

# The Minimum Volume Covering Ellipsoid Estimation in Kernel-Defined Feature Spaces

Alexander N. Dolia<sup>1</sup>, Tijl De Bie<sup>2</sup>, Chris J. Harris<sup>1</sup>,  
John Shawe-Taylor<sup>1</sup> and D.M. Titterton<sup>3</sup>

<sup>1</sup> University of Southampton, Southampton, SO17 1BJ, UK

<sup>2</sup> OKP Research group, Katholieke Universiteit Leuven, Belgium

<sup>3</sup> Department of Statistics, University of Glasgow

**Abstract.** Minimum volume covering ellipsoid estimation is important in areas such as systems identification, control, video tracking, sensor management, and novelty detection. It is well known that finding the minimum volume covering ellipsoid (MVCE) reduces to a convex optimisation problem. We propose a regularised version of the MVCE problem, and derive its dual formulation. This makes it possible to apply the MVCE problem in kernel-defined feature spaces. The solution is generally sparse, in the sense that the solution depends on a limited set of points. We argue that the MVCE is a valuable alternative to the minimum volume enclosing hypersphere for novelty detection. It is clearly a less conservative method. Besides this, we can show using statistical learning theory that the probability of a typical point being misidentified as a novelty is generally small. We illustrate our results on real data.

## 1 Introduction

The minimum volume covering ellipsoid (MVCE) [2, 3, 10, 12], the ellipsoid smallest in volume that covers all of a given set of points, has many applications in areas ranging from systems and control to robust statistics. In this paper we focus on its application to novelty detection (also known as support estimation or domain description): then, all data points from a training set  $\{\mathbf{x}_i\}_{i=1}^{\ell}$  are specified to be sampled from an unknown distribution  $\mathcal{D}$ , and the support of  $\mathcal{D}$  is estimated as the inside region of the MVCE. Points lying outside of the ellipsoid can then subsequently be judged to be *novelties*.

Recently, several results in the machine learning domain have attacked this problem by means of the minimum volume covering hypersphere (MVCS) [7, 8, 11], fitting a tight hypersphere around the data. A hypersphere being a special type of ellipsoid, the volume of the MVCE will never be larger than the volume of the MVCS. The motivation for the current work is that the additional flexibility in using an ellipsoid is likely to be more sensitive in identifying novelties.

However, specificity problems should be expected for high-dimensional spaces. Indeed, the MVCE becomes vanishingly small for data sets smaller in size than their dimension, and the method would reject (nearly) all test points from  $\mathcal{D}$  as outliers, judged not to belong to the support of the distribution. To overcome

this problem, we propose a regularised MVCE (RMVCE) method. This allows us to derive the main result of this paper, which is the RMVCE problem in a (possibly infinite-dimensionally) kernel-defined feature space.

Additionally, we present an in depth statistical analysis of the novelty detection method that is based on the RMVCE problem, and an extension of the RMVCE problem and its kernel version towards a soft-margin variant.

## 2 The Minimum Volume Ellipsoid

Assume that we have a training dataset containing  $\ell$  samples,  $\{\mathbf{x}_i \in \mathfrak{R}^{k \times 1}\}_{i=1}^{\ell}$ . The MVCE is specified by the positive definite matrix  $\mathbf{M} \in \mathfrak{R}^{k \times k}$  that solves the optimization problem (for conciseness, in this paper we assume the ellipsoid is centred at the origin—extending to a variable centre is trivial [12]):

$$\begin{aligned} \min_{\mathbf{M}, \mu} \quad & \log \det \mathbf{M} + \mu, \\ \text{s.t.} \quad & \mathbf{x}_i' \mathbf{M}^{-1} \mathbf{x}_i \leq \mu, \text{ for all } i. \end{aligned} \tag{1}$$

The objective consists of two terms: the logarithm of the volume of the ellipsoid, and the maximal *Mahalanobis distance*  $\mathbf{x}_i' \mathbf{M}^{-1} \mathbf{x}_i$  over all data points  $\mathbf{x}_i$ . This objective as well as the constraints which constrain the data points to be within a Mahalanobis distance  $\mu$  from the centre of the ellipsoid are both convex in  $\mathbf{M}^{-1}$  and  $\mu$ . Therefore, the optimization problem has a unique optimal solution.

The dual of optimisation problem (1) can be written as [12]:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \mathbf{M}} \quad & \log \det (\mathbf{M}), \\ \text{s.t.} \quad & \boldsymbol{\alpha} \geq \mathbf{0}, \quad \boldsymbol{\alpha}' \mathbf{e} = 1, \\ & \mathbf{M} = \sum_{i=1}^{\ell} \alpha_i \mathbf{x}_i \mathbf{x}_i'. \end{aligned} \tag{2}$$

from which the variable  $\mathbf{M}$  can directly be eliminated to yield an optimisation problem in  $\boldsymbol{\alpha}$  only. In the following section we propose a regularised version of the MVCE problem.

## 3 Regularised minimum volume covering ellipsoid

As explained in the introduction, we should prevent the ellipsoid to collapse to zero volume in large dimensional spaces. This can be achieved by changing the constraint  $\mathbf{M} = \sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i'$  in (2) into  $\mathbf{M} = \sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i' + \gamma \mathbf{I}$ , which guarantees a minimal diameter of the ellipsoid in all directions. This gives:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \mathbf{M}} \quad & \log \det (\mathbf{M}), \\ \text{s.t.} \quad & \boldsymbol{\alpha} \geq \mathbf{0}, \quad \boldsymbol{\alpha}' \mathbf{e} = 1, \\ & \mathbf{M} = \sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i' + \gamma \mathbf{I}. \end{aligned} \tag{3}$$

The dual of this optimisation problem is given by (without derivation due to space restrictions):

$$\begin{aligned} \min_{\mathbf{M}, \mu} \quad & \log \det(\mathbf{M}) + \mu + \gamma \text{trace}(\mathbf{M}^{-1}), \\ \text{s.t.} \quad & \mathbf{x}'_i \mathbf{M}^{-1} \mathbf{x}_i \leq \mu, \text{ for all } i. \end{aligned} \quad (4)$$

For  $\gamma = 0$ , this is equal to the standard MVCE centred at the origin formulation as discussed in the previous section. Different from the standard formulation is the additional *regularization term*  $\gamma \text{trace}(\mathbf{M}^{-1})$ . This term ensures that the ellipsoids axes are never extremely small. Indeed, a small diameter in one dimension would result in a small eigenvalue of  $\mathbf{M}$ , which in turn leads to a large trace of  $\mathbf{M}^{-1}$ . As we can learn from  $\mathbf{M} = \sum_i \alpha_i \mathbf{x}_i \mathbf{x}'_i + \gamma \mathbf{I}$  in (3), the effect is that the diameter along each of the dimensions is at least equal to  $\gamma$ .

*Soft margin RMVCE formulation.* In the presence of outliers it can be appropriate to introduce slack variables  $\xi_i$  and add a corresponding penalty term to the objective:

$$\begin{aligned} \min_{\mathbf{M}, \mu} \quad & \log \det(\mathbf{M}) + \mu + \gamma \text{trace}(\mathbf{M}^{-1}) + \frac{1}{\nu \ell} \sum_{i=1}^{\ell} \xi_i, \\ \text{s.t.} \quad & \mathbf{x}'_i \mathbf{M}^{-1} \mathbf{x}_i \leq \mu + \xi_i, \text{ for all } i, \quad \boldsymbol{\xi} \geq 0. \end{aligned} \quad (5)$$

where  $\nu \in (0, 1]$ . The dual problem can be written as follows:

$$\begin{aligned} \boldsymbol{\alpha}_\gamma^* = \text{argmin}_{\boldsymbol{\alpha}} \quad & -\log \det \left( \sum_i \alpha_i \mathbf{x}_i \mathbf{x}'_i + \gamma \mathbf{I} \right), \\ \text{s.t.} \quad & \mathbf{e} \geq \nu \ell \boldsymbol{\alpha} \geq 0, \quad \boldsymbol{\alpha}' \mathbf{e} = 1. \end{aligned}$$

## 4 Kernel regularised minimum volume covering ellipsoid

Let us first define the diagonal matrix  $\mathbf{A}$ , with  $\mathbf{A}_{ii} = a_i = \sqrt{\alpha_i} \geq 0$ , such that (with  $\mathbf{a} = (a_1 \ a_2 \ \dots \ a_\ell)'$ ) from  $\mathbf{e}' \boldsymbol{\alpha} = 1$  we have that  $\mathbf{a}' \mathbf{a} = 1$ . Then we can write  $\sum_i \alpha_i \mathbf{x}_i \mathbf{x}'_i + \gamma \mathbf{I} = \mathbf{X}' \mathbf{A}^2 \mathbf{X} + \gamma \mathbf{I}$ . Note that the matrices  $(\mathbf{A}\mathbf{X})'(\mathbf{A}\mathbf{X}) = \mathbf{X}' \mathbf{A}^2 \mathbf{X}$  and  $(\mathbf{A}\mathbf{X})(\mathbf{A}\mathbf{X})' = \mathbf{A}\mathbf{X}\mathbf{X}'\mathbf{A} = \mathbf{A}\mathbf{K}\mathbf{A}$  have the same nonzero eigenvalues  $\lambda_i$ , equal to the squares of the singular values of  $\mathbf{A}\mathbf{X}$  [2]. With  $d$  the dimensionality of the space and  $\ell$  the number of data points  $\mathbf{x}_i$ , it is now easy to show that:

$$\log \det (\mathbf{A}\mathbf{K}\mathbf{A} + \gamma \mathbf{I}) = \log \det (\mathbf{X}\mathbf{A}^2 \mathbf{X} + \gamma \mathbf{I}) + (\ell - d) \log(\gamma).$$

Hence we can optimize  $\log \det (\mathbf{A}\mathbf{K}\mathbf{A} + \gamma \mathbf{I})$  instead of  $\log \det (\mathbf{X}\mathbf{A}^2 \mathbf{X} + \gamma \mathbf{I})$ . Now define  $\mathbf{C}$  to be a Cholesky factor of  $\mathbf{K}$  (i.e.  $\mathbf{K} = \mathbf{C}\mathbf{C}'$ ). Then,  $\mathbf{A}\mathbf{K}\mathbf{A} = \mathbf{A}\mathbf{C}\mathbf{C}'\mathbf{A}$  and  $\mathbf{C}'\mathbf{A}^2\mathbf{C} = \sum_{i=1}^{\ell} \alpha_i \mathbf{c}_i \mathbf{c}'_i$  with  $\mathbf{c}_i$  the  $i$ th row of  $\mathbf{C}$  have the same eigenvalues, such that  $\log \det (\mathbf{A}\mathbf{K}\mathbf{A} + \gamma \mathbf{I}) = \log \det \left( \sum_{i=1}^{\ell} \alpha_i \mathbf{c}_i \mathbf{c}'_i + \gamma \mathbf{I} \right)$ . Hence, we obtain the kernel version of the regularized MVCE:

$$\begin{aligned} \boldsymbol{\alpha}_\gamma^* = \text{argmin}_{\boldsymbol{\alpha}} \quad & -\log \det \left( \sum_{i=1}^{\ell} \alpha_i \mathbf{c}_i \mathbf{c}'_i + \gamma \mathbf{I} \right), \\ \text{s.t.} \quad & \mathbf{e} \geq \nu \ell \boldsymbol{\alpha} \geq 0, \quad \boldsymbol{\alpha}' \mathbf{e} = 1. \end{aligned} \quad (6)$$

*Computing the Mahalanobis distance for a test point.* We should be able to compute the Mahalanobis distance for a test point exclusively using kernel evaluations and the vector  $\boldsymbol{\alpha}$ . Recall the eigenvalue decompositions of  $\sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i' = \mathbf{X}' \boldsymbol{\Lambda}^2 \mathbf{X} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}'$  and  $\mathbf{A} \mathbf{X} \mathbf{X}' \mathbf{A} = \mathbf{A} \mathbf{K} \mathbf{A} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}'$  [2]. We then have that  $\sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i' + \gamma \mathbf{I} = \mathbf{U} (\boldsymbol{\Lambda} + \gamma \mathbf{I}) \mathbf{U}' + \mathbf{U}^\perp (\gamma \mathbf{I}) \mathbf{U}^{\perp'}$  (where  $\mathbf{U}^\perp$  is an orthonormal basis for the space orthogonal to the column space of  $\mathbf{U}$ ). Thus we can write the Mahalanobis distance as (and we introduce the notation  $d_\gamma(\cdot, \boldsymbol{\alpha})$ ):

$$\begin{aligned} d_\gamma(\mathbf{x}, \boldsymbol{\alpha}) &\triangleq \mathbf{x}' \mathbf{M}^{-1} \mathbf{x} = \mathbf{x}' (\sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i' + \gamma \mathbf{I})^{-1} \mathbf{x} \\ &= \mathbf{x}' \left( \mathbf{U} (\boldsymbol{\Lambda} + \gamma \mathbf{I})^{-1} \mathbf{U}' + \mathbf{U}^\perp (\gamma \mathbf{I})^{-1} \mathbf{U}^{\perp'} \right) \mathbf{x} \\ &= \frac{1}{\gamma} \mathbf{x}' \mathbf{x} + \mathbf{x}' \mathbf{U} \left( (\boldsymbol{\Lambda} + \gamma \mathbf{I})^{-1} - (\gamma \mathbf{I})^{-1} \right) \mathbf{U}' \mathbf{x} \\ &= \frac{1}{\gamma} k(\mathbf{x}, \mathbf{x}) - \frac{1}{\gamma} \mathbf{x}' \mathbf{U} \left( \boldsymbol{\Lambda} (\boldsymbol{\Lambda} + \gamma \mathbf{I})^{-1} \right) \mathbf{U}' \mathbf{x}, \\ &= \frac{1}{\gamma} \left( k(\mathbf{x}, \mathbf{x}) - \mathbf{k}' \mathbf{A} \mathbf{V} \boldsymbol{\Lambda} (\boldsymbol{\Lambda} + \gamma \mathbf{I})^{-1} \mathbf{V}' \mathbf{A} \mathbf{k} \right), \end{aligned}$$

using  $\mathbf{U} = \mathbf{X}' \mathbf{A} \mathbf{V} \boldsymbol{\Lambda}^{-\frac{1}{2}}$  and  $\mathbf{X} \mathbf{x} = \mathbf{k}$ . This is expressed entirely in terms of kernels, since  $\mathbf{V}$  and  $\boldsymbol{\Lambda}$  can be found using the eigenvalue decomposition of  $\mathbf{A} \mathbf{K} \mathbf{A}$ .

## 5 Statistical Learning Analysis

Theoretically we can view the novelty detection problem in a space  $\mathcal{X}$  as the task of finding a set  $A \subset \mathcal{X}$  such that most of the support  $\text{supp}(\mathcal{D})$  of the distribution  $\mathcal{D}$  generating the data is contained in  $A$ ; that is

$$P_{\mathbf{x} \sim \mathcal{D}}(\mathbf{x} \in \text{supp}(\mathcal{D}) \setminus A) \leq \epsilon, \quad (7)$$

for some small  $\epsilon$ . This must be achieved while keeping the volume of  $A$  as small as possible, where in general the volume could be measured according to some prior distribution though in our case we consider the input space volume.

The motivation for this definition is to ‘shrink wrap’ the support of the training distribution as tightly as possible to increase the likelihood of detecting novel outliers. The bound of equation (7) upper bounds the probability that a point detected as an outlier (or novelty) is actually generated according to the original training distribution.

Earlier analyses of this type are based on covering number arguments [7] or Rademacher complexities [8], and deal with the case where the set  $A$  can be viewed as a hypersphere. However, it seems unnatural to use a spherical shape if the variance of the data varies significantly across different dimensions of the space. One would expect that we can use an elliptical shape with smaller diameters in the dimensions of low variance. The algorithm described in this paper implements just such a shape for the set  $A$  through the use of the Mahalanobis distance relative to the matrix  $\mathbf{M}$ . Introducing such flexibility into the shape of

the set  $A$  raises the question of whether the algorithm may not be overfitting the data and jeopardizing the confidence with which equation (7) can be asserted. This section will confirm that this concern is unfounded: that is we will prove a bound of the type given in equation (7) that holds with high confidence over the random selection of training sets according to the underlying distribution.

We first observe that the Mahalanobis distance  $d_\gamma(\mathbf{x}, \alpha)$  can be viewed as a linear function in the space defined by the kernel  $k(\mathbf{x}, \mathbf{z})^2$  where  $k(\mathbf{x}, \mathbf{z})$  is the kernel defining the feature space. This follows from the observation that

$$d_\gamma(\mathbf{x}, \alpha) = \text{trace}(\mathbf{M}^{-1}\mathbf{x}\mathbf{x}') = \langle \mathbf{M}^{-1}, \mathbf{x}\mathbf{x}' \rangle_F,$$

while:  $\langle \mathbf{x}\mathbf{x}', \mathbf{z}\mathbf{z}' \rangle = \langle \mathbf{x}, \mathbf{z} \rangle^2 = k(\mathbf{x}, \mathbf{z})^2$ .

Therefore, the critical quantity in analysing the generalization would appear to be the norm of the matrix  $\mathbf{M}^{-1}$ . Unfortunately this scales with the dimension of the space and so a naive application of standard Rademacher bounds would lead to a bound unsuitable for kernel defined feature spaces.

We will present a bound that uses the PAC-Bayes approach to generalization analysis in order to overcome this difficulty. As far as we are aware this is the first application of this technique to novelty detection.

The general PAC-Bayes theorem assumes a pre-specified ‘prior’ distribution  $P(c)$  over the class of classifiers. The learning algorithm returns a distribution  $Q(c)$  over the class and classification of an example  $\mathbf{x}$  is performed by drawing a classifier  $c$  randomly according to  $c \sim Q$  and using it to return the label  $c(\mathbf{x})$ . We denote by  $Q_{\mathcal{D}}$  the misclassification probability of  $Q$  on an example drawn according to  $\mathcal{D}$ . For a training set  $S$  of  $n$  examples, we denote by  $\hat{Q}_S$  the empirical misclassification error of  $Q$ . We will describe later how such a bound can be applied to the deterministic outlier detector that we consider. We use KL to denote the Kullback-Leibler divergence between two distributions:

$$\text{KL}(Q\|P) = E_{c \sim Q} \ln \frac{Q(c)}{P(c)}.$$

For  $p \in [0, 1]$  we overload the notation by using  $p$  to represent the binary distribution  $\{p, 1 - p\}$ . We can now state the theorem in a form due to Langford.

**Theorem 1.** [5] *For all  $\mathcal{D}$ , for all priors  $P(c)$ , and for all  $\delta \in (0, 1)$ ,*

$$P_{S \sim \mathcal{D}^n} \left( \forall Q(c) : \text{KL}(\hat{Q}_S \| Q_{\mathcal{D}}) \leq \frac{\text{KL}(Q\|P) + \ln \frac{n+1}{\delta}}{n} \right) \geq 1 - \delta.$$

Our application of the theorem to the novelty detector will follow closely the application to support vector machines as described in [6] and [5]. This involves choosing  $P$  to be a symmetric Gaussian prior of variance 1 but rather than being centered on the origin as in those papers, we choose the prior distribution to be centered at the point  $(\mu\gamma^{-1}I, 0)$  for some  $\mu > 0$ . Note that we are viewing the space as a Euclidean space with the Frobenius inner product with one extra dimension for the threshold. We augment the examples by adding a

coordinate equal to  $-1$  in this extra dimension. The posterior distribution  $Q(\mu)$  is now a spherically symmetric Gaussian with variance 1 centered at the point  $(\mu\mathbf{M}^{-1}, \mu\theta)$ , and  $\theta$  is a threshold such that a novelty is indicated if

$$d_\gamma(\mathbf{x}, \boldsymbol{\alpha}_\gamma^*) \geq \theta. \quad (8)$$

Clearly, equation (8) can be written as a linear function thresholded at 0 with weight vector  $(\mathbf{M}^{-1}, \theta)$ . If equation (8) holds for  $\mathbf{x}$  then  $Q(\mu)$  has probability at least 0.5 of being 1, hence

$$P(d_\gamma(\mathbf{x}, \boldsymbol{\alpha}_\gamma^*) \geq \theta) \leq 2Q(\mu)_\mathcal{D}.$$

It will therefore suffice to obtain an implicit bound on  $Q(\mu)_\mathcal{D}$  using Theorem 1.

We describe the critical quantities required in the theorem. Following [5] we require the function

$$\tilde{F}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

We denote the weight vector  $\mathbf{W} = (\mu\mathbf{M}^{-1}, \mu\theta)$ . The normalized margin of an example  $\mathbf{x}$  is given by

$$g(\mathbf{x}) = \frac{d_\gamma(\mathbf{x}, \boldsymbol{\alpha}_\gamma^*) - \theta}{\sqrt{\|\mathbf{x}\|^2 + 1}\|\mathbf{W}\|}.$$

The stochastic error rate is then

$$\hat{Q}(\mu)_S = E_{\mathbf{x} \sim S} \tilde{F}(\mu\|\mathbf{W}\|g(\mathbf{x})).$$

Finally, the KL-divergence between prior and posterior is given by

$$\text{KL}(Q\|P) = \frac{\mu^2}{2} (\|\gamma^{-1}\mathbf{I} - \mathbf{M}^{-1}\|^2 + \theta^2) = \frac{\mu^2}{2} \left( \sum_{i=1}^n \frac{\lambda_i^2}{\gamma^2(\lambda_i + \gamma)^2} + \theta^2 \right)$$

which critically is independent of the dimension of the feature space.

Putting the pieces together we obtain the following bound on the probability of misidentifying an outlier.

**Theorem 2.** *Fix  $\gamma > 0$  and  $\mu > 0$ . For all distributions  $\mathcal{D}$  and all  $\delta \in (0, 1)$ , we have with probability at least  $1 - \delta$  over the draw of an  $n$ -sample  $S$ , if  $\boldsymbol{\alpha}_\gamma^*$  is the solution of the novelty detection optimization then*

$$P_{\mathbf{x} \sim \mathcal{D}}(d_\gamma(\mathbf{x}, \boldsymbol{\alpha}_\gamma^*) \geq \theta) \leq 2Q(\mu)_\mathcal{D} \quad (9)$$

where  $Q(\mu)_\mathcal{D}$  satisfies

$$\text{KL}(\hat{Q}(\mu)_S\|Q(\mu)_\mathcal{D}) \leq \frac{\frac{\mu^2}{2} \left( \sum_{i=1}^n \frac{\lambda_i^2}{\gamma^2(\lambda_i + \gamma)^2} + \theta^2 \right) + \ln \frac{n+1}{\delta}}{n},$$

$$\text{and: } \hat{Q}(\mu)_S = E_{\mathbf{x} \sim S} \tilde{F} \left( \mu \frac{d_\gamma(\mathbf{x}, \boldsymbol{\alpha}_\gamma^*) - \theta}{\sqrt{\|\mathbf{x}\|^2 + 1}} \right).$$

Note that in practice one would apply the theorem for a number of different values of  $\mu$  and possibly different regularization parameter choices. If  $N$  applications are made then we should substitute  $\delta/N$  for  $\delta$  in the expression for the KL-divergence, but this only enters into the  $\ln$  term and so has a limited effect.

*Proof.* The only unresolved part of the proof is the verification of the expression for the stochastic error. We decompose the example  $(\mathbf{x}\mathbf{x}', -1)$  into two components  $\mathbf{X}_{\parallel}$  parallel to  $\mathbf{W}$  and  $\mathbf{X}_{\perp}$  perpendicular. The randomly drawn weight vector can be decomposed into three components  $\mathbf{U}_{\parallel}$  parallel to  $\mathbf{W}$  and distributed according to  $N(\mu\|\mathbf{W}\|, 1)$ ,  $\mathbf{U}_{\perp}$  parallel to  $\mathbf{X}_{\perp}$  distributed according to  $N(0, 1)$  and  $\mathbf{W}_{\perp\perp}$ . Let  $w = \|\mathbf{W}\|$ ,  $u_{\parallel} = \|\mathbf{U}_{\parallel}\|$ ,  $u_{\perp} = \|\mathbf{U}_{\perp}\|$ ,  $x_{\parallel} = \|\mathbf{X}_{\parallel}\|$ , and  $x_{\perp} = \|\mathbf{X}_{\perp}\|$ . Then we have, as required:

$$\begin{aligned}\hat{Q}(\mu)_S &= E_{\mathbf{x}\sim S, u_{\parallel}\sim N(\mu w, 1), u_{\perp}\sim N(0, 1)} I(u_{\parallel}x_{\parallel} + u_{\perp}x_{\perp} \geq 0) \\ &= E_{\mathbf{x}\sim S, z\sim N(0, 1), v\sim N(0, 1)} I((\mu w + z)x_{\parallel} + vx_{\perp} \geq 0) \\ &= E_{\mathbf{x}\sim S, z\sim N(0, 1), v\sim N(0, 1)} I\left(\mu w \geq z + v\frac{x_{\perp}}{x_{\parallel}}\right) \\ &= E_{\mathbf{x}\sim S, z\sim N\left(0, 1 + \frac{x_{\perp}^2}{x_{\parallel}^2}\right)} I(\mu w \geq z) = E_{\mathbf{x}\sim S} E_{z\sim N\left(0, \frac{1}{g(\mathbf{x})^2}\right)} I(\mu w \geq z) \\ &= E_{\mathbf{x}\sim S} \tilde{F}(\mu w g(\mathbf{x})).\end{aligned}$$

## 6 Experiment: condition monitoring

The purpose of this section is to analyse the comparative performance of the proposed soft margin kernel RMVCE algorithm and the one-class SVM algorithm on a real-life dataset from the Structural Integrity and Damage Assessment Network [1]. There are vibration measurements in this dataset that correspond to “healthy” measurements (without fault) and 4 types of malfunction of machinery: Fault 1, Fault 2, Fault 3 and Fault 4 (see [1] for details). In order to compare the proposed RMVCE method (see (6)) with the one-class SVM method [7, 11], we performed experiments in the similar manner as described in [1]: 1) “Healthy” measurements ( $\ell = 150$ ) are used to train the RMVCE (see (6)) and the one-class SVM [7]; 2) one hundred fifty samples (Fault 1) are used to validate the results of training. It can be seen that the proposed RMVCE can be successfully used for novelty detection as a valuable alternative to the minimum volume enclosing hypersphere for novelty detection (see Table 1). The RMVCE method can be also applied to Gaussian Processes to perform optimal experimental design [4].

## 7 Conclusions

We have tackled the novelty detection problem using the MVCE. While the MVCE can directly be used in low dimensional spaces, it is problematic in high dimensional spaces. To resolve this, we introduced regularisation, which allowed us to derive a learning theory bound guaranteeing a maximal probability of

**Table 1.** The percentage of correctly labeled classes using one-class SVM and RMVCE methods with Gaussian kernel,  $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-0.5\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$

| Method     | $\sigma$ | $\nu$ | $\gamma$ | Healthy     | Fault 1     | Fault 2     | Fault 3    | Fault 4    |
|------------|----------|-------|----------|-------------|-------------|-------------|------------|------------|
| RMVCE, 320 | 0.3      | 0.02  |          | <b>100%</b> | <b>91%</b>  | <b>100%</b> | <b>90%</b> | <b>61%</b> |
| RMVCE, 320 | 0.25     | 0.02  |          | 92%         | 100%        | 85%         | 55%        | 75%        |
| 1-SVM      | 320      | 0.25  | -        | 79%         | 100%        | 98%         | 85%        | 93%        |
| 1-SVM      | 320      | 0.001 | -        | <b>90%</b>  | <b>100%</b> | <b>95%</b>  | <b>68%</b> | <b>85%</b> |

misidentifying an outlier. Finally, we presented a kernel version allowing to model nonlinearly shaped supports and supports for structured data types.

*Acknowledgements.* T.De Bie acknowledges support from the CoE EF/05/007 SymBioSys, and from GOA/2005/04, both from the Research Council K.U.Leuven.

## References

1. Campbell, C., Bennett, K.P. A Linear Programming Approach to Novelty Detection, Advances in Neural Information Processing Systems 14 (NIPS01), 2002.
2. Dolia, A. N., Harris, C. J., Shawe-Taylor, J. and Titterton, D. M. (2005) Kernel Ellipsoidal Trimming, Internal Report T8.11.10-01/05, School of Electronics and Computer Science, University of Southampton, 11 October 2005.
3. Dolia, A.N., Page, S.F., White, N.M., Harris, C.J. D-optimality for Minimum Volume Ellipsoid with Outliers, Proc. of the 7th International Conference on Signal/Image Processing and Pattern Recognition, (UkrOBRAZ'2004), 73–76, 2004.
4. Dolia, A.N., Harris, C.J., Shawe-Taylor, J., De Bie, T. (2006) Gaussian Processes for Active Sensor Management, Gaussian Processes in Practice Workshop, Bletchley Park, UK 12 - 13 June 2006
5. Langford, J. Tutorial on Practical Prediction Theory for Classification, Journal of Machine Learning Research, **6**:273–306, 2005.
6. Langford, J. and Shawe-Taylor, J. "PAC Bayes and Margins, Advances in Neural Information Processing Systems 15 (NIPS02), 2003.
7. B. Schölkopf and J.C. Platt and J.S. Shawe-Taylor and A.J. Smola and R.C. Williamson. Estimating the Support of a High-Dimensional Distribution, Neural Computation, **13**(7):1443–1471, 2001.
8. Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK (2004)
9. 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. Proc. of the 13th International Conference on Algorithmic Learning Theory (ALT2002) **2533**, 2002.
10. Sun, P. and Freund, R.M. Computation of Minimum-Volume Covering Ellipsoids, Operations Research **52**(5):690–706, 2004.
11. Tax, D.M.J., Duin, R.P.W. Data Domain Description by Support vectors. Verleysen, M. (ed.). Proceedings, ESANN. Brussels, 251–256, 1999.
12. Titterton, D.M. Estimation of Correlation Coefficients by Ellipsoidal Trimming, Journal of Royal Statistical Society **C27**(3):227–234, 1978.