2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Learning the Kernel in Mahalanobis One-Class Support Vector Machines

Ivor W. Tsang, James T. Kwok, Shutao Li

Abstract— **In this paper, we show that one-class SVMs can also utilize data covariance in a robust manner to improve performance. Furthermore, by constraining the desired kernel function as a convex combination of base kernels, we show that the weighting coefficients can be learned via quadratically constrained quadratic programming (QCQP) or second order cone programming (SOCP) methods. Performance on both toy and real-world data sets show promising results. This paper thus offers another demonstration of the synergy between convex optimization and kernel methods.**

I. Introduction

In recent years, kernel methods have been successfully used in various aspects of machine learning, such as classification, regression and clustering [1]. In this paper, we will focus on the use of one-class support vector machines (SVMs) [2] for novelty detection, in which only a set of unlabeled patterns are given. The one-class SVM, like other kernel methods, first maps the data from the input space to a feature space $\mathcal{H}$ via some map φ, and then constructs a hyperplane in $\mathcal{H}$ that separates the φ-mapped patterns from the origin with maximum margin. The computations do not require φ explicitly, but depend only on the inner product defined in $\mathcal{H}$, which in turn can be obtained efficiently from a suitable kernel function (the "kernel trick"). The one-class SVM also closely resembles the support vector data description [3], which uses balls (instead of hyperplanes) to describe the data in $\mathcal{H}$. In fact, these two approaches are equivalent when stationary kernels are used [2].

However, one-class SVMs rely on the Euclidean distance, which is often sub-optimal. A standard alternative is to utilize information from the data, such as the readily accessible sample covariance matrix. For example, the single-class minimax probability machine (MPM) [4], which is another kernel-based technique for novelty detection, maximizes the Mahalanobis distance of the hyperplane to the origin instead. In the context of supervised learning, the covariance of different classes have also been used to improve the performance of the SVM [5]. Moreover, to alleviate the undesirable effects of estimation error in the covariance matrix, [4] adopted an uncertainty model for the sample mean and covariance matrix, and then used robust optimization to address this estimation problem.

Another issue in using one-class SVMs is the choice of kernels. As in other kernel methods, because of the central role of the kernel, a poor kernel choice can lead to significantly impaired performance. As reported in [6], one-class SVMs can be very sensitive in this aspect. In the supervised learning setting, progress has been made in the past few years on how to choose the parameters of a kernel with fixed parametric form. Typically, this is performed by optimizing a quality functional of the kernel [7], such as the kernel target alignment, generalization error bounds, Bayesian probabilities and cross-validation error. Recently, instead of adapting only the kernel parameters, one also attempts to adapt the form of the kernel directly. As all information on the feature space is encoded in the kernel matrix, one can bypass learning of the kernel function by just learning the kernel matrix instead [8], [9], [10], [11]. These methods, however, usually work better in a transductive setting. For induction, a novel approach that selects the kernel function directly is by using the hyperkernel [7]. However, all these results are designed for supervised learning and not readily applicable to one-class SVMs.

In this paper, we first show that covariance information can also be utilized in a robust manner by one-class SVMs. This includes an uncertainty model on the covariance matrix which is more general than the one used by single-class MPMs. Furthermore, by constraining the kernel function in the one-class SVM as a convex combination of some fixed base kernels, we show that the weighting coefficients can be learned by convex programming techniques. The rest of this paper is organized as follows. Section II describes the robust use of covariance information in one-class SVMs. Section III then addresses the problem of kernel learning in one-class SVMs. Experimental results are presented in Section IV, and the last section gives some concluding remarks. Because of the lack of space, detailed proofs cannot be included in this paper.

II. One-Class SVM with the Mahalanobis Distance

Given a set of unlabeled patterns $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the one-class SVM first maps them to the feature space $\mathcal{H}$ via a nonlinear map φ. In the sequel, for simplicity, we will abuse the notation and still write $\varphi(\mathbf{x})$ as $\mathbf{x}$. The data is then separated from the origin by solving

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}\mathbf{w}'\mathbf{w} + \frac{1}{\nu n}\sum_i \xi_i - \rho \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{x}_i \geq \rho - \xi_i, \\ & \xi_i \geq 0, \end{aligned}$$

where $\mathbf{w}'\mathbf{x} = \rho$ is the desired hyperplane and $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n]'$. The corresponding dual (with $\boldsymbol{\alpha} =$

Ivor W. Tsang and James T. Kwok are with the Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. Shutao Li is with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China.

0-7803-9490-9/06/$20.00/©2006 IEEE

$[\alpha_1, \ldots, \alpha_n]', \mathbf{1} = [1, \ldots, 1]'$ and kernel matrix $\mathbf{K}$)

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} && (1) \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}'\mathbf{1} = 1, && (2) \end{aligned}$$

is a quadratic programming (QP) problem.

A. Using the Covariance Information by Robust Optimization

As mentioned in Section I, it is often beneficial to utilize the covariance matrix $\boldsymbol{\Sigma}$ and use the Mahalanobis distance instead. Writing $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, a common estimator for $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma}^0 = c\mathbf{X}\mathbf{H}\mathbf{X}',$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$[1], $\mathbf{I}$ is the identity matrix, and $c = \frac{1}{n}$ (or $\frac{1}{n-1}$) for the maximum likelihood (or sample) covariance matrix. The primal now becomes:

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}^{-1}\mathbf{w} + \frac{1}{\nu n}\sum_i \xi_i - \rho && (3) \\ \text{s.t.} \quad & \mathbf{w}'\boldsymbol{\Sigma}^{-1}\mathbf{x}_i \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

Putting $\mathbf{w} = \boldsymbol{\Sigma}\mathbf{u}$, (3) is equivalent to

$$\begin{aligned} \min_{\mathbf{u},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u} + \frac{1}{\nu n}\sum_i \xi_i - \rho && (4) \\ \text{s.t.} \quad & \mathbf{u}'\mathbf{x}_i \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

(3) is thus the same as still using the Euclidean metric, but maximizes instead the Mahalanobis distance of the plane $\mathbf{u}'\mathbf{x} = \rho$ to the origin (which is given by $\rho/\sqrt{\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}}$ [4]). In the sequel, we will use the formulation in (4) (and write $\mathbf{w}$ instead of $\mathbf{u}$).

In general, there is uncertainty in the estimation of $\boldsymbol{\Sigma}$. As in [4], we assume that $\boldsymbol{\Sigma}$ is only known to be within the set

$$\{\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^0\|_F \leq r\},$$

where $r > 0$ is fixed and $\|\cdot\|_F$ denotes the Frobenius norm. The primal in (4) can then be modified as

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},\rho} \max_{\boldsymbol{\Sigma}} \quad & \frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} + \frac{1}{\nu n}\sum_i \xi_i - \rho && (5) \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{x}_i \geq \rho - \xi_i, \\ & \xi_i \geq 0, \\ & \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^0\|_F \leq r. \end{aligned}$$

Now,

$$\max_{\boldsymbol{\Sigma}:\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^0\|_F \leq r} \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} = \mathbf{w}'(r\mathbf{I} + \boldsymbol{\Sigma}^0)\mathbf{w}$$

[1]The following properties on $\mathbf{H}$ can be easily verified: $\mathbf{H} = \mathbf{H}', \mathbf{H}\mathbf{H} = \mathbf{H}$ and $\mathbf{H}\mathbf{1} = \mathbf{0}$.

[4]. Therefore, (5) becomes

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}_r\mathbf{w} + \frac{1}{\nu n}\sum_i \xi_i - \rho \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{x}_i \geq \rho - \xi_i, \\ & \xi_i \geq 0, \end{aligned}$$

where $\boldsymbol{\Sigma}_r = r\mathbf{I} + \boldsymbol{\Sigma}^0 = r\mathbf{I} + c\mathbf{X}\mathbf{H}\mathbf{X}'$ is always non-singular for $r > 0$. In effect, this is similar to the common trick of making $\boldsymbol{\Sigma}^0$ non-singular. The corresponding dual is then:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}'\mathbf{X}'\boldsymbol{\Sigma}_r^{-1}\mathbf{X}\boldsymbol{\alpha} && (6) \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}'\mathbf{1} = 1. \end{aligned}$$

As we would expect, when the covariance information is not used (i.e., $c = 0$), (6) reduces to the original dual in (1). Using the Woodbury formula [12]

$$(\mathbf{A} + \mathbf{B}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

and $\mathbf{H}\mathbf{H} = \mathbf{H}$, we obtain

$$\begin{aligned} \boldsymbol{\Sigma}_r^{-1} &= (r\mathbf{I} + c\mathbf{X}\mathbf{H}\mathbf{H}\mathbf{X}')^{-1} \\ &= \frac{1}{r}\mathbf{I} - \frac{c}{r}\mathbf{X}\mathbf{H}\left(r\mathbf{I} + c\mathbf{H}\mathbf{X}'\mathbf{X}\mathbf{H}\right)^{-1}\mathbf{H}\mathbf{X}'. \end{aligned}$$

(6) then becomes

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2r}\boldsymbol{\alpha}'\left(\mathbf{K} - c\mathbf{K}\mathbf{H}(r\mathbf{I} + c\mathbf{H}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}\mathbf{K}\right)\boldsymbol{\alpha} && (7) \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}'\mathbf{1} = 1, \end{aligned}$$

where $\mathbf{K} = \mathbf{X}'\mathbf{X}$ is the kernel matrix[2]. This is again a standard QP. Moreover, when $\mathbf{K}$ is invertible, (7) can be further simplified to

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}'\left(r\mathbf{K}^{-1} + c\mathbf{H}\right)^{-1}\boldsymbol{\alpha} && (8) \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}'\mathbf{1} = 1, \end{aligned}$$

by using the Woodbury formula. Besides, as for the original one-class SVM, $\nu \in (0, 1)$ is an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors.

B. A More General Uncertainty Model

In this Section, the uncertainty set takes a more general form, as

$$\{\boldsymbol{\Sigma} : 0 \preceq \boldsymbol{\Sigma} \preceq \boldsymbol{\Sigma}^0 + \boldsymbol{\Delta}\},$$

where $\boldsymbol{\Delta} \succ 0$. Here, the notation $\mathbf{A} \succeq 0$ ($\mathbf{A} \succ 0$) means that the matrix $\mathbf{A}$ is symmetric and positive semi-definite (definite). Similarly, $\preceq$ and $\prec$ means negative semi-definite (definite). When $\boldsymbol{\Delta} = \frac{r}{\sqrt{n}}\mathbf{I}$, this reduces to the

[2]Recall that our $\mathbf{x}_i$'s here are in fact $\varphi(\mathbf{x}_i)$'s in the kernel-induced feature space.

uncertainty model in Section II-A. Alternatively, $\mathbf{\Delta}$ can also be considered as a more general prior on $\mathbf{w}$ [4]. Now, for any $\mathbf{w}$,

$$\begin{aligned} & \mathbf{\Sigma}^0 + \mathbf{\Delta} \succeq \mathbf{\Sigma} \\ \Rightarrow \quad & \mathbf{w}'(\mathbf{\Sigma}^0 + \mathbf{\Delta})\mathbf{w} \geq \mathbf{w}'\mathbf{\Sigma}\mathbf{w}, \end{aligned}$$

with the equality attained when $\mathbf{\Sigma} = \mathbf{\Sigma}^0 + \mathbf{\Delta}$. Hence,

$$\max_{0 \preceq \mathbf{\Sigma} \preceq \mathbf{\Sigma}^0 + \mathbf{\Delta}} \mathbf{w}'\mathbf{\Sigma}\mathbf{w} = \mathbf{w}'(\mathbf{\Sigma}^0 + \mathbf{\Delta})\mathbf{w}.$$

In other words, we can follow the same steps in Section II-A by simply replacing $\mathbf{\Sigma}_r$ by $\mathbf{\Sigma}^0 + \mathbf{\Delta}$, and obtain the primal as

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}\mathbf{w}'(\mathbf{\Delta} + c\mathbf{X}\mathbf{H}\mathbf{X}')\mathbf{w} + \frac{1}{\nu n}\sum_i \xi_i - \rho \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{x}_i \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

By using the Woodbury formula and recalling that $\mathbf{H}\mathbf{H} = \mathbf{H}$, we have

$$\begin{aligned} & (\mathbf{\Delta} + c\mathbf{X}\mathbf{H}\mathbf{H}\mathbf{X}')^{-1}| \\ = \quad & \mathbf{\Delta}^{-1} - \mathbf{\Delta}^{-1}\mathbf{X}\mathbf{H}(\frac{1}{c}\mathbf{I} + \mathbf{H}\mathbf{X}'\mathbf{\Delta}^{-1}\mathbf{X}\mathbf{H})^{-1}\mathbf{H}\mathbf{X}'\mathbf{\Delta}^{-1}, \end{aligned}$$

and the dual becomes

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}'\left(\tilde{\mathbf{K}} - \tilde{\mathbf{K}}\mathbf{H}(\frac{1}{c}\mathbf{I} + \mathbf{H}\tilde{\mathbf{K}}\mathbf{H})^{-1}\mathbf{H}\tilde{\mathbf{K}}\right)\boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}'\mathbf{1} = 1, \end{aligned}$$

where $\tilde{\mathbf{K}} = \mathbf{X}'\mathbf{\Delta}^{-1}\mathbf{X}$. This, again, is a QP. When $\tilde{\mathbf{K}}$ is invertible, by using the Woodbury formula, the dual can be reduced to

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}'(\tilde{\mathbf{K}}^{-1} + c\mathbf{H})^{-1}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}'\mathbf{1} = 1. \end{aligned} \tag{9}$$

III. Learning the Kernel Matrix

Notice that the objectives in (1), (8) and (9) are of the same form, namely,

$$\frac{1}{2}\boldsymbol{\alpha}'(\hat{\mathbf{K}}^{-1} + c\mathbf{H})^{-1}\boldsymbol{\alpha}. \tag{10}$$

$\hat{\mathbf{K}}$ thus embodies information on both the original kernel $\mathbf{K}$ and the uncertainty model of the data covariance. In this Section, we consider learning this $\hat{\mathbf{K}}$ directly. As the uncertainty model corresponds to a prior on $\mathbf{w}$, learning $\hat{\mathbf{K}}$ also learns this prior from the empirical data, in the same spirit as empirical Bayes methods. We constrain the target kernel function $\hat{K}$ to be a convex combination of some fixed base kernels K_i's, i.e.,

$$\hat{K} = \sum_{i=1}^m \mu_i K_i, \tag{11}$$

where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_m]' \geq \mathbf{0}$, and $\boldsymbol{\mu}'\mathbf{1} = 1$. As usual, the corresponding kernel matrices defined on the training set will be denoted in bold. Obviously, $\mathbf{K}_i \succeq 0$ for all base kernels implies $\hat{\mathbf{K}} \succeq 0$. While one may want to directly minimize the objective in (10) over the allowable $\hat{\mathbf{K}}$'s, this is not desirable as different kernels will induce different feature spaces with different scales. A kernel can easily "cheat" by simply expanding the data distribution (in the feature space) and thus obtain a large margin. Hence, some normalization is necessary in order to compare the margins in a meaningful manner.

A. Modified One-Class SVM Formulation

In this Section, we offer a simple remedy by modifying the primal in (4) to

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},\rho,R} \quad & \frac{1}{2}\mathbf{w}'\mathbf{\Sigma}\mathbf{w} + \frac{1}{\nu n}\sum_i \xi_i - \rho + CR \\ \text{s.t.} \quad & 1 \leq \mathbf{w}'\mathbf{x}_i \leq R, \\ & \mathbf{w}'\mathbf{x}_i \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

The constraint $1 \leq \mathbf{w}'\mathbf{x}_i \leq R$ sets a scale in the kernel-induced feature space. Those kernels that achieve a large margin by simply having a large R will get penalized in the primal objective. By introducing Lagrange multipliers

$$\begin{aligned} \boldsymbol{\alpha}_h &= [\alpha_{h1}, \ldots, \alpha_{hn}]' \geq \mathbf{0}, \\ \boldsymbol{\alpha}_r &= [\alpha_{r1}, \ldots, \alpha_{rn}]' \geq \mathbf{0}, \\ \boldsymbol{\alpha}_s &= [\alpha_{s1}, \ldots, \alpha_{sn}]' \geq \mathbf{0}, \\ \boldsymbol{\eta} &= [\eta_1, \ldots, \eta_l]' \geq \mathbf{0}, \end{aligned}$$

the Lagrangian is then:

$$\begin{aligned} & \mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \rho, R, \boldsymbol{\alpha}, \boldsymbol{\eta}) \\ = \quad & \frac{1}{2}\mathbf{w}'\mathbf{\Sigma}\mathbf{w} + \frac{1}{\nu n}\sum_i \xi_i - \rho + CR - \sum_i \alpha_{hi}(\mathbf{w}'\mathbf{x}_i - 1) \\ & - \sum_i \alpha_{ri}(R - \mathbf{w}'\mathbf{x}_i) - \sum_i \alpha_{si}(\mathbf{w}'\mathbf{x}_i - \rho + \xi_i) \\ & - \sum_i \eta_i \xi_i. \end{aligned}$$

where, for simplicity of notation, we have encapsulated $\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_h$ and $\boldsymbol{\alpha}_r$ together as $\boldsymbol{\alpha}$. Setting the derivatives of $\mathcal{L}$ w.r.t. all primal variables to zero, and assuming that $\mathbf{\Sigma}$ is non-singular, the dual becomes:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}_h'\mathbf{1} \\ & -\frac{1}{2}(\boldsymbol{\alpha}_s + \boldsymbol{\alpha}_h - \boldsymbol{\alpha}_r)'\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}(\boldsymbol{\alpha}_s + \boldsymbol{\alpha}_h - \boldsymbol{\alpha}_r) \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha}_s \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\ & \boldsymbol{\alpha}_h \geq \mathbf{0}, \\ & \boldsymbol{\alpha}_r \geq \mathbf{0}, \\ & \boldsymbol{\alpha}_r'\mathbf{1} = C. \end{aligned}$$

Proceeding as in Section II-B with the uncertainty models and together with (11), we obtain

$$
\begin{aligned}
\min_{\hat{\mathbf{K}}} \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}_h'\mathbf{1} \\
& -\frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'(\hat{\mathbf{K}}^{-1}+c\mathbf{H})^{-1}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r) \\
\text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha}_s \le \frac{1}{\nu n}\mathbf{1}, \\
& \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\
& \boldsymbol{\alpha}_h \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r'\mathbf{1} = C, \\
& \hat{\mathbf{K}} = \sum_i \mu_i \mathbf{K}_i, \\
& \boldsymbol{\mu}'\mathbf{1} = 1, \\
& \boldsymbol{\mu} \ge \mathbf{0}.
\end{aligned} \tag{12}
$$

B. Without Use of Covariance Information: A QCQP Formulation

First, consider the special case when covariance is not used ($c = 0$). (12) then reduces to

$$
\begin{aligned}
\min_{\hat{\mathbf{K}}} \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}_h'\mathbf{1} \\
& -\frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'\hat{\mathbf{K}}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r) \\
\text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha}_s \le \frac{1}{\nu n}\mathbf{1}, \\
& \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\
& \boldsymbol{\alpha}_h \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r'\mathbf{1} = C, \\
& \hat{\mathbf{K}} = \sum_i \mu_i \mathbf{K}_i, \\
& \boldsymbol{\mu}'\mathbf{1} = 1, \\
& \boldsymbol{\mu} \ge \mathbf{0} \\
= \quad \min_{\boldsymbol{\mu}'\mathbf{1}=1,\ \boldsymbol{\mu}\ge\mathbf{0}} \quad & \max_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}_h'\mathbf{1} \\
& -\sum_i \mu_i \left(\frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'\mathbf{K}_i(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)\right) \\
& \mathbf{0} \le \boldsymbol{\alpha}_s \le \frac{1}{\nu n}\mathbf{1}, \\
& \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\
& \boldsymbol{\alpha}_h \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r'\mathbf{1} = C.
\end{aligned}
$$

The Slater's condition [11] is satisfied and we can interchange $\min$ and $\max$, as:

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \min_{\boldsymbol{\mu}'\mathbf{1}=1,\ \boldsymbol{\mu}\ge\mathbf{0}} \boldsymbol{\alpha}_h'\mathbf{1} \\
& -\sum_i \mu_i \left(\frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'\mathbf{K}_i(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)\right) \\
\text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha}_s \le \frac{1}{\nu n}\mathbf{1}, \\
& \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\
& \boldsymbol{\alpha}_h \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r'\mathbf{1} = C \\
= \quad \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}_h'\mathbf{1} - \left[\max_{\boldsymbol{\mu}'\mathbf{1}=1,\ \boldsymbol{\mu}\ge\mathbf{0}} \right. \\
& \left. \sum_i \mu_i \left(\frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'\mathbf{K}_i(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)\right)\right] \\
\text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha}_s \le \frac{1}{\nu n}\mathbf{1}, \\
& \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\
& \boldsymbol{\alpha}_h \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r'\mathbf{1} = C \\
= \quad \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}_h'\mathbf{1} \\
& -\max_i \frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'\mathbf{K}_i(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r) \\
\text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha}_s \le \frac{1}{\nu n}\mathbf{1}, \\
& \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\
& \boldsymbol{\alpha}_h \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r'\mathbf{1} = C \\
= \quad \max_{\boldsymbol{\alpha},t} \quad & \boldsymbol{\alpha}_h'\mathbf{1} - t \\
\text{s.t.} \quad & t \ge \frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'\mathbf{K}_i(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r), \\
& \mathbf{0} \le \boldsymbol{\alpha}_s \le \frac{1}{\nu n}\mathbf{1}, \\
& \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\
& \boldsymbol{\alpha}_h \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r \ge \mathbf{0}, \\
& \boldsymbol{\alpha}_r'\mathbf{1} = C,
\end{aligned}
$$

which is a quadratically constrained quadratic programming (QCQP) problem.

C. With the Use of Covariance Information: A SOCP Formulation

We now return to (12) (with $c \neq 0$). First, consider the sub-problem involving $\boldsymbol{\alpha}$,

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}_h'\mathbf{1} \\ & -\frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'(\hat{\mathbf{K}}^{-1}+c\mathbf{H})^{-1}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r) \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha}_s \leq \frac{1}{\nu n}\mathbf{1}, \\ & \boldsymbol{\alpha}_s'\mathbf{1} = 1, \\ & \boldsymbol{\alpha}_h \geq \mathbf{0}, \\ & \boldsymbol{\alpha}_r \geq \mathbf{0}, \\ & \boldsymbol{\alpha}_r'\mathbf{1} = C. \end{aligned} \tag{13}$$

By introducing Lagrange multipliers

$$\begin{aligned} \boldsymbol{\gamma}_h &= [\gamma_{h1},\ldots,\gamma_{hn}]' \geq \mathbf{0}, \\ \boldsymbol{\gamma}_r &= [\gamma_{r1},\ldots,\gamma_{rn}]' \geq \mathbf{0}, \\ \boldsymbol{\gamma}_s &= [\gamma_{s1},\ldots,\gamma_{sn}]' \geq \mathbf{0}, \\ \boldsymbol{\beta} &= [\beta_1,\ldots,\beta_n]' \geq \mathbf{0} \end{aligned}$$

and λ_s, λ_r, the Lagrangian is then:

$$\begin{aligned} & \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda}) \\ = \quad & \boldsymbol{\alpha}_h'\mathbf{1} \\ & -\frac{1}{2}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r)'(\hat{\mathbf{K}}^{-1}+c\mathbf{H})^{-1}(\boldsymbol{\alpha}_s+\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_r) \\ & +\boldsymbol{\gamma}_s'\boldsymbol{\alpha}_s+\boldsymbol{\beta}'(\frac{1}{\nu n}\mathbf{1}-\boldsymbol{\alpha}_s)+\lambda_s(\boldsymbol{\alpha}_s'\mathbf{1}-1) \\ & +\boldsymbol{\gamma}_h'\boldsymbol{\alpha}_h+\boldsymbol{\gamma}_r'\boldsymbol{\alpha}_r+\lambda_r(C-\boldsymbol{\alpha}_r'\mathbf{1}), \end{aligned}$$

where, again, we have used $\boldsymbol{\gamma}$ to represent $(\boldsymbol{\gamma}_s,\boldsymbol{\gamma}_h,\boldsymbol{\gamma}_r)$ and $\boldsymbol{\lambda}$ for (λ_s,λ_r). As (13) is a QP,

$$\begin{aligned} & \max_{\boldsymbol{\alpha}} \min_{\boldsymbol{\gamma},\boldsymbol{\beta}\geq\mathbf{0},\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda}) \\ = \quad & \min_{\boldsymbol{\gamma},\boldsymbol{\beta}\geq\mathbf{0},\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda}). \end{aligned}$$

For $\max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda})$, the derivatives of $\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda})$ w.r.t. $\boldsymbol{\alpha}$ are zero. Substituting these back into (13) and on using $\mathbf{H}\mathbf{1} = \mathbf{0}$, the dual becomes:

$$\begin{aligned} & \min_{\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda}} \ \frac{1}{2}(\boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1})'(\hat{\mathbf{K}}^{-1}+c\mathbf{H})(\boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1}) \\ & \qquad +\frac{1}{\nu n}\boldsymbol{\beta}'\mathbf{1}-\lambda_s+C\lambda_r \\ & \qquad \text{s.t.} \quad \boldsymbol{\gamma}_s \geq \mathbf{0}, \\ & \qquad \boldsymbol{\gamma}_h \geq \mathbf{0}, \\ & \qquad \boldsymbol{\gamma}_r \geq \mathbf{0}, \\ & \qquad \boldsymbol{\beta} \geq \mathbf{0}, \\ & \qquad \boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1} = -\boldsymbol{\gamma}_r+\lambda_r\mathbf{1}, \\ & \qquad \boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1} = \boldsymbol{\gamma}_h+\mathbf{1} \\ = \quad & \min_{\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda},t_1,t_2} \ \frac{1}{2}t_1+\frac{c}{2}t_2+\frac{1}{\nu n}\boldsymbol{\beta}'\mathbf{1}-\lambda_s+C\lambda_r \\ & \qquad \boldsymbol{\gamma}_s \geq \mathbf{0}, \\ & \qquad \boldsymbol{\gamma}_h \geq \mathbf{0}, \\ & \qquad \boldsymbol{\gamma}_r \geq \mathbf{0}, \\ & \qquad \boldsymbol{\beta} \geq \mathbf{0}, \\ & \qquad \boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1} = -\boldsymbol{\gamma}_r+\lambda_r\mathbf{1}, \\ & \qquad \boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1} = \boldsymbol{\gamma}_h+\mathbf{1}, \\ & \qquad t_1 \geq (\boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1})'\hat{\mathbf{K}}^{-1}(\boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1}), \\ & \qquad t_2 \geq (\boldsymbol{\gamma}_s-\boldsymbol{\beta})'\mathbf{H}(\boldsymbol{\gamma}_s-\boldsymbol{\beta}). \end{aligned} \tag{14}$$

Recall that $\hat{\mathbf{K}}$ is of the form in (11), [13], [14] show that the constraint

$$t_1 \geq (\boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1})'\hat{\mathbf{K}}^{-1}(\boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1})$$

above can then be replaced by

$$\begin{aligned} \sum_i \mathbf{K}_i^{\frac{1}{2}}\mathbf{c}_i &= \boldsymbol{\gamma}_s-\boldsymbol{\beta}+\lambda_s\mathbf{1}, \\ \sum_i \tau_i &\leq t_1, \\ \boldsymbol{\tau} &\geq \mathbf{0}, \\ \mathbf{c}_i'\mathbf{c}_i &\leq \mu_i\tau_i. \end{aligned}$$

Moreover, the constraints

$$\mu_i\tau_i \geq \mathbf{c}_i'\mathbf{c}_i$$

and

$$t_2 \geq (\boldsymbol{\gamma}_s-\boldsymbol{\beta})'\mathbf{H}(\boldsymbol{\gamma}_s-\boldsymbol{\beta})$$

can be converted to second-order cone constraints by using the fact that the constraint $\mathbf{w}'\mathbf{w} \leq xy$ (where $x, y \geq 0$) is equivalent to the constraint

$$\left\| \begin{bmatrix} 2\mathbf{w} \\ x-y \end{bmatrix} \right\| \leq x+y$$

[13]. Applying these conversions, and together with the optimization w.r.t. $\boldsymbol{\mu}$ in (12), we finally obtain

$$
\begin{aligned}
\min_{\boldsymbol{\mu},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\lambda},t_1,t_2,\boldsymbol{\tau},\mathbf{c}_i} \quad & \frac{1}{2}t_1 + \frac{c}{2}t_2 + \frac{1}{\nu n}\boldsymbol{\beta}'\mathbf{1} - \lambda_s + C\lambda_r \\
\text{s.t.} \quad & \boldsymbol{\gamma}_s \geq \mathbf{0}, \\
& \boldsymbol{\gamma}_h \geq \mathbf{0}, \\
& \boldsymbol{\gamma}_r \geq \mathbf{0}, \\
& \boldsymbol{\beta} \geq \mathbf{0}, \\
& \boldsymbol{\gamma}_s - \boldsymbol{\beta} + \lambda_s\mathbf{1} = -\boldsymbol{\gamma}_r + \lambda_r\mathbf{1}, \\
& \boldsymbol{\gamma}_s - \boldsymbol{\beta} + \lambda_s\mathbf{1} = \boldsymbol{\gamma}_h + \mathbf{1}, \\
& \sum_i \mathbf{K}_i^{\frac{1}{2}}\mathbf{c}_i = \boldsymbol{\gamma}_s - \boldsymbol{\beta} + \lambda_s\mathbf{1}, \\
& t_1 \geq \sum_i \tau_i, \\
& \boldsymbol{\mu} \geq \mathbf{0}, \\
& \boldsymbol{\tau} \geq \mathbf{0}, \\
& \boldsymbol{\mu}'\mathbf{1} = 1, \\
& \mu_i + \tau_i \geq \left\| \begin{bmatrix} 2\mathbf{c}_i \\ \mu_i - \tau_i \end{bmatrix} \right\|, \\
& t_2 + 1 \geq \left\| \begin{bmatrix} 2\mathbf{H}(\boldsymbol{\gamma}_s - \boldsymbol{\beta}) \\ t_2 - 1 \end{bmatrix} \right\|,
\end{aligned}
$$

which is a second order cone programming (SOCP) problem.

IV. Experiments

We first perform experiments on a toy problem, with the "normal" data coming from a banana-shaped set. 50 "normal" points are used for training the (Mahalanobis) one-class SVM with RBF kernel $k(\mathbf{x},\mathbf{y}) = \exp(-\beta\|\mathbf{x}-\mathbf{y}\|^2)$. Here, we set $\beta = \beta_0$ where $1/\beta_0$ is the mean distance between points. For testing, we use another 200 "normal" points and 200 outliers outside the banana-shaped region. Table I shows the improvements on classification accuracies (averaged over 50 repetitions) when different amounts of covariance information are used. Next, we use four RBF kernels, with $\beta = 2\beta_0, \beta_0, \beta_0/2$ and $\beta_0/3$ respectively, as base kernels in (11). Figure 1 compares the resultant data descriptions and Table II shows the corresponding accuracies. As can be seen, the learned kernel can obtain a good data description and almost the best accuracy over the range of ν experimented.

Experiments are then performed on three real-world data sets (ionosphere, heart and sonar) from the UCI machine learning repository. For each data set, we treat each class as the "normal" data in separate experiments. We randomly choose 90% of points as training and the remaining 10% as testing, lumping the latter with the points of the opposite class. Results are averaged over 10 repetitions. Table III shows that the learned kernel is often competitive with the kernel having the "best" β, particularly on the sonar data set.

V. Conclusion

In this paper, we extended the one-class SVMs so that covariance information from the data can be utilized in a robust manner. Furthermore, by constraining the desired kernel function as a convex combination of some base kernels, we showed that the weighting coefficients can be obtained by solving a QCQP or SOCP problem. Results on both toy and real-world data sets show promising results. In the future, we will explore using other forms for the target kernel function.

Acknowledgment

This paper is supported by the Research Grants Council of the Hong Kong Special Administrative Region under grants 615005 and DAG03/04.EG28, the National Nature Science Foundation of China (No. 6040204) and the Program for New Century Excellent Talents in University.

References

[1] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[2] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, July 2001.

[3] D. Tax and R. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, no. 14, pp. 1191–1199, 1999.

[4] G. Lanckriet, L. El Ghaoui, and M. Jordan, "Robust novelty detection with single-class MPM," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.

[5] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal risk minimization," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, pp. 416–422.

[6] L. Manevitz and M. Yousef, "One-class SVMs for document classification," *Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2001.

[7] C. Ong, A. Smola, and R. Williamson, "Hyperkernels," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.

[8] O. Bousquet and D. Herrmann, "On the complexity of learning the kernel matrix," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.

[9] K. Crammer, J. Keshet, and Y. Singer, "Kernel design using boosting," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.

[10] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Advances in Neural Information Processing Systems 14*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002.

[11] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan, "Learning the kernel matrix with semi-definite programming," in *Proceedings of the Nineteenth International Conference on Machine Learning*, Sydney, Australia, 2002, pp. 323–330.

[12] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, 2nd ed. New York: Cambridge University Press, 1992.

[13] F. Alizadeh and D. Goldfarb, "Second-order cone programming," Rutgers Center for Operations Research, Rutgers University, Tech. Rep. RRR 51-2001, 2001.

[14] Y. Nesterov and A. Nemirovskii, *Interior-point Polynomial Algorithms in Convex Programming*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1994.

TABLE I

TEST SET CLASSIFICATION ACCURACIES ON THE TOY DATA WITH DIFFERENT AMOUNTS OF COVARIANCE INFORMATION ($\nu = 0.25$).

c/r in (8)	0	$1/n$	$10/n$	$100/n$	$1000/n$	$10000/n$
accuracy	86.07%	86.09%	86.11%	86.56%	91.29%	89.98%

TABLE II

TEST SET CLASSIFICATION ACCURACIES ON THE TOY DATA AT DIFFERENT ν'S.

ν	learned kernel	base kernels			
		$\beta = 2\beta_0$	$\beta = \beta_0$	$\beta = \beta_0/2$	$\beta = \beta_0/3$
0.1	90.50%	80.75%	78.25%	86.00%	77.50%
0.2	79.50%	78.25%	79.75%	77.75%	85.75%
0.3	84.75%	74.75%	81.50%	77.25%	78.00%
0.4	80.75%	75.00%	79.00%	78.50%	77.25%

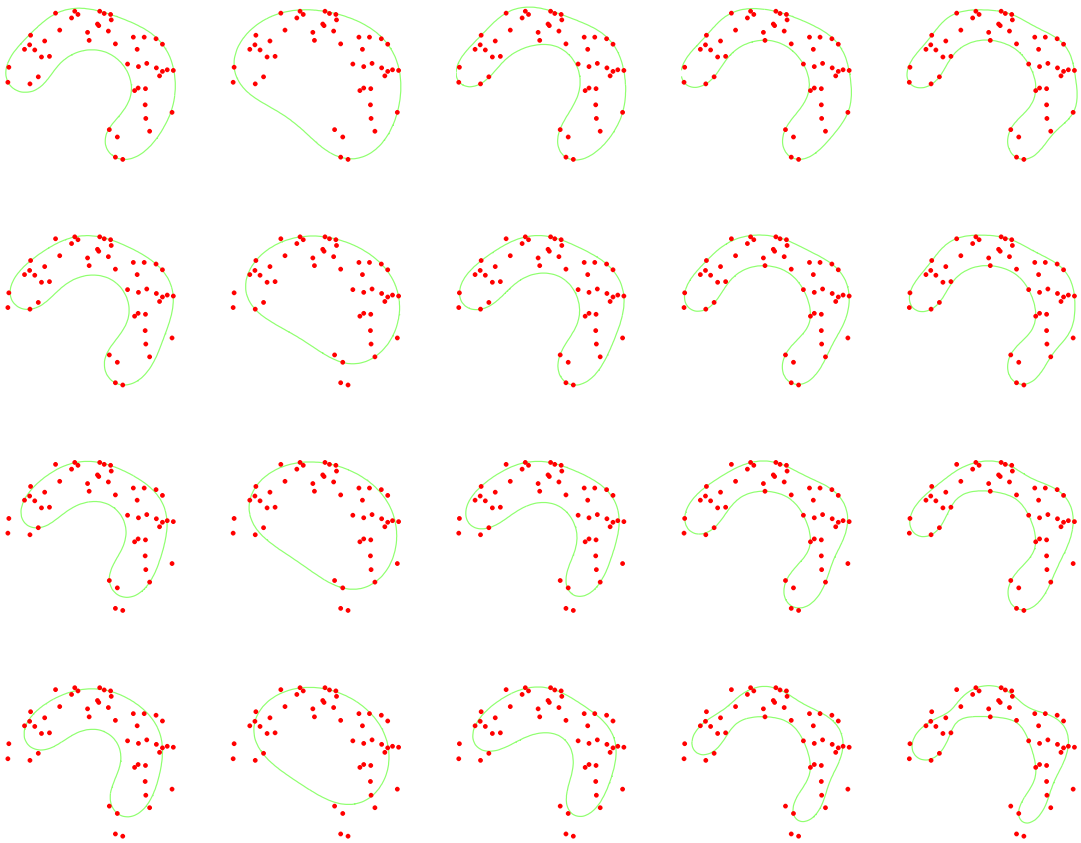

Fig. 1. Data descriptions of the toy data (Top to bottom: $\nu = 0.1, 0.2, 0.3, 0.4$. Left to right: learned kernel, base kernels with $\beta = 2\beta_0, \beta_0, \beta_0/2, \beta_0/3$).

TABLE III

TEST SET CLASSIFICATION ACCURACIES ON THE UCI DATA.

data set		learned kernel	base kernels			
			$\beta = 2\beta_0$	$\beta = \beta_0$	$\beta = \beta_0/2$	$\beta = \beta_0/3$
ionosphere	class +	66.05%	93.29%	22.27%	66.95%	21.43%
	class –	70.99%	28.61%	21.43%	53.49%	73.53%
heart	class +	69.96%	76.18%	28.61%	21.43%	72.10%
	class –	71.78%	55.74%	50.17%	21.43%	76.33%
sonar	class +	93.29%	46.93%	42.99%	40.60%	70.85%
	class –	90.49%	55.66%	42.99%	21.43%	68.25%