UNIVERSITY OF SOUTHAMPTON

Kernel Ellipsoidal Trimming

T8.11.10-01/05

by

A.N. Dolia, C.J. Harris, J. Shawe-Taylor, D.M. Tittetington

T8.11.10-01/05

Technical Report

Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

October 11, 2005

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

by A.N. Dolia, C.J. Harris, J. Shawe-Taylor, D.M. Tittetington

T8.11.10-01/05

Ellipsoid estimation is an issue of primary importance in many practical areas such as control, system identification, visual/audio tracking, experimental design, data mining, robust statistics and novelty/outlier detection. This paper presents a new method of kernel information matrix ellipsoid estimation (KIMEE) that finds an ellipsoid in a kernel defined feature space based on a centered information matrix. Although the method is very general and can be applied to many of the aforementioned problems, the main focus in this paper is the problem of novelty or outlier detection associated with fault detection. A simple iterative algorithm based on Titterington's minimum volume ellipsoid method is proposed for practical implementation. The KIMEE method demonstrates very good performance on a set of real-life and simulated datasets compared with support vector machine methods.

Contents

1 Introduction **1**

2 The Minimum Volume Ellipsoid **3**

3 Working in high-dimensional feature spaces **5**

4 Some implementation details **8**
4.1 Some implementation details 8

5 Experiments **10**
5.0.1 Simulated dataset 10
5.0.2 Condition monitoring 12

6 Conclusions **16**

Bibliography **17**

List of Figures

5.1 In this example 4 outliers are ignored by the KIMEE method after retraining. Both the larger and smaller ellipses were found by Algorithm 1 using the simple inner product kernel $\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y}$ and $t = 0.001$. Support points on the boundaries of these two ellipsoids are denoted by *triangles* and *circles*, respectively. 11

5.2 In this example the sample of 100 points from a Gaussian distribution was contaminated by 10 outliers that belong to a Gaussian distribution with a different mean and covariance matrix. By the tuning parameter ν we can exclude the outliers but the ellipse is still not centered around the original data. 11

5.3 Illustration of the banana dataset with $\ell = 50$. In this example we use Algorithm 1 with the threshold $t = 0.001$ and a Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\rho^2)$, with (a) $\rho = 21$, (b) $\rho = 7$. 12

5.4 Illustration of dimensionality reduction and feature selection using a Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\rho^2)$, for (a) the banana and (b) condition monitoring datasets. 12

5.5 Condition monitoring experiment. In this example measurements, one channel, are obtained from a healthy machine and when four different faults are present [4, 39]. 13

5.6 Illustration of the features used in the condition monitoring experiment [4, 39]. 14

Chapter 1

Introduction

In practice, there are many applications such as control, system identification, visual/audio tracking, experimental design, data mining, robust statistics and novelty/outlier detection that can be solved by computing the minimum volume covering ellipsoid (MVCE) [32, 36, 34, 10] from training data $\{\mathbf{x}_i\}_{i=1}^{\ell}$. The MVCE must contain the entire training dataset. The problem becomes extremely difficult in multi-dimensional (possibly, infinite dimensional) space and in the presence of outliers [30, 24].

The MVCE problem is significant and has been extensively studied for over 50 years [32]. It can be considered as a special case of the more general maximum determinant problem [32, 36] and is related to D-optimum experimental design when the ellipsoid is centered at the origin [34, 36] (see [16] for an application of D-optimum experimental design to regression and model selection problems). There are a number of problems, such as novelty detection, in which it is possible to fit the minimum volume covering hypersphere (MVCS) around the data [27, 26, 28, 30]. Obviously, a hypersphere is a particular type of hyperellipsoid and therefore the volume contained by the hypersphere is usually larger than that estimated by the MVCE algorithm. Therefore, more outliers could be accepted by the MVCS algorithm than by the MVCE. In order to address this problem one can perform pre-whitening of the data points $\{\mathbf{x}_i\}_{i=1}^{\ell}$. Howerever, in the presence of outliers it is difficult to find such a linear transformation that will map the data with ellipsoidal support in the original space or in the kernel defined feature space to spherical-like support. The main reason for this observation is that, in practice, any method that finds this linear mapping, such as principal component analysis (PCA)[1, 7], is sensitive to outliers. The MVCE problem involves many interesting and sound theoretical results and optimization algorithms [32, 36, 34].

Our approach is mainly motivated by the theory of optimal experimental design, minimum volume ellipsoid estimation, ellipsoidal trimming, outlier detection [35, 34, 19, 20, 23, 15, 2] and treatments of the novelty detection or so-called one-class classification

problem [9, 3, 27, 26, 28, 4, 38, 21, 33]. The proposed KIMEE method finds an ellipsoid in the kernel defined feature space based on the centered information matrix. In this paper we do not try to find the most robust method for dealing with outliers, for example using the so-called soft margin approach and slack variables [28, 4, 32, 30, 10] or methods based on robust statistics [6, 7, 8, 24, 23, 25, 18]. A choice of robust method should be based on the problem we are trying to solve. The main objective of this paper is to show how the minimum volume covering ellipsoid can be found in the kernel-defined feature space [28, 4, 30]. We have chosen the outlier or novelty detection [3] context in order to demonstrate our method but it can be adapted to optimal experimental design for kernel ridge regression [19, 30], for example.

The paper is organized as follows. The MVCE model is described in Section 2. The main theoretical results of this work are given in Section 3. The iterative algorithm based on Titterington's minimum volume ellipsoid method is proposed in Section 4. The performance of the one-class support vector machine (SVM), the linear programming novelty detection algorithm (LPND) and the KIMEE method for the novelty/outlier detection problem is analyzed in Section 5 followed by conclusions and acknowledgements in Section 6.

Chapter 2

The Minimum Volume Ellipsoid

Assume that we have a training dataset containing ℓ samples, $\{\mathbf{x}_i \in \boldsymbol{\mathcal{R}}^{k\times 1}\}_{i=1}^{\ell}$. In order to solve the MVCE problem we need to obtain a $(k \times k)$ positive definite matrix $\mathbf{M} \in \boldsymbol{\mathcal{R}}^{k\times k}$ and the center of the ellipsoid $\mathbf{c}$ so as to minimize $\det \mathbf{M}$ subject to [35, 34, 32, 36]

$$(\mathbf{x}_i - \mathbf{c})'\mathbf{M}^{-1}(\mathbf{x}_i - \mathbf{c}) \leq k. \tag{2.1}$$

The dual optimization problem of the MVCE calculation (see (2.1)) has its roots in D-optimum experimental design and is that of maximizing $\log\det \mathbf{M}$ with respect to $\boldsymbol{\alpha}$, where $\mathbf{M} = \sum_{i=1}^{\ell} \alpha_i(\mathbf{x}_i - \mathbf{c})(\mathbf{x}_i - \mathbf{c})'$ and $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i\ \mathbf{x}_i$; $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_\ell\}$ are nonnegative numbers summing to 1. The matrix $\mathbf{M}$ might be called the "corrected" information matrix for the probability measure $\boldsymbol{\alpha}$ (see [35, 34, 32, 36] for details). The MVCE for the dataset $\{\mathbf{x}_i\}_{i=1}^{\ell}$ must go through at least $k+1$ and at most $\frac{1}{2}k(k+3)+1$ support points, that is, points $\mathbf{x}_i$ such that the corresponding α_i is greater than zero [35, 34, 19]. There could be more than $\frac{1}{2}k(k+3)+1$ points, for example, if all the data were on the surface of an ellipsoid. However $\frac{1}{2}k(k+3)+1$ is the largest number that are necessary.

The MVCE model with slack variables allows a fraction ν of data be outside the ellipsoid [10, 32]. In this case we obtain the following dual optimization problem (for a primal problem see the so-called MVCEP optimization method in [32], page 701) to be satisfied by the Lagrangian multipliers α [10]:

$$\begin{aligned} \min \quad & \varepsilon(\alpha) = \Psi(\mathbf{M}) \\ \text{s.t.} \quad & \\ & \sum_{i=1}^{\ell} \alpha_i = 1,\ 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \end{aligned} \tag{2.2}$$

where $\Psi(\cdot) = -\log\det(\cdot)$, $\mathbf{M} = \left\{\sum_{i=1}^{\ell} \alpha_i \mathbf{x}_i\mathbf{x}_i' - \mathbf{c}\mathbf{c}'\right\}$ and $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i\mathbf{x}_i$. The optimization criterion $-\log\det \mathbf{M}$ is strictly convex on the set of possible nonnegative

definite matrices $\mathbf{M}$ and therefore the optimization problem has a unique optimal solution for $\mathbf{M}$ and $\mathbf{c}$ but not necessarily for α. After obtaining values α (see (2.2)) in order to check if the test point $\mathbf{x}_t$ belongs to the estimated support we should employ the following rule $(\mathbf{x}_t - \mathbf{c})'\mathbf{M}^{-1}(\mathbf{x}_t - \mathbf{c}) > R^2$, where R^2 is the threshold; for example, $R^2 = k$.

Note that, if the function $\Psi(\cdot)$ is the trace of he centered information matrix $\mathbf{M}$, $\Psi(\mathbf{M}) = -\text{trace}(\mathbf{M})$, then the optimization problem (2.2) is a standard quadratic optimization problem that finds the MVCS in an original [14, 13, 35] or a kernel-defined feature space [28, 33]:

$$\begin{aligned} \min \quad & -\text{trace}(\mathbf{M}) = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j \mathbf{x}_i' \mathbf{x}_j - \sum_{i=1}^{\ell} \alpha_i \mathbf{x}_i' \mathbf{x}_i \\ \text{s.t.} \quad & \\ & \sum_{i=1}^{\ell} \alpha_i = 1, \ 0 \leq \alpha_i \leq \frac{1}{\nu \ell}. \end{aligned} \tag{2.3}$$

In the following section and section 4 we propose a method of kernelising the optimization problem (2.2) and a decision rule (of whether or not a test point is an outlier) that are different from the method used to find the MVCS [28, 33].

Chapter 3

Working in high-dimensional feature spaces

One way of applying the ellipsoid algorithm in high-dimensional feature spaces is first to project the data into a low-dimensional subspace and then to apply the primal algorithm to the projected data. Perhaps the most natural way to do this is to use PCA. The kernel PCA algorithm makes it possible to perform such a projection in a kernel defined feature space by noting that the eigenvectors of the kernel matrix form a dual representation of the eigenvectors of the outer product matrix when suitably rescaled [11]:

$$\mathbf{u}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{X}'\mathbf{v}_i,$$

where $\mathbf{v}_i$ is the ith eigenvector of the kernel matrix with eigenvalue λ_i, $\mathbf{u}_i$ is the ith eigenvector of the outer product matrix and the rows of $\mathbf{X}$ are the feature vectors $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_\ell)$. Here, we assume a feature projection ϕ with corresponding kernel κ:

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \phi(\mathbf{x})'\phi(\mathbf{z}).$$

Here we consider applying the ellipsoid algorithm directly in the kernel-defined feature space. As it has been shown in [34] the key computation required in Titterington's algorithm is the computation of the Mahalanobis norm

$$\|\mathbf{x}\|_\alpha^2 = \phi(\mathbf{x})'\mathbf{M}^{-1}\phi(\mathbf{x}),$$

defined by the matrix

$$\mathbf{M} = \sum_{i=1}^{\ell} \alpha_i \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)' = \mathbf{X}'\mathbf{A}^2\mathbf{X},$$

where $\mathbf{A}$ is a diagonal matrix with $\mathbf{A}_{ii} = \sqrt{\alpha_i}$. This is the outer product matrix for the training set

$$S_\alpha = \{\sqrt{\alpha_i}\phi(\mathbf{x}_i) : i = 1, \ldots, \ell\}\,.$$

It follows that we can obtain a dual representation of its ith eigenvector $\mathbf{u}_i$ using kernel PCA of S_α:

$$\mathbf{u}_i = \frac{1}{\lambda_i}\mathbf{X}'\mathbf{A}\mathbf{v}_i,$$

where $\mathbf{v}_i$ is the eigenvector corresponding to the eigenvalue λ_i of the corresponding kernel matrix

$$\mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A} = \mathbf{A}\mathbf{K}\mathbf{A},$$

with $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. Note that the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are now dependent on α. We have suppressed this dependence to enhance readability, though we leave α as an index of the norm. Now consider the Mahalanobis norm computed using the matrix $\mathbf{M}$ for a point $\phi(\mathbf{x})$:

$$\|\mathbf{x}\|_\alpha^2 = \phi(\mathbf{x})'\mathbf{M}^{-1}\phi(\mathbf{x}) = \phi(\mathbf{x})'\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}'\phi(\mathbf{x}),$$

where $\mathbf{U}$ is the matrix with the eigenvectors $\mathbf{u}_i$ as columns and $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{\Lambda}_{ii} = \lambda_i$. Hence,

$$\begin{aligned}\|\mathbf{x}\|_\alpha^2 &= \sum_{i=1}^{k} \lambda_i^{-1}\left(\mathbf{u}_i'\phi(\mathbf{x})\right)^2 \\ &= \sum_{i=1}^{k} \lambda_i^{-2}\left(\mathbf{v}_i'\mathbf{A}\mathbf{k}\right)^2,\end{aligned}$$

where $\mathbf{k}_j = \kappa(\mathbf{x}_j, \mathbf{x})$, for $j = 1, \ldots, \ell$, and we assume truncation at some $k \leq \ell$.

This computation forms the basis of the kernel version of the ellipsoid algorithm used in the experiments described below.

Following the approach adopted in the analysis of [31] we can view $\|\mathbf{x}\|_\alpha^2$ as a linear function in the space defined by the kernel $\hat{\kappa}(\mathbf{x}, \mathbf{z}) = \kappa(\mathbf{x}, \mathbf{z})^2$ since

$$\begin{aligned}\|\mathbf{x}\|_\alpha &= \sum_{i=1}^{k} \lambda^{-2}\mathbf{u}_i\phi(\mathbf{x})\phi(\mathbf{x})'\mathbf{u}_i \\ &= \left\langle \sum_{i=1}^{k} \lambda^{-2}\mathbf{u}_i\mathbf{u}_i, \phi(\mathbf{x})\phi(\mathbf{x})' \right\rangle_F =: \left\langle \mathbf{w}, \phi(\mathbf{x})\phi(\mathbf{x})' \right\rangle_F,\end{aligned}$$

where $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product and

$$\hat{\kappa}(\mathbf{x}, \mathbf{z}) = \langle\phi(\mathbf{x})\phi(\mathbf{x})', \phi(\mathbf{z})\phi(\mathbf{z})'\rangle_F = \left(\phi(\mathbf{z})'\phi(\mathbf{x})\right)^2 = \kappa(\mathbf{x}, \mathbf{z})^2.$$

The norm of the weight vector $\mathbf{w}$ is given by

$$\begin{aligned}\|\mathbf{w}\|^2 &= \left\langle \sum_{i=1}^{k} \lambda_i^{-2}\mathbf{u}_i\mathbf{u}_i', \sum_{j=1}^{k} \lambda_j^{-2}\mathbf{u}_j\mathbf{u}_j' \right\rangle_F \\ &= \sum_{i=1}^{k} \lambda_i^{-4}\|\mathbf{u}_i\|^2 = \sum_{i=1}^{k} \lambda_i^{-4}.\end{aligned}$$

With k as the dimensionality of the space and ℓ as the number of data points, it is now easy to show the following identities:

$$\log\det\left(\mathbf{X}'\mathbf{A}^2\mathbf{X}\right) = \sum_{i:\lambda_i\neq 0} \log(\lambda_i) + (k - \#\{\lambda_i \neq 0\})\log(0) \tag{3.1}$$

$$\log\det\left(\mathbf{AKA}\right) = \sum_{i:\lambda_i\neq 0} \log(\lambda_i) + (\ell - \#\{\lambda_i \neq 0\})\log(0) \tag{3.2}$$

These equations suggest that we need to deal with the problem that a few eigenvalues λ_i are equal to zero. Therefore, we need to introduce some form of regularisation or to reduce the dimensionality of the space under considaration. In the case of the smallest enclosing sphere this problem does not arise because the MVCS uses a different objective function, $\Psi(\mathbf{M}) = -\text{trace}(\mathbf{M})$ (see (2.2) and (2.3)). But in both cases, the MVCE and the MVCS, the objective function depends on the centered information matrix only.

Note that, if we use trace as the objective function, then we obtain

$$\text{trace}\left(\mathbf{X}'\mathbf{A}^2\mathbf{X}\right) = \sum_{i:\lambda_i\neq 0} (\lambda_i) + (k - \#\{\lambda_i \neq 0\})\,(0) = \sum_{i:\lambda_i\neq 0} (\lambda_i), \tag{3.3}$$

$$\text{trace}\left(\mathbf{AKA}\right) = \sum_{i:\lambda_i\neq 0} (\lambda_i) + (\ell - \#\{\lambda_i \neq 0\})\,(0) = \sum_{i:\lambda_i\neq 0} (\lambda_i), \tag{3.4}$$

which implies that (3.5)

$$\Rightarrow \text{trace}\left(\mathbf{AKA}\right) = \text{trace}\left(\mathbf{X}'\mathbf{A}^2\mathbf{X}\right) = \text{trace}\left(\mathbf{KA}^2\right) = \sum_{i=1}^{\ell} \alpha_i k(\mathbf{x}_i, \mathbf{x}_i). \tag{3.6}$$

Essentially, for appropriate choice of the objective function $\Psi(\mathbf{M})$, the kernelisation method proposed in this paper contains the MVCS method [13, 33] as a special case. Our method can also be used for the criterion that is used in A-optimal experimental design, namely $\text{trace}(\mathbf{M})^{-1}$.

Recall that the MVCE for the dataset $\{\mathbf{x}_i\}_{i=1}^{\ell}$ must go through at least $k+1$ and at most $\frac{1}{2}k(k+3)+1$ support points, for which $\alpha_i > 0$ [35, 34, 19]. Therefore, in the case of the MVCE method the choice of k can be also dictated by the following two extreme cases: 1) $k+1 \leq \ell$; and 2) $\left\lfloor \frac{1}{2}k(k+3)+1 \right\rfloor \leq \ell$. We will also discuss the choice of k in the applications to artificial and real life datasets in Section 5.

Chapter 4

Some implementation details

4.1 Some implementation details

The optimization algorithm for the KIMEE method without upper bound for Lagrangian multipliers α ($\nu = 0$) can be derived using the Titterington's simple but effective algorithm for the minimum volume ellipsoid [34]. In this case, in order to update α_j and to find supports points $\mathbf{x}_{SVs}$ it is necessary to evaluate the only Mahalanobis norm $\|\mathbf{x}_j\|_\alpha^2$ of each training point $\mathbf{x}_j$. After a few iterations the algorithm converges. Note that the Mahalanobis norm $\|\mathbf{x}_j\|_\alpha^2$ of the training points $\mathbf{x}_j$ is a gradient of the objective function $\log\det\mathbf{M}$ (see (2.2)): $\frac{\partial \log\det\mathbf{M}}{\partial\alpha_j} = (\phi(\mathbf{x}_t) - \phi(\mathbf{c}))'\mathbf{M}^{-1}(\phi(\mathbf{x}_t) - \phi(\mathbf{c})) = \|\mathbf{x}_t\|_\alpha^2$, where $\mathbf{M} = \left\{\sum_{i=1}^{\ell}\alpha_i\phi(\mathbf{x}_i)\phi(\mathbf{x}_i') - \phi(\mathbf{c})\phi(\mathbf{c})'\right\}$ and $\phi(\mathbf{c}) = \sum_{i=1}^{\ell}\alpha_i\phi(\mathbf{x}_i)$. We propose the following algorithm in which $\mathbf{1}_\ell$ denotes the column vector of ℓ ones.

If optimization functions require the calculation of the objective function $\log\det\mathbf{M}$ then $\log\det\mathbf{M} = \sum_{i=1}^{k}\log\lambda_i$ where λ_i are the eigenvalues of the matrix $\mathbf{A}\hat{\mathbf{K}}\mathbf{A}$; see Algorithm 1.

After obtaining the values α we can use the following rules to check if the test point $\mathbf{x}_t$ belongs to the estimated support

$$(\phi(\mathbf{x}_t) - \phi(\mathbf{c}))'\mathbf{M}^{-1}(\phi(\mathbf{x}_t) - \phi(\mathbf{c})) = \|\mathbf{x}_t\|_\alpha^2 < R^2 \qquad \text{indicates a target point;} \quad (4.1)$$

$$(\phi(\mathbf{x}_t) - \phi(\mathbf{c}))'\mathbf{M}^{-1}(\phi(\mathbf{x}_t) - \phi(\mathbf{c})) = \|\mathbf{x}_t\|_\alpha^2 = R^2 \qquad \text{indicates a boundary point;} \quad (4.2)$$

$$(\phi(\mathbf{x}_t) - \phi(\mathbf{c}))'\mathbf{M}^{-1}(\phi(\mathbf{x}_t) - \phi(\mathbf{c})) = \|\mathbf{x}_t\|_\alpha^2 > R^2 \qquad \text{indicates an outlier.} \quad (4.3)$$

Here R^2 is a squared distance, normalized by the matrix $\mathbf{M}$, from the center of the ellipsoid in feature space to one of the support vectors $\mathbf{x}_{bsv}$ that lies on its boundary [35, 34]:

$$R^2 = (\phi(\mathbf{x}_{bsv}) - \phi(\mathbf{c}))'\mathbf{M}^{-1}(\phi(\mathbf{x}_{bsv}) - \phi(\mathbf{c})) = \|\mathbf{x}_{bsv}\|_\alpha^2. \quad (4.4)$$

Algorithm 1 Kernel Information Matrix Ellipsoid Estimation (KIMEE) Algorithm

Initialization: Define the kernel matrix $\mathbf{K}$, the number of iterations r_{max}, the threshold t and α. For example, in the condition monitoring experiment (see Section 5.2) $\mathbf{K}$ was a Gaussian kernel with $\rho = 320$, $r_{max} = 150$, $t = 0.0001$, and $\alpha_j = 1/\ell$, for $j = 1, \ldots, \ell$).

for $r = 1, \ldots, r_{max}$ **do**

1. $b = \sum_i^\ell \sum_j^\ell \mathbf{B}_{ij}$, where $\mathbf{B} = \hat{\mathbf{A}}\mathbf{K}\hat{\mathbf{A}}$ and $\hat{\mathbf{A}} = diag(\alpha_1, ..., \alpha_\ell)$.
2. Find the 'centered' kernel $\hat{\mathbf{K}}$
   $\hat{\mathbf{K}} = \mathbf{K} - \mathbf{1}_\ell \alpha' \mathbf{K} - \mathbf{K}' \alpha \mathbf{1}'_\ell + b\mathbf{1}_\ell \mathbf{1}'_\ell$.
3. $\mathbf{A} = diag(\sqrt{\alpha_1}, ..., \sqrt{\alpha_\ell})$.
4. Obtain eigenvectors $\mathbf{v}_i \in \mathbf{V}$ and eigenvalues $\lambda_i \in \lambda$ of the matrix $\mathbf{A}\hat{\mathbf{K}}\mathbf{A}$:
   $\mathbf{V}\mathbf{D}\mathbf{V}' = \mathbf{A}\hat{\mathbf{K}}\mathbf{A}$, where $\mathbf{V}$ is an $\ell \times \ell$ matrix and $\mathbf{D} = \text{diag}(\lambda)$.
5. Sort λ in decreasing order and the correspondingly permute the columns of the matrix $\mathbf{V}$.

if r=1 **then**
Set k equal to the number of eigenvalues that are greater or equal to t.
end if

6. $\alpha^{old} \leftarrow \alpha$.

for $j = 1, \ldots, \ell$ **do**

7. Calculate the Mahalanobis norm $\|\mathbf{x}_j\|^2_\alpha$ for the training point $\mathbf{x}_j$:
   $\|\mathbf{x}_j\|^2_\alpha = \sum_{i=1}^k \lambda_i^{-2} \left(\mathbf{v}'_i \mathbf{A}\hat{\mathbf{k}}^j\right)^2$, where $\hat{\mathbf{k}}^j_s = \hat{\kappa}(\mathbf{x}_s, \mathbf{x}_j)$, for $s = 1, \ldots, \ell$.
8. Obtain new values for α_j using the Titterington's algorithm [34, 19]: $\alpha_j = \alpha_j^{old} \|\mathbf{x}_j\|^2_\alpha / k$.

end for
end for
return α

There are many other choices for R^2, for example, there could be the following three methods: (a) $R^2 = k$ [34]; (b) R^2 based on $\chi^2(k)$ [24, 22] or on extreme value statistics [21]; (c) R^2 obtained by cross validation.

If the point of the training dataset is inside the ellipsoid then the corresponding Lagrangian multiplier α_j is 0 but for a support point on the boundary (outside the ellipsoid) we have $0 < \alpha_j < \frac{1}{\nu\ell}$ ($\alpha_j = \frac{1}{\nu\ell}$).

Chapter 5

Experiments

In this section the performance of the KIMEE method is analyzed on simulated and real-life datasets. Two datasets contain outliers in the training sample.

5.0.1 Simulated dataset

In order to illustrate the performance of the proposed KIMEE for novelty/outlier detection we use an artificial dataset that is similar to that previously reported by C. Campbell and K.P. Bennett for novelty detection based on the linear programming approach [4]. In our experiment we generate a sample from the Gaussian distribution with mean $(10, 5)'$ and covariance matrix $\mathbf{M} = (0.0163, -0.0062; -0.0062, 0.0163)$. There are 4 outliers (see Fig.5.1). We use the simple inner product kernel $\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y}$ with $t = 0.001$ (which gives $k = 2$). We first find the ellipse that encloses all points. Then we remove from the training sample points on the boundary of that ellipse. We then find the ellipse that does not include the outliers. This approach of removing outliers is extensively used in statistics [32, 34].Clearly our method successfully removed the four outliers. We used this approach to remove outliers from the training dataset for the condition monitoring example that we describe below.

The novelty detection approach based on slack variables [33, 37, 28, 30] is sensitive to outliers [32]. In some cases it can be useful to re-train the model as described above even when using the MVCE with slack variables [32, 10]. Fig.5.2 demonstrates this idea. Note that, in practice, the novelty detection methods [33] based on a Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\rho^2)$, can be more robust than, for example, a polynomial kernel, because $\|\mathbf{x}\|^2 = 1$ for the Gaussian kernel. Therefore, the values of slack variables are implicitly bounded above.

An application of our approach is also demonstrated using data with a nonconvex support (see Fig.5.3) and a Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\rho^2)$. The sparsity

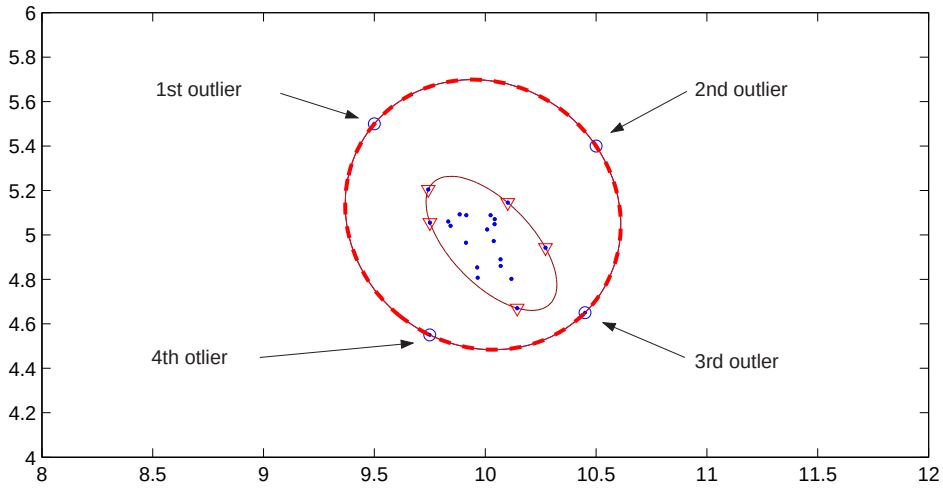

FIGURE 5.1: In this example 4 outliers are ignored by the KIMEE method after retraining. Both the larger and smaller ellipses were found by Algorithm 1 using the simple inner product kernel $\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y}$ and $t = 0.001$. Support points on the boundaries of these two ellipsoids are denoted by *triangles* and *circles*, respectively.

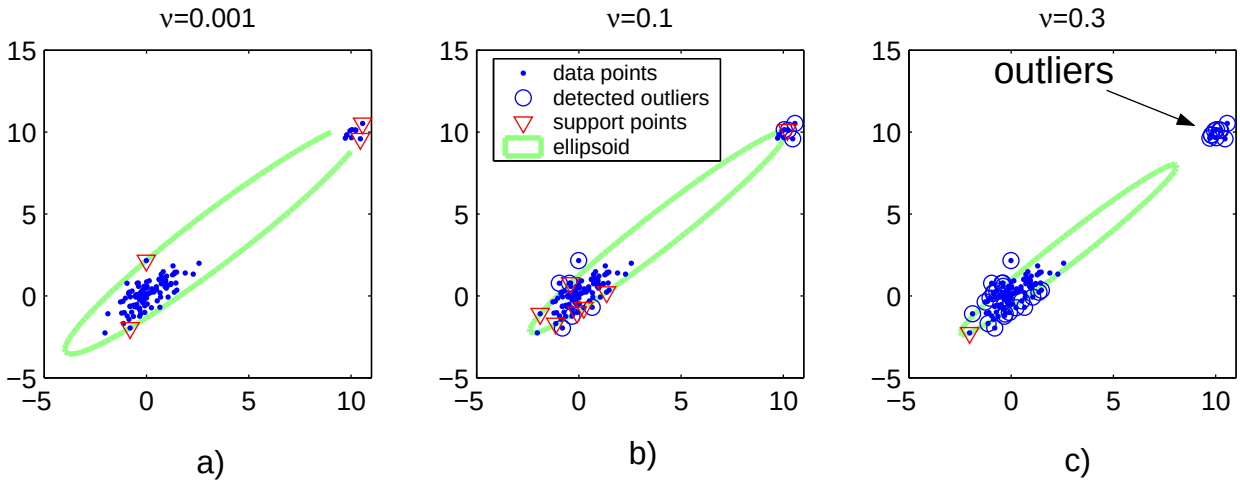

FIGURE 5.2: In this example the sample of 100 points from a Gaussian distribution was contaminated by 10 outliers that belong to a Gaussian distribution with a different mean and covariance matrix. By the tuning parameter ν we can exclude the outliers but the ellipse is still not centered around the original data.

of the solution (i.e. the number of support points) of the MVCE method depends on the value of the smoothing parameter ρ. This has a very simply explanation. By increasing the value ρ we decrease the number of eigenvalues that are above t (see (3.1), Fig. 5.3,5.4 (a) and Algorithm 1). In other words, by increasing the value ρ we decrease the dimensionality k of the space of the data. However as mentioned before the number of support points depends on the dimensionality of the space: the MVCE for the dataset $\{\mathbf{x}_i\}_{i=1}^{\ell}$ must go through at least $k+1$ and at most $\frac{1}{2}k(k+3)+1$ support points. It can be seen that by tuning the parameter ρ we can find the more tight data support (see Fig. 5.3).

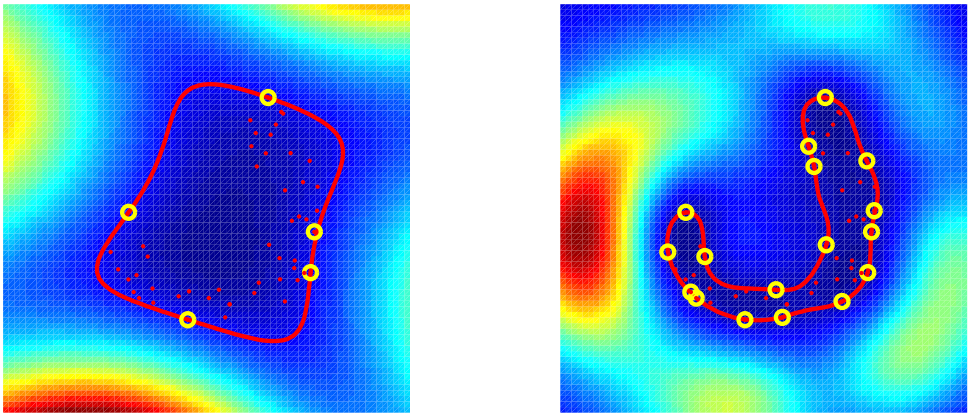

FIGURE 5.3: Illustration of the banana dataset with $\ell = 50$. In this example we use Algorithm 1 with the threshold $t = 0.001$ and a Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\rho^2)$, with (a) $\rho = 21$, (b) $\rho = 7$.

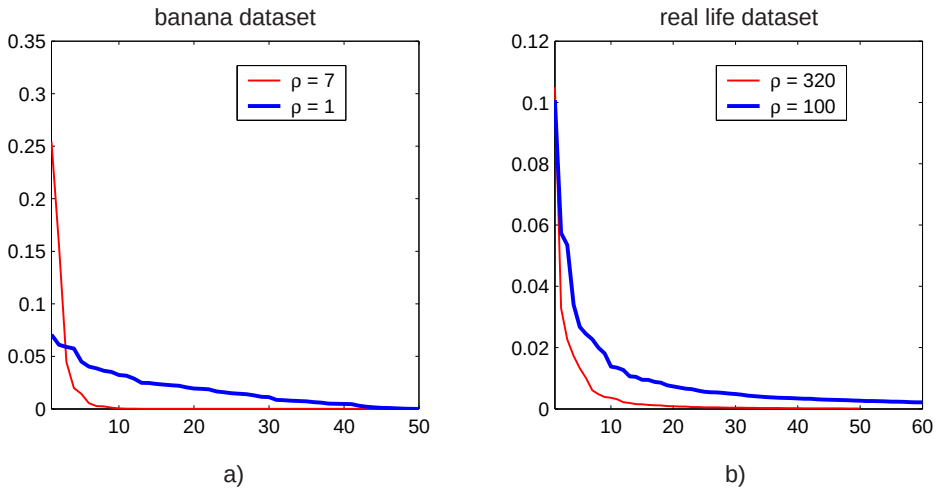

FIGURE 5.4: Illustration of dimensionality reduction and feature selection using a Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\rho^2)$, for (a) the banana and (b) condition monitoring datasets.

5.0.2 Condition monitoring

We analyse the comparative performance of one-class SVM using LIBSVM implementation [5], the proposed KIMEE method (see Algorithm 1) using Titterington's algorithm [34] and the LPND algorithm [4] on a real-life dataset from the Structural Integrity and Damage Assessment Network [39, 4]. The LPND method used on this dataset was reported in [4]. There are vibration measurements in this dataset that correspond to "healthy" measurements (without fault) and 4 types of malfunction of machinery (see Fig.5.5): 1) Fault 1 (the bearing had an outer race completely broken); 2) Fault 2 (broken cage with one loose element); 3) Fault 3 (broken cage with four loose elements); 4) Fault 4 (a badly worn ball-bearing with no apparent damage). Seven hundred data samples of time series are shown in Fig.5.5 for each of these five situations together with an example of extracted features in Fig.5.6

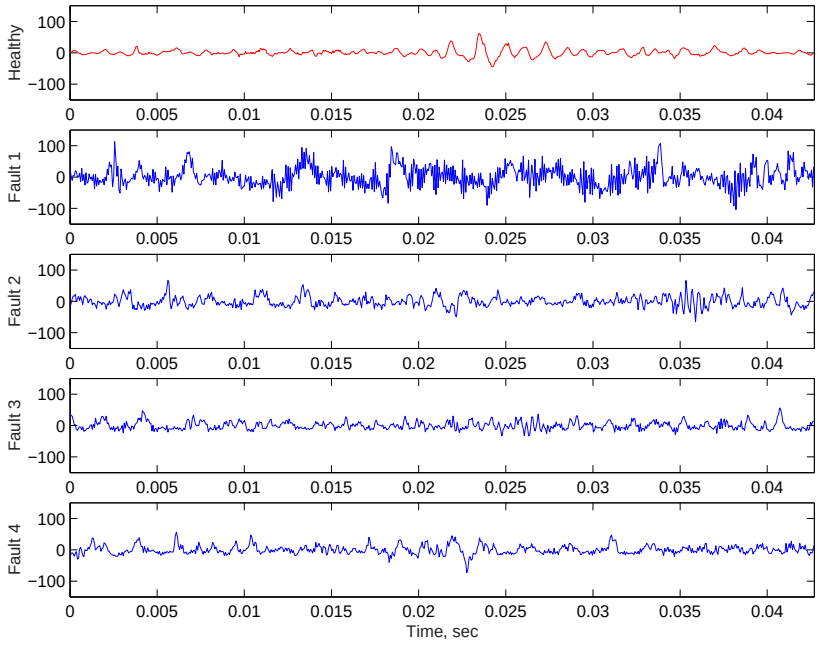

FIGURE 5.5: Condition monitoring experiment. In this example measurements, one channel, are obtained from a healthy machine and when four different faults are present [4, 39].

In order to compare our KIMEE method with the LPND method and the one-class SVM method, we performed experiments in the same way as described in [4], using the same Gaussian kernel, training set, validation set and test set. In the experiments we use the same kernel matrix $\mathbf{K}$ (Gaussian kernel with standard deviation equal to 320) for LPND, the one-class SVM and the KIMEE method. After training, the KIMEE algorithm showed really poor performance for Fault 1 (see Fig.5.5,5.6). The method labels almost 100% of the points from this class as 'Healthy' (no fault). It suggests that the training sample contains outliers (for example, there are large vibrations between 0.2s and 0.3s, see Fig,5.5).

In order to remove these outliers we carried out the same steps as with the artificial dataset in Section 5.1 (Fig.5.1): 1) we removed the points on the boundary; 2) we re-trained our novelty detector based on KIMEE using the remaining 793 points, 60 of the 793 eigenvalues of the kernel matrix are shown in Fig.5.4(b); 3) we scaled a newly obtained ellipse using a different R^2 in order to achieve desirable errors of the first and second kinds. It can be seen that for approximately the same correct classification of the 'Healthy' class ($R^2 = 140$) our method performs better than the soft margin one-class SVM method, and when $R^2 = 180$ or $R^2 = 190$ our KIMEE method is significantly better than the LPND method [4]. Note that one-class SVM using a Gaussian kernel is equivalent to finding the hypersphere around the data points.

In robust statistics and signal processing there are two approaches to remove outliers: 1) apply a robust method that uses the outliers to obtain the estimate of the parameter; and 2) detect the outliers by another method, exclude them from the training dataset

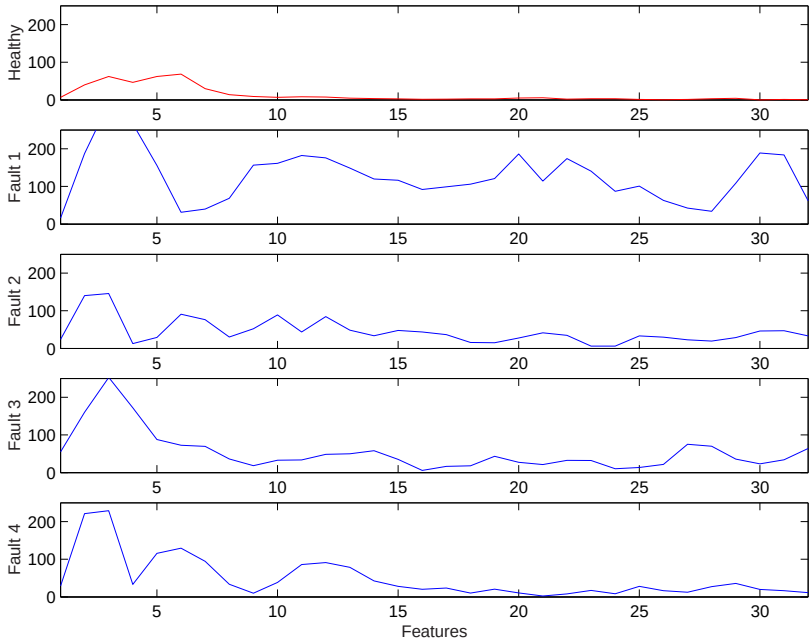

FIGURE 5.6: Illustration of the features used in the condition monitoring experiment [4, 39].

TABLE 5.1: The percentage of correctly labelled classes using LPND, one-class SVM and KIMEE methods

Method	**Healthy**	**Fault 1**	**Fault 2**	**Fault 3**	**Fault 4**
LPND	98.7%	100%	53.3%	28.3%	25.5%
one-class SVM	97.9%	100%	84.3%	57.3%	61.1%
KIMEE, $R^2 = 190$	99.8%	100%	79.2%	50.5%	52.4%
KIMEE, $R^2 = 180$	99.6%	100%	82.9%	54.4%	57.7%
KIMEE, $R^2 = 140$	97.7%	100%	93.9%	72.8%	76.8%
KIMEE, $R^2 = 64$	79%	100%	100%	97.5%	98.8%

and estimate the parameter using this reduced dataset and a non-robust estimator. In this case soft margin algorithms belong to the first approach but ellipsoid trimming and hard margin methods belong to the second approach. Both approaches are valid to some extent.

All three methods (LPND, 1-class SVM and our method) were trained using, cross validation in such a way that they about 2% of the target data are rejected rate (see the column "Healthy", Table 1) and such that all cases of "Fault 1" are detected. Thus all three methods behave equally in this sense while using different approaches to deal with outliers. We do not use any manual or visual inspection to remove outliers before applying our method. To deal with outliers in our simulations we use the soft margin LPND method [4] and the soft margin 1-class SVM algorithm [28].

Similarly to the minimum volume covering sphere R^2 is equal to the distance (in our case Mahalanobis distance) to the support point $\mathbf{x}_i$ on the boundary of the ellipsoid, that is, corresponding to $\alpha_i > 0$ [28, 33].

It has been reported by many authors a Gaussian kernel is a good choice for novelty detection but the choice of the kernel could depend on the problem. For example, if it is known that the target data cloud is are ellipsoidal then we can use the simple inner product kernel (see Fig. 1). In a high-dimensional feature space, the algorithm benefits significantly if the distribution of the data is nonconvex or multimodal.

Chapter 6

Conclusions

We have proposed the new Kernel Information Matrix Ellipsoid Estimation method and have shown how it can be applied for the outlier detection. The proposed algorithm is a very general method. For example, if we use a different form of regularisation the proposed method can be further developed for the D-optimum experimental design for the kernel ridge regression model. We have demonstrated that it is not always necessary to specify an accurate estimate of the proportion of outliers in advance. We have proposed a very simple but effective iterative algorithm, based on Titterington's minimum volume ellipsoid method, that can be used both for novelty detection and experimental design problems. The KIMEE method has demonstrated a better or similar performance on the real-life condition monitoring problem compared to the one-class SVM model and the LPND method. We have that for the appropriately chosen objective function the method of kernelisation proposed in this paper includes the MVCS as a special case and therefore it is related to the one-class SVM model as well. In future work it would be interesting to see how much we can increase robustness of the proposed method if we adapt the methodology that is successfully used by Rousseeuw's minimum covariance determinant estimator [25].

Acknowledgments

This research is partially supported by the Data Information Fusion Defence Technology Center, United Kingdom, under DTC Projects 8.1: "Active multi-sensor management" and the PASCAL network of excellence. We would like to thank Tijl De Bie for fruitful discussions.

Bibliography

[1] Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, (2nd ed.) New York: Wiley (1984)

[2] Barnett, V., Lewis, T. *Outliers in Statistical Data*. 3rd ed, Chichester : Wiley (1994)

[3] Bishop, C. Novelty Detection and Neural Network Validation, Proceedings, IEE Conference on Vision and Image Signal Processing (1994) 217–222

[4] Campbell, C., Bennett, K.P. A Linear Programming Approach to Novelty Detection, Advances in Neural Information Processing Systems **14** MIT Press, Cambridge, MA (2001)

[5] Chang, C.-C. and C.-J. Lin. LIBSVM: a Library for Support Vector Machines (2001). Software available at http://www.csie.ntu.edu.tw/ ˜cjlin/libsvm

[6] Croux, C. and Haesbroeck, G. Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator, Preprint, University of Brussels (1998)

[7] Croux, C., and Haesbroeck, G. Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. Biometrika **8**(3) (2000) 603–618

[8] Croux, C. and Haesbroeck, G., Rousseeuw, P. J. Location Adjustment for the Minimum Volume Ellipsoid Estimator. Statistical Computation **12**(3) (2002) 191–200

[9] Devroye, L., Wise, G.L. Detection of Abnormal Behaviour via Nonparametric Estimation the Support, SIAM Journal on Applied Mathematics **38**(3) (1980) 480–488

[10] Dolia, A.N., Page, S.F., White, N.M., Harris, C.J. D-optimality for Minimum Volume Ellipsoid with Outliers, Proceedings of the Seventh International Conference on Signal/Image Processing and Pattern Recognition, (UkrOBRAZ'2004), Kiev, Ukraine, October 11-15 (2004) 73–76

[11] Dolia, A.N., Page, S.F., White, N.M., Harris, C.J. ν-MCD Approach to Novelty Detection, Machine Learning, Support Vector Machines, and Large Scale Optimization Workshop, Wissenschaftszentrum Schlob Thurnau, Germany 16 - 18 March (2005)

[12] Duda, R.O., Hart, P.E., Stork, D.G. *Pattern classification*, (2nd ed.) New York : Wiley (2001)

[13] Elzinga, D.J. and Hearn, D.W. The Minimum Covering Sphere Problem, Management Science **19**(1) (1972) 96–104

[14] Elfving, G. Optimum Allocation in Linear Regression Theory, The Annals of Mathematical Statistics **23**(2) (1952) 255–262

[15] Jobson, J.D. *Applied Multivariate Data Analysis*, Vol.1 Regression and Experimental Design, Vol.2 Categorical and Multivariate Methods, New York : Springer (1991-1992)

[16] Hong, X., Brown, M., Chen, S., Harris, C. J. Sparse Model Identification Using Orthogonal Forward Regression with Basis Pursuit and D-optimality, IEE Proceedings - Control Theory and Applications **151**(4) (2004) 491–498

[17] Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., Tarassenko L. A System for the Analysis of Jet Engine Vibration Data, Integrated Computer Aided Engineering **6** (1999) 53–65

[18] Olive, D.J. Applications of Robust Distances, Technometrics **44**(1) (2002) 1009–1052

[19] Pronzato, L. Acceleration of D-Optimum Design Algorithms by Removing Non-Optimal Support Points. Technical Report I3S/RR-2002-05- FR, Laboratoire I3S Informatiques Signaux et Systèmes de Sophia Antipolis (2002)

[20] Pronzato, L., Wynn, H., Zhigljavsky, A. Kantarovich-type Inequalities for Operators via D-optimal Design Theory, Linear Algebra and Its Applications **410** (2005) 160–169

[21] Roberts, S.J. Novelty Detection Using Extreme Value Statistics. IEE Proceedings on Vision, Image and Signal Processing **146**(3) (1999) 124–129

[22] Roth, V. Outlier Detection with One-class Kernel Fisher Discriminants. Advances in Neural Information Processing Systems 17, MIT Press (2004)

[23] Rousseeuw, P.J. and Leroy, A.M. *Robust Regression and Outlier Detection*. New York: Wiley-Interscience (1987)

[24] Rousseeuw, P.J. Least Median of Squares Regression, Journal of the American Statistical Association, **79** (1984) 871–881

[25] Rousseeuw, P.J. and Van Driessen, K. A Fast Algorithm for the Minimum Covariance Determinant Estimator, Technometrics **41**(3) (1999) 212–223

[26] Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J.S., Platt, J. Support Vector Method for Novelty Detection, In: Solla, S.A., Leen, T.K., Muller, K.R. (eds.) Neural Information Processing Systems (2000) 582–588

[27] Schölkopf, B., Burges, C., Vapnik, V. Extracting Support Data for a Given Task, Fayyad, U.M., Uthurusamy, R. (eds.) Proceedings, First International Conference on Knowledge Discovery & Data Mining. AAAI Press, Menlo Park (1995) 252–257

[28] Schölkopf, B., Smola, A. *Learning with Kernels*. MIT Press, Cambridge, MA (2001)

[29] Silverman, B. W., and Titterington, D. M. Minimum Covering Ellipses, SIAM Journal on Scientific and Statistical Computing **1** (1980) 401-409

[30] Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK (2004)

[31] Shawe-Taylor, J. and Williams, C. and Cristianini, N. and Kandola, J. S. On the Eigenspectrum of the Gram Matrix and Its Relationship to the Operator Eigenspectrum. Proceedings of the 13th International Conference on Algorithmic Learning Theory (ALT2002) **2533** (2002) 23–40

[32] Sun, P. and Freund, R.M. Computation of Minimum-Volume Covering Ellipsoids, Operations Research **52**(5) (2004) 690–706

[33] Tax, D.M.J., Duin, R.P.W. Data Domain Description by Support vectors. Verleysen, M. (ed.). Proceedings, ESANN. Brussels (1999) 251–256

[34] Titterington, D.M. Estimation of Correlation Coefficients by Ellipsoidal Trimming, Journal of Royal Statistical Society **C27**(3) (1978) 227–234

[35] Titterington, D.M. Optimal Design: Some Geometrical Aspect of D-optimality, Biometrika **62**(2) (1975) 313–320

[36] Vandenberghe, L., Boyd, S., and Wu, S.-P. Determinant Maximization with Linear Matrix Inequality Constraints, SIAM Journal on Matrix Analysis and Applications **19**(2):499-533, 1998

[37] Vapnik, V. *Statistical Learning Theory*. Wiley NY (1998)

[38] Ypma, A., Duin, R.P.W. Novelty Detection Using Self-organising Maps.: Progress in Connectionist Based Information Systems **2** (1998) 1322–1325

[39] http://www.sidanet.org