+^\top\mathbf{w}_+ - \mathbf{w}_-^\top\mathbf{w}_-\right) - \frac{\gamma}{2}\sum_{i=1}^m \xi_i^2 \\ & - \sum_{i=1}^m \alpha_i\left(\mathbf{w}_+^\top(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) + \mathbf{w}_-^\top(\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) - \xi_i\right). \end{aligned}$$

Then from the conditions of a stationary point, we have

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}_+} &= \mathbf{w}_+ - \sum\nolimits_{i=1}^{m} \alpha_i (\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) = 0,\\
\frac{\partial \mathcal{L}}{\partial \mathbf{w}_-} &= -\mathbf{w}_- - \sum\nolimits_{i=1}^{m} \alpha_i (\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) = 0,\\
\frac{\partial \mathcal{L}}{\partial \xi_i} &= -\gamma \xi_i + \alpha_i = 0,\\
\frac{\partial \mathcal{L}}{\partial \alpha_i} &= \mathbf{w}_+^\top (\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) + \mathbf{w}_-^\top (\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) - \xi_i = 0.
\end{aligned}
$$

Elimination of the primal variables results in the following optimality condition,

$$
\begin{aligned}
&\frac{1}{\gamma}\alpha_i - \sum\nolimits_{j=1}^{m} \alpha_j (\phi_+(\mathbf{x}_j) - \hat{\mu}_{\phi_+})^\top (\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+})\\
&\quad - \sum\nolimits_{j=1}^{m} \alpha_j (\phi_-(\mathbf{x}_j) - \hat{\mu}_{\phi_-})^\top (\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) = 0, \quad \forall i \in \{1, 2, \ldots, m\}.
\end{aligned}
\tag{11}
$$

Applying the kernel trick (7) and (8), we know that

$$
\begin{aligned}
&(\phi_\pm(\mathbf{x}_i) - \hat{\mu}_{\phi_+})^\top (\phi_\pm(\mathbf{x}_j) - \hat{\mu}_{\phi_\pm})\\
&= \mathcal{K}_\pm(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^{m} \mathcal{K}_\pm(\mathbf{x}_i, \mathbf{x}_r) - \sum_{r=1}^{m} \mathcal{K}_\pm(x_j, x_r) + \frac{1}{m^2}\sum_{r=1}^{m}\sum_{s=1}^{m} \mathcal{K}_\pm(x_r, x_s).
\end{aligned}
$$

Additionally with (6), the optimality condition (11) can be formulated as the eigenvalue problem (9). Therefore, (9) is the dual problem of (10) and gives a stationary solution for (10), which aims at having maximal variance as the same as kPCA with PSD kernels.

5. Numerical experiments

In the preceding sections, we discussed the use of indefinite kernels in the framework of LS-SVM for classification and kernel principal component analysis, respectively. The general conclusions are: i) indefinite LS-SVM shares the same optimization model as the PSD ones and hence the same toolbox, namely LS-SVMlab [35], is applicable; ii) on the computational load of LS-SVM, there is no difference between a PSD kernel and an indefinite kernel, i.e., using an indefinite kernel in LS-SVM will not additionally bring computational burden; iii) the feature space interpretation of LS-SVM for an indefinite kernel is extended to a pE space and only a stationary point can be obtained.

In theory, indefinite kernels are the generalization of PSD ones, which are constrained to have zero-negative parts in (6). In practice, there are indefinite kernels successfully applied in specific applications [29–31]. Now with the feature space interpretation given in this paper, one can use LS-SVM and its modifications to learn from an indefinite kernel. In general, algorithmic properties holding for LS-SVM with PSD kernels are still valid when an indefinite kernel is used.

In this section, we will test the performance of LS-SVM with indefinite kernels on some benchmark problems. It should be noticed that the performance heavily relies on the choice of kernel. Though there are already some indefinite kernels designed to specific tasks, it is still hard to find an indefinite kernel for a wide range of problems. Therefore, PSD kernels, especially the radial basis function (RBF) kernel and the polynomial kernel, are currently dominant in kernel learning. One challenger from indefinite kernels is the tanh kernel, which has been evaluated in the framework of C-SVM [8,16]. Another possible indefinite kernel

Table 1
Test accuracy of LS-SVM with PSD and indefinite kernels.

Dataset	m	n	RBF	poly	tanh	TL1	TL1
			(CV)	(CV)	(CV)	$\rho = 0.7n$	(CV)
DBWords	32	242	84.3%	85.6%	75.0%	85.2%	84.4%
Fertility	50	9	86.7%	80.4%	83.8%	86.7%	87.8%
Planning	91	12	70.2%	67.9%	71.6%	70.6%	73.6%
Sonar	104	60	84.5%	83.1%	72.9%	84.3%	83.6%
Statlog	135	13	81.4%	75.2%	82.7%	83.8%	83.5%
Monk1	124	6	79.1%	78.3%	76.6%	73.4%	85.2%
Monk2	169	6	84.1%	75.6%	69.9%	53.4%	83.7%
Monk3	122	6	93.5%	93.5%	88.0%	97.2%	97.2%
Climate	270	20	93.2%	91.7%	92.6%	91.9%	92.0%
Liver	292	10	69.0%	67.9%	67.2%	69.7%	71.8%
Austr.	345	14	85.0%	85.1%	86.6%	86.0%	86.7%
Breast	349	10	96.4%	95.7%	96.6%	97.0%	97.1%
Trans.	374	4	78.3%	78.5%	78.9%	78.4%	78.4%
Splice	1000	121	89.4%	87.3%	90.1%	93.6%	94.9%
Spamb.	2300	57	93.1%	92.4%	91.7%	94.1%	94.2%
ML-prove	3059	51	72.5%	74.6%	71.8%	79.1%	79.3%

is a truncated ℓ_1 distance (TL1) kernel, which has been recently proposed in [32]. The mentioned kernels are listed below:

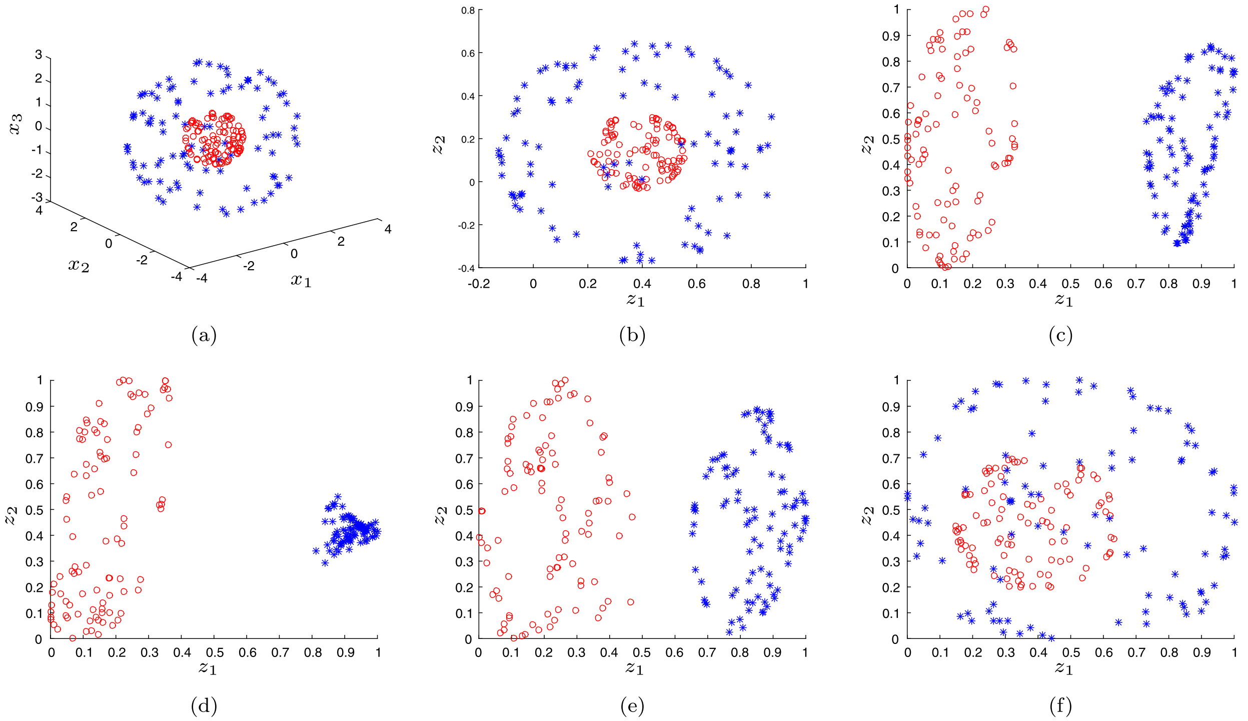
- PSD kernels:
  - linear kernel: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}$,
  - RBF kernel with parameter σ: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = \exp\left(-\|\mathbf{u} - \mathbf{v}\|_{\ell_2}^2 / \sigma^2\right)$,
  - polynomial kernel with parameters $c \geq 0$, $d \in \mathcal{N}^+$: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^\top \mathbf{v} + c)^d$.
- indefinite kernels:
  - tanh kernel with parameters c, d[1]: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = \tanh(c\mathbf{u}^\top \mathbf{v} + d)$,
  - TL1 kernel with parameter ρ: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = \max\{\rho - \|\mathbf{u} - \mathbf{v}\|_{\ell_1}, 0\}$.

These kernels will be compared in the framework of LS-SVM for both classification and principal component analysis. First, consider binary classification problems, for which the data are downloaded from UCI Repository of Machine Learning Datasets [36]. For some datasets, there are both training and test data. Otherwise, we randomly pick half of the data for training and the rest for test. All training data are normalized to $[0, 1]^n$ in advance. In training procedure, there are a regularization coefficient and kernel parameters, which are tuned by 10-fold cross validation. Specifically, we randomly partition the training data into 10 subsets. One of these subsets is used for validation in turn and the remaining ones for training. As discussed in [32], the performance of the TL1 kernel is not very sensitive to the value of ρ and $\rho = 0.7n$ was suggested. We thus also evaluate the TL1 kernel with $\rho = 0.7n$. With one parameter less, the training time can be largely saved.

The above procedure is repeated 10 times, and then the average classification accuracy on test data are reported in Table 1, where the number of training set m and the problem dimension n are given as well. The best one for each dataset in the sense of average accuracy is underlined. The results confirm the potential use of indefinite kernels in LS-SVM: an indefinite kernel can achieve similar accuracy as a PSD kernel in most of the problems and can have better performance in some specific problems. This does not mean that indefinite kernel surely improves the performance from PSD ones but for some datasets, e.g., Monk1, Monk3, and Splice, it is worthy to consider indefinite learning with LS-SVM which may have better accuracy within almost the same training time. Moreover, this experiment, for which the performance of

[1] The tanh kernel is conditionally positive definite (CPD) when $c \geq 0$ and is indefinite otherwise; see, e.g., [33,34]. In our experiments, we consider both positive and negative c, and hence the tanh kernel is regarded as an indefinite kernel.

Please cite this article in press as: X. Huang et al., Indefinite kernels in least squares support vector machines and principal component analysis, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.09.001



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1. (a) Data points of one class come from the unit sphere and are marked by red circles. The other data points, shown by blue stars, come from a sphere with radius 3. (b) This dataset is not linearly separable and thus linear PCA is not helpful for distinguishing the two classes. Instead, kernel PCA is needed and if the parameter is suitably chosen, the reduced data can be correctly classified by a linear classifier. (c) the RBF kernel with $\sigma = 0.05$; (d) the TL1 kernel with $\rho = 0.1n$; (e) the TL1 kernel with $\rho = 0.2n$; (f) the TL1 kernel with $\rho = 0.5n$, which is similar to linear PCA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the TL1 kernel with $\rho = 0.7n$ being satisfactory for many datasets, illustrates the good parameter stability of the TL1 kernel.

In the following, we use indefinite kernels for principal component analysis. As an intuitive example, we consider a 3-dimensional sphere problem that distinguishs data from the sphere with radius equal to 1 and 3. The data are shown in Fig. 1(a). To reduce the dimension, we apply PCA, kPCA with the RBF kernel, and kPCA with the TL1 kernel, respectively. The obtained two dimensional data are displayed in Fig. 1(b)–(f), which roughly imply that a suitable indefinite kernel can be used for kernel principal component analysis.

To quantitatively evaluate kPCA with indefinite kernels, we choose the problems of which the dimension is higher than 20 from Table 1 and then apply kPCA to reduce the data into n_r dimension. For the reduced data, linear classifiers, trained from linear C-SVM with libsvm [37], are used to classify the test data. The parameters, including kernel parameters and the regularization constant in linear C-SVM, are tuned based on 10-fold cross-validation. In Table 2, the average classification accuracy of 10 trials for different reduction ratios n_r/n is listed. The results illustrate that indefinite kernels can be used for kPCA. Its performance in general is comparable to PSD kernels and for some datasets the performance is significantly improved.

Summarizing all the experiments above, we observe the potential use of indefinite kernels in LS-SVM for classification and kPCA. For example, the TL1 kernel has similar performance as the RBF kernel in many problems and has much better results for several datasets. Our aim in this experiment is not to claim which kernel is the best, which actually depends on the specific problem. Instead, we show that for some problems, a proper indefinite kernel can significantly improve the performance from PSD ones, which may motivate the researchers to design indefinite kernels and use them in LS-SVMs.

Please cite this article in press as: X. Huang et al., Indefinite kernels in least squares support vector machines and principal component analysis, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.09.001

Table 2
Test accuracy based on kPCA with different reduction ratios.

Dataset	m	n	Ratio	Linear	RBF	poly	tanh	TL1
Sonar	104	60	10%	72.6%	75.6%	75.2%	63.8%	77.9%
			30%	73.1%	79.1%	78.2%	71.0%	80.4%
			50%	75.9%	80.7%	79.0%	71.9%	81.9%
Climate	270	21	10%	90.4%	90.5%	91.4%	91.5%	90.5%
			30%	90.9%	90.8%	91.4%	91.6%	90.9%
			50%	91.6%	91.4%	93.9%	91.6%	91.9%
Qsar	528	41	10%	74.4%	77.8%	75.5%	77.5%	78.8%
			30%	85.4%	86.4%	84.1%	84.5%	85.9%
			50%	85.9%	86.7%	85.4%	86.0%	86.2%
Splice	1000	60	10%	83.7%	86.6%	85.5%	83.7%	91.9%
			30%	83.9%	87.7%	85.3%	83.0%	91.1%
			50%	84.1%	87.8%	86.5%	85.2%	91.3%
Spamb.	2300	57	10%	84.7%	86.5%	84.9%	86.4%	88.4%
			30%	87.7%	89.9%	88.3%	89.9%	91.8%
			50%	90.7%	91.0%	91.5%	92.8%	92.8%
ML-prove	3059	51	10%	59.2%	69.7%	64.0%	63.3%	70.1%
			30%	67.9%	70.3%	72.8%	69.3%	71.3%
			50%	70.2%	71.0%	73.1%	68.9%	75.5%

6. Conclusion

In this paper, we proposed to use indefinite kernels in the framework of least squares support vector machines. In the training problem itself, there is no difference between definite kernels and indefinite kernels. Thus, one can easily use an indefinite kernel in LS-SVM by simply changing the kernel evaluation function. Numerically, the indefinite kernels achieve good performance compared with commonly used PSD kernels for both classification and kernel principal component analysis. The good performance motivates us to investigate the feature space interpretation for an indefinite kernel in LS-SVM, which is the main theoretical contribution of this paper. We hope that the theoretical analysis and good performance shown in this paper can attract research and application interests on indefinite LS-SVM and indefinite kPCA in the future.

Acknowledgments

The authors would like to thank Prof. Lei Shi in Fudan University for helpful discussion. The authors are grateful to the anonymous reviewer for insightful comments.

References

[1] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[2] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.
[3] J.A.K. Suykens, T. Van Gestel, B. De Moor, J. De Brabanter, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, 2002.
[4] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[5] C.S. Ong, X. Mary, S. Canu, A.J. Smola, Learning with non-positive kernels, in: Proceeding of the 21st International Conference on Machine Learning (ICML), 2004, pp. 639–646.
[6] Y. Ying, C. Campbell, M. Girolami, Analysis of SVM with indefinite kernels, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22, 2009, pp. 2205–2213.
[7] S. Gu, Y. Guo, Learning SVM classifiers with indefinite kernels, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012, pp. 942–948.
[8] G. Loosli, S. Canu, C.S. Ong, Learning SVM in Kreĭn spaces, IEEE Trans. Pattern Anal. Mach. Intell. 38 (6) (2016) 1204–1216.
[9] E. Pekalska, B. Haasdonk, Kernel discriminant analysis for positive definite and indefinite kernels, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 1017–1032.
[10] B. Haasdonk, E. Pekalska, Indefinite kernel discriminant analysis, in: Proceedings of the 16th International Conference on Computational Statistics (COMPSTAT), 2010, pp. 221–230.

[11] S. Zafeiriou, Subspace learning in Kreĭn spaces: complete kernel Fisher discriminant analysis with indefinite kernels, in: Proceedings of European Conference on Computer Vision (ECCV) 2012, 2012, pp. 488–501.

[12] E. Pekalska, P. Paclik, R.P. Duin, A generalized kernel approach to dissimilarity-based classification, J. Mach. Learn. Res. 2 (2002) 175–211.

[13] V. Roth, J. Laub, M. Kawanabe, J.M. Buhmann, Optimal cluster preserving embedding of nonmetric proximity data, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1540–1551.

[14] R. Luss, A. d'Aspremont, Support vector machine classification with indefinite kernels, in: J.C. Platt, D. Koller, Y. Singer, S.T. Roweis (Eds.), Advances in Neural Information Processing Systems 20, 2007, pp. 953–960.

[15] F. Schleif, P. Tino, Indefinite proximity learning: A review, Neural Comput. 27 (2015) 2039–2096.

[16] H.-T. Lin, C.-J. Lin, A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods, technical report, Department of Computer Science, National Taiwan University, 2003, http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf.

[17] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in: Advances in Kernel Methods – Support Vector Learning, MIT Press, 1999, pp. 185–208.

[18] R.-E. Fan, P.-H. Chen, C.-J. Lin, Working set selection using second order information for training support vector machines, J. Mach. Learn. Res. 6 (2005) 1889–1918.

[19] J.A.K. Suykens, T. Van Gestel, J. Vandewalle, B. De Moor, A support vector machine formulation to PCA analysis and its kernel version, IEEE Trans. Neural Netw. 14 (2) (2003) 447–450.

[20] H. Ling, D.W. Jacobs, Using the inner-distance for classification of articulated shapes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005, 2005, pp. 719–726.

[21] M.M. Deza, E. Deza, Encyclopedia of Distances, Springer, New York, 2009.

[22] W. Xu, W. Richard, E. Hancock, Determining the cause of negative dissimilarity eigenvalues, in: Computer Analysis of Images and Patterns, Springer, New York, 2011, pp. 589–597.

[23] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, K. Obermayer, Classification on pairwise proximity data, in: M.S. Kearns, S.A. Solla, D.A. Cohn (Eds.), Advances in Neural Information Processing Systems 11, 1998, pp. 438–444.

[24] B. Haasdonk, Feature space interpretation of SVMs with indefinite kernels, IEEE Trans. Pattern Anal. Mach. Intell. 27 (4) (2005) 482–492.

[25] I. Alabdulmohsin, X. Gao, X. Zhang, Support vector machines with indefinite kernels, in: Proceedings of the 6th Asian Conference on Machine Learning (ACML), 2014, pp. 32–47.

[26] H. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularization, Appl. Comput. Harmon. Anal. 30 (1) (2011) 96–109.

[27] Q. Wu, Regularization networks with indefinite kernels, J. Approx. Theory 166 (2013) 1–18.

[28] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, J. Vandewalle, Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis, Neural Comput. 14 (5) (2002) 1115–1147.

[29] A.J. Smola, Z.L. Ovari, R.C. Williamson, Regularization with dot-product kernels, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems 13, 2000, pp. 308–314.

[30] H. Saigo, J.P. Vert, U. Ueda, T. Akutsu, Protein homology detection using string alignment kernels, Bioinformatics 20 (11) (2004) 1682–1689.

[31] B. Haasdonk, H. Burkhardt, Invariant kernel functions for pattern analysis and machine learning, Mach. Learn. 68 (1) (2007) 35–61.

[32] X. Huang, J.A.K. Suykens, S. Wang, A. Maier, J. Hornegger, Classification with truncated ℓ_1 distance kernel, internal report 15-211, ESAT-SISTA, KU Leuven, 2015.

[33] M.D. Buhmann, Radial Basis Functions: Theory and Implementations, Cambridge Monographs on Applied and Computational Mathematics, vol. 12, 2004, pp. 147–165.

[34] H. Wendland, Scattered Data Approximation, Cambridge University Press, Cambridge, UK, 2004.

[35] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, J.A.K. Suykens, LS-SVMlab toolbox user's guide, internal report 10-146, ESAT-SISTA, KU Leuven, 2011.

[36] A. Frank, A. Asuncion, UCI Machine Learning Repository, School of Information and Computer Science, University of California, Irvine, CA, 2010, http://archive.ics.uci.edu/ml.

[37] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.