A Naive Solution to the One-Class Problem and Its Extension to Kernel Methods

Alberto Muñoz[1] and Javier M. Moguerza[2]

[1] University Carlos III, c/ Madrid 126, 28903 Getafe, Spain
alberto.munoz@uc3m.es
[2] University Rey Juan Carlos, c/ Tulipán s/n, 28933 Móstoles, Spain
javier.moguerza@urjc.es

Abstract. In this work, the problem of estimating high density regions from univariate or multivariate data samples is studied. To be more precise, we estimate minimum volume sets whose probability is specified in advance. This problem arises in outlier detection and cluster analysis, and is strongly related to One-Class Support Vector Machines (SVM). In this paper we propose a new simpler method to solve this problem. We show its properties and introduce a new class of kernels, relating the proposed method to One-Class SVMs.

1 Introduction

The task of estimating high density regions from data samples arises explicitly in a number of works involving interesting problems such as outlier detection or cluster analysis (see for instance [5,7] and references herein). One-Class Support Vector Machines (SVM) [10,12] are designed to solve this problem with tractable computational complexity. We refer to [10] and references therein for a complete description of the problem and its ramifications.

In the recent years papers showing failures in the estimations found by One-Class SVM have appeared [4,6]. In this work, a new algorithm to estimate high density regions from data samples is presented. The algorithm relaxes the density estimation problem in the following sense: instead of trying to estimate the density function at each data point, an easier to calculate data-based measure is introduced in order to establish a density ranking among the sample points.

The concrete problem to solve is the estimation of minimum volume sets of the form $S_\alpha(f) = \{x | f(x) \geq \alpha\}$, such that $P(S_\alpha(f)) = \nu$, where f is the density function and $0 < \nu < 1$. Throughout the paper, sufficient regularity conditions on f are assumed.

The rest of the paper is organized as follows. Section 2 introduces the method and its properties. In Section 3, a kernel formulation of the proposed algorithm is shown. Section 4 shows the numerical advantages of the new method over One-Class SVM. Section 5 concludes.

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 193–204, 2005.

2 The Naive One-Class Algorithm

There are data analysis problems where the knowledge of an accurate estimator of the density function $f(x)$ is sufficient to solve them, for instance, mode estimation [2], or the present task of estimating $S_\alpha(f)$. However, density estimation is far from trivial [11,10]. The next definition is introduced to relax the density estimation problem: the task of estimating the density function at each data point is replaced by a simpler measure that asymptotically preserves the order induced by f.

Definition 1. Neighbourhood Measures. *Consider a random variable X with density function $f(x)$ defined on $\mathbb{R}^d$. Let S_n denote the set of random independent identically distributed (iid) samples of size n (drawn from f). The elements of S_n take the form $s_n = (x_1, \cdots, x_n)$, where $x_i \in \mathbb{R}^d$. Let $M : \mathbb{R}^d \times S_n \longrightarrow \mathbb{R}$ be a real-valued function defined for all $n \in \mathbb{N}$. (a) If $f(x) < f(y)$ implies $\lim_{n\to\infty} P(M(x, s_n) > M(y, s_n)) = 1$, then M is a* ***sparsity measure****. (b) If $f(x) < f(y)$ implies $\lim_{n\to\infty} P(M(x, s_n) < M(y, s_n)) = 1$, then M is a* ***concentration measure****.*

Example 1. $M(x, s_n) \propto 1/\hat{f}(x, s_n)$, where $\hat{f}$ can be any consistent non-parametric density estimator, is a sparsity measure; while $M(x, s_n) \propto \hat{f}(x, s_n)$ is a concentration measure. A commonly used estimator is the kernel density one $\hat{f}(x, s_n) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\|x-x_i\|}{h})$.

Example 2. Consider the distance from a point x to its k^{th}-nearest neighbour in s_n, $x^{(k)}$: $M(x, s_n) = d_k(x, s_n) = d(x, x^{(k)})$: it is a sparsity measure. Note that d_k is neither a density estimator nor is it one-to-one related to a density estimator. Thus, the definition of 'sparsity measure' is not trivial. Another valid choice is given by the average distance over all the k nearest neighbours: $M(x, s_n) = \bar{d}_k = \frac{1}{k} \sum_{j=1}^{k} d_j = \frac{1}{k} \sum_{j=1}^{k} d(x, x^{(j)})$. Extensions to other centrality measures, such as trimmed-means are straightforward.

Our goal is to obtain some decision function $h(x)$ which solves the problem stated in the introduction, that is, $h(x) = +1$ if $x \in S_\alpha(f)$ and $h(x) = -1$ otherwise. We will show how to use sparsity measures to build $h(x)$.

Consider a sample $s_n = \{x_1, \ldots, x_n\}$. Consider the function $g(x) = M(x_i, s_n)$, where M is a sparsity measure. For the sake of simplicity we assume $g(x_i) \neq g(x_j)$ if $i \neq j$ (the complementary event has zero probability).

To solve the One-Class problem, the following algorithm is introduced:

Naive One-Class Algorithm

```
(1) Choose a constant $\nu \in [0, 1]$.
(2) Consider the order induced in $s_n$ by the sparsity measure $g(x)$, that is,
    $g(x_{\{1\}}) \leq g(x_{\{2\}}) \leq \ldots \leq g(x_{\{n\}})$, where $x_{\{i\}}$ denotes the $i^{th}$-sample.
(3) Consider the value $\rho^* = g(x_{\{\nu n\}})$ if $\nu n \in \mathbb{N}$, $\rho^* = g(x_{\{[\nu n]+1\}})$ otherwise,
    where $[x]$ stands for the largest integer not greater than $x$.
(4) Define $h(x) = sign(\rho^* - g(x))$
```

Note that the choice of the function $g(x)$ is not involved in the algorithm; it has to be determined in advance. The role of ρ^* and ν will become clear with the next proposition, which shows that the decision function $h(x) = sign(\rho^* - g(x))$ will be non-negative for at least a proportion equal to ν of the training s_n sample. Following [10], this result is called ν-property.

Proposition 1. ν-property. *The following two statements hold for the value ρ^*:*

1. $\frac{1}{n}\sum_{i=1}^{n} I(g(x_i) < \rho) \leq \nu \leq \frac{1}{n}\sum_{i=1}^{n} I(g(x_i) \leq \rho)$*, where I stands for the indicator function and $x_i \in s_n$.*
2. *With probability 1, asymptotically, the preceding inequalities become equalities.*

Proof. 1. Regarding the right-hand side of the inequality, $\frac{1}{n}\sum_{i=1}^{n} I(g(x_i) \leq \rho) = \frac{\nu n}{n} = \nu$ if $\nu n \in \mathbb{N}$ and equals $\frac{[\nu n]+1}{n} > \nu$ if $\nu n \notin \mathbb{N}$. For the left-hand side a similar argument applies. 2. Regarding the right-hand side inequality, if $\nu n \in \mathbb{N}$ the result is immediate from the preceding argument. If $\nu n \notin \mathbb{N}$, $\frac{[\nu n]+1}{n} \longrightarrow \nu$ as $n \longrightarrow \infty$. Again, for the left-hand side a similar argument applies. □

Remark 1. If $g(x)$ is chosen to be a concentration measure, then the decision function has to be defined as $h(x) = sign(g(x) - \rho^*)$.

Notice that in the naive algorithm ν represents the fraction of points inside the support of the distribution if $g(x)$ is a sparsity measure. If a concentration measure is used, ν represents the fraction of outlying points. The role of ρ^* becomes now clear: it represents the decision value which, induced by the sparsity measure, determines if a given point belongs to the support of the distribution. As the next theorem states an asymptotical result, we will denote every quantity depending on the sample s_n with the subscript n. Also we will suppose $\nu n \in \mathbb{N}$. The theorem goes one step further from the ν-property, showing that, asymptotically, the naive One-Class algorithm finds the desired α-level sets. In order to formulate the theorem, we need a measure to estimate the difference between two sets. We will use the d_μ-distance. Given two sets A and B

$$d_\mu(A, B) = \mu(A \Delta B)\,,$$

where μ is a measure on $\mathbb{R}^d$, Δ is the symmetric difference $A \Delta B = (A \cap B^c) \cup (B \cap A^c)$, and A^c denotes the complementary set of A.

Theorem 1. *Consider a measure μ absolutely continuous with respect to the Lebesgue measure. The set $R_n = \{x : h_n(x) = sign(\rho_n^* - g_n(x)) \geq 0\}$ d_μ-converges to a region of the form $S_\alpha(f) = \{x | f(x) \geq \alpha\}$, such that $P(S_\alpha(f)) = \nu$. Therefore, the naive One-Class method estimates a density contour cluster $S_\alpha(f)$ (which, in probability, includes the mode).*

Proof. For space reasons, we omit some mechanical steps. Consider the set $C_\nu = \{x_\nu \in \mathbb{R}^d : f(x_\nu) = \alpha\}$, where $\nu = P(S_\alpha(f))$. By Proposition 1, point 2, $\lim_{n\to\infty} P(g_n(x) < g_n(x_{\{\nu n\}})) = \nu$ (fact 1). Besides, it is easy to prove that given $C \subset S_{f(y)}(f)$ with $\mu(C) < \infty$, then $\mu(x \in C : g_n(x) < g_n(y))$ tends to $\mu(C)$. Thus $\lim_{n\to\infty} P(g_n(x) < g_n(x_\nu)) = \nu$, $\forall x_\nu \in C_\nu$ (fact 2). From facts 1 and 2, and using standard arguments from probability theory, it follows that $\forall \varepsilon > 0$, $\lim_{n\to\infty} P(\left|f(x_{\{\nu n\}}) - f(x_\nu)\right| > \varepsilon) = 0$, that is, $\lim_{n\to\infty} f(x_{\{\nu n\}}) = f(x_\nu)$ in probability.

Now consider $x \in S_\alpha(f)) \cap R_n^c$. From $f(x) > f(x_\nu)$ and Definition 1, it holds that $\lim_{n\to\infty} P(g_n(x) < g_n(x_\nu)) = 1$. Given that $\lim_{n\to\infty} f(x_{\{\nu n\}}) = f(x_\nu)$ in probability, it follows that $\lim_{n\to\infty} P(g_n(x) < g_n(x_{\{\nu n\}})) = 1$, that is, $P(h_n(x) < 0) \to 1$. Therefore, $\mu\left(S_\alpha(f)) \cap R_n^c\right) \to 0$.

Let now $x \in R_n \cap S_\alpha(f)^c$. From $f(x) < f(x_\nu)$, Definition 1 and $\lim_{n\to\infty} f(x_{\{\nu n\}}) = f(x_\nu)$ in probability, it holds that $P(g_n(x) \geq g_n(x_{\{\nu n\}})) \to 1$, that is, $P(h_n(x) > 0) \to 1$. Thus $\mu\left(R_n \cap S_\alpha(f)^c\right) \to 0$, which concludes the proof. □

We provide an estimate of a region $S_\alpha(f)$ with the property $P(S_\alpha(f)) = \nu$. Among regions S with the property $P(S) = \nu$, the region $S_\alpha(f)$ will have minimum volume as it has the form $S_\alpha(f) = \{x | f(x) \geq \alpha\}$. Therefore we provide an estimate that asymptotically, in probability, has minimum volume.

Finally, it is important to remark that the quality of the estimation procedure heavily depends on using a sparsity or a concentration measure (the particular choice is not – asymptotically – relevant). If the measure used is neither a concentration nor a sparsity measure, there is no reason why the method should work.

3 Kernel Formulation of the Naive Algorithm

In this section we will show the relation between the naive algorithm and One-Class SVM. In order to do so we have to define a class of neighbourhood measures.

Definition 2. Positive and Negative Neighbourhood Measures. *$MP(x, s_n)$ is said to be a* ***positive sparsity (concentration) measure*** *if $MP(x, s_n)$ is a sparsity (concentration) measure and $MP(x, s_n) \geq 0$. $MN(x, s_n)$ is said to be a* ***negative sparsity (concentration) measure*** *if $-MN(x, s_n)$ is a positive concentration (sparsity) measure.*

Given that negative neighbourhood measures are in one-to-one correspondence to positive neighbourhood measures, only positive neighbourhood measures need to be considered. The following classes of kernels can be defined using positive neighbourhood measures.

Definition 3. Neighbourhood Kernels. *Consider the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^+$ defined by $\phi(x) = MP(x, s_n)$, where $MP(x, s_n)$ is a positive neighbourhood measure. The function $K(x, y) = \phi(x)\phi(y)$ is called a* ***neighbourhood kernel****. If $MP(x, s_n)$ is a positive sparsity (concentration) measure, $K(x, y)$ is a* ***sparsity (concentration) kernel****.*

Note that the set $\{\phi(x_i)\}$ is trivially separable from the origin in the sense of [10], since each $\phi(x_i) \in \mathbb{R}^+$. Separability is guaranteed by Definition 2.

The strategy of One-Class support vector methods is to map the data points into a feature space determined by a kernel function, and to separate them from the origin with maximum margin (see [10] for details). In order to build a separating hyperplane between the origin and the points $\{\phi(x_i)\}$, the quadratic One-Class SVM method solves the following problem:

$$\begin{aligned} &\min_{w,\rho,\xi} \frac{1}{2}\|w\|^2 - \nu n\rho + \sum_{i=1}^{n} \xi_i \\ &\text{s.t.} \;\; \langle w, \phi(x_i)\rangle \geq \rho - \xi_i \,, \\ &\qquad \xi_i \geq 0, \qquad\qquad i = 1, \ldots, n \,, \end{aligned} \tag{1}$$

where ϕ is the mapping defining the kernel function, ξ_i are slack variables, $\nu \in [0, 1]$ is an a priori fixed constant, and ρ is a decision variable which determines if a given point belongs to the estimated high density region.

The next theorem illustrates the relation between our naive algorithm and One-Class SVMs when neighbourhood kernels are used.

Theorem 2. *Define the mapping $\phi(x) = MP(x, s_n)$. The decision function $h_V(x) = sign(\rho_V^* - w^*\phi(x))$ obtained from the solution ρ_V^* and w^* to the One-Class SVM problem (1) using the sparsity kernel $K(x, y) = \phi(x)\phi(y)$ coincides with the decision function $h(x)$ obtained by the naive algorithm.*

Proof. Consider the dual problem of (1):

$$\begin{aligned} &\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j K(x_i, x_j) \\ &\text{s.t.} \;\; \sum_{i=1}^{n} \alpha_i = \nu n \,, \\ &\qquad 0 \leq \alpha_i \leq 1 \,, \qquad\qquad i = 1, \ldots, n \,, \end{aligned} \tag{2}$$

where $x_i \in s_n$. For the sake of simplicity we assume $\phi(x_i) \neq \phi(x_j)$ if $i \neq j$ (the complementary event has zero probability) and that $\nu n \in \mathbb{N}$ (the proof for $\nu n \notin \mathbb{N}$ can be derived with similar arguments to those in the proof of Proposition 1). Consider the order induced in s_n by the mapping $\phi(x)$ and denote $x_{\{i\}}$ the i^{th}-sample and $\alpha_{\{i\}}$ the corresponding dual variable. Therefore $\phi(x_{\{1\}}) \leq \phi(x_{\{2\}}) \leq \ldots \leq \phi(x_{\{n\}})$. Since $K(x_i, x_j) = \phi(x_i)\phi(x_j)$ and, by Definition 2, $\phi(x_i) \in \mathbb{R}^+$, the maximum of the objective function of problem (2) will be attained for $\alpha_{\{i\}} = 1$, $i \in \{1, \ldots, \nu n\}$ and $\alpha_j = 0$ otherwise. At the solution, the objective function takes the value $-\frac{1}{2}\sum_{i=1}^{\nu n}\sum_{j=1}^{\nu n} K(x_{\{i\}}, x_{\{j\}})$. By the weak theorem of duality, the value of the objective function of problem (1) has to be equal or greater than the value of the objective function of problem (2) at the solution. Consider the solution $w^* = \sum_{i=1}^{\nu n} \phi(x_{\{i\}})$, $\rho_V^* = w^*\phi(x_{\{\nu n\}})$, $\xi_{\{i\}} = w^*\left[\phi(x_{\{\nu n\}}) - \phi(x_{\{i\}})\right]$ for $i \in \{1, \ldots, [\nu n]\}$. For the remaining indexes $\xi_j = 0$.

At this point the solution to problem (1) coincides with the solution to problem (2), that is, the duality gap is zero. The decision function takes the form $h_V(x) = sign(w^* \left[\phi(x_{\{\nu n\}}) - \phi(x)\right])$ which coincides with the decision function of the naive algorithm (the scalar $w^* > 0$ does not affect the sign). So the theorem holds. □

It remains open to show if the decision function obtained from One-Class SVM algorithms within the framework in [10,8] can be stated in terms of positive sparsity or concentration measures. The next remark provides the answer.

Remark 2. The exponential kernel $K_c(x, y) = e^{-\|x-y\|^2/c}$ is neither a sparsity kernel nor a concentration kernel. For instance, consider a univariate bimodal density f with finite modes m_1 and m_2 such that $f(m_1) = f(m_2)$. Consider any positive sparsity measure $MP(x, s_n)$ and the induced mapping $\phi(x) = MP(x, s_n)$. As $n \to \infty$, the sparsity kernel $K(x, y) = \phi(x)\phi(y)$ would attain its minimum at (m_1, m_2) (or at two points in the sample s_n near to the modes). On the other hand, as the exponential kernel $K_c(x, y)$ depends exclusively on the distance between x and y, any pair of points (a, b) whose distance is larger than $\|m_1 - m_2\|$ will provide a value $K_c(a, b) < K_c(m_1, m_2)$, which asymptotically can not happen for kernels induced by positive sparsity measures. In this case, the neighbourhood kernel has four minima while the exponential kernel has the whole diagonal as minima. The reasoning for concentration kernels is analogous. A similar argument applies for polinomial kernels with even degrees (odd degrees induce mapped data sets that are non separable from the origin, which discards them).

Note that, while the naive algorithm works with every neighbourhood measure, the separability condition of the mapped data is necessary when One-Class SVM are being used, restricting the use of neighbourhood measures to positive or negative ones. This restriction and the fact that our method provides a simpler approach make the use of the naive algorithm advisable when neighbourhood measures are being used.

4 Experiments

In this section we compare the performance of One-Class SVM and the naive algorithm for a variety of artificial and real data sets. Systematic comparisons of the two methods as data dimension increases are carried out. First of all we describe the implementation details concerning both algorithms.

With regards to One-Class SVM we adopt the proposal in [10], that is, the exponential kernel $K_c(x, y) = e^{-\|x-y\|^2/c}$ is used. This is the only kernel used for experimentation in [10], and it is also the only (non neighbourhood) kernel for which a clear relation to density estimation has been demonstrated (see [6]). To perform the experiments, a range of values for c has been chosen, following the widely used rule $c = hd$ (see [9,10]), where d is the data dimension and $h \in \{0.1, 0.2, 0.5, 0.8, 1.0\}$.

Concerning the naive algorithm, three different sparsity measures have been considered:

- $M_1(x, s_n) = d_k = d(x, x^{(k)})$, the distance from a point x to its k^{th}-nearest neighbour $x^{(k)}$ in the sample s_n. The only parameter in M_1 is k, which takes a finite number of values (in the set $\{1, \cdots, n\}$). We have chosen k to cover a representative range of values, namely, k will equal the 10%, 20%, 30%, 40% and 50% sample proportions. Therefore we choose k as the closest integer to hn, where n is the sample size and $h \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.
- $M_2(x, s_n) = \dfrac{1}{\sum_{i=1}^{n} \exp\left(-\frac{\|x - x_i\|^2}{2\sigma}\right)}$, where $\sigma \in \mathbb{R}^+$. The only parameter in M_2 is σ. We want σ to be related to the sample variability and, at the same time, to scale well with respect to the data sample distances. We choose $\sigma = hs$, where $s = \max d_{ij}^2/\varepsilon$, $h \in \{0.1, 0.2, 0.5, 0.8, 1.0\}$, $d_{ij}^2 = \|x_i - x_j\|^2$ and ε is a small value which preserves scalability in M_2. For all the experiments we have chosen $\varepsilon = 10^{-8}$.
- $M_3(x, s_n) = \log\left(\dfrac{1}{\sum_{i=1}^{n} \frac{1}{\|x - x_i\|^p}}\right)$, where $p \in \mathbb{R}^+$. Parameter p in M_3 is related to data dimension [3]. We choose $p = hd$, where d is the data dimension and $h \in \{0.01, 0.02, 0.05, 0.08, 0.1\}$. In this case the values of h are smaller for smoothing reasons (see [3] for details).

Measure M_1 has been described in Example 2 in Section 2. Measures M_2 and M_3 are of the type described in Examples 1 and 4 in the same section. M_2 uses as density estimator the Parzen window [11], while M_3 is based on the Hilbert kernel density estimator [3] and could take negative values. Note that Theorem 1 guarantees that asymptotically every sparsity measure (and in particular the three chosen here) will lead to sets containing the true mode.

4.1 Artificial Data Sets

An Asymmetric Distribution. In the first experiment we have generated 2000 points from a gamma $\Gamma(\alpha, \beta)$ distribution, with $\alpha = 1.5$ and $\beta = 3$. Figure 1 shows the histogram, the gamma density curve, the true mode $(\alpha - 1)/\beta$ as a bold vertical line, the naive algorithm estimations with sparsity measure M_1 (five upper lines) and the One-Class SVM (five lower lines) estimations of the 50% highest density region. The parameters have been chosen as described at the beginning of Section 4, and lines are drawn for each method in increasing order in the h parameter, starting from the bottom. Being our goal to detect the shortest region of the form $S_\alpha(f) = \{x : f(x) > \alpha\}$ (that must contain the mode), it is apparent that the naive regions improve upon the One-Class SVM regions. All the naive regions regions contain the true mode and are connected. All the One-Class SVM regions are wider and show a strong bias towards less dense zones. Furthermore, only in two cases the true mode is included in the estimated SVM regions, but in these cases the intervals obtained are not simply connected. The naive algorithm using measures M_2 and M_3 provide similar intervals to those obtained using measure M_1, and are not shown for space reasons.

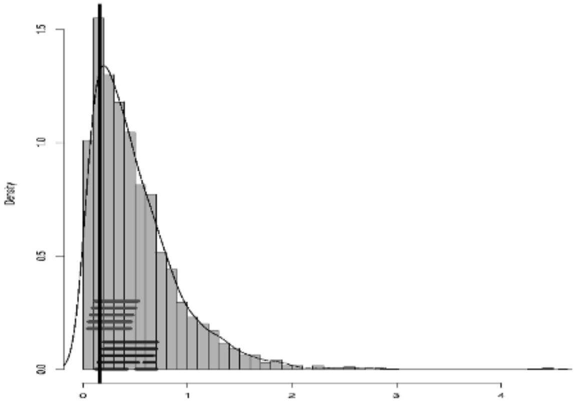

Fig. 1. Gamma sample with 2000 points. The figure shows the histogram, the density curve, a vertical line at the true mode, the naive estimations with sparsity measure M_1 (five upper lines) and One-Class SVM (five lower lines) estimations of the 50% highest density region.

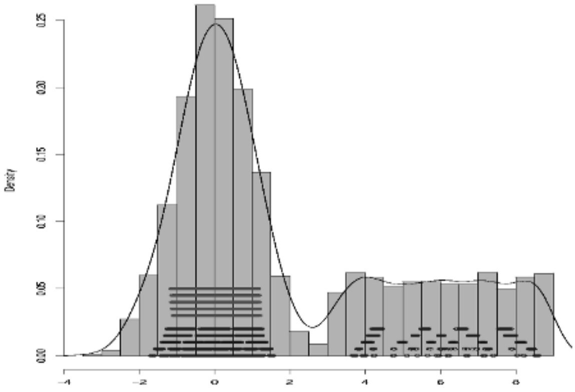

Fig. 2. Mixture sample with 3000 points. The figure shows the histogram, the estimated density curve, the naive estimations with sparsity measure M_1 (five upper lines) and One-Class SVM (five lower lines) estimations of the 50% highest density region.

A Mixture of Distributions. This second experiment considers a mixture of a normal $N(0,1)$ and a uniform $U(6,9)$ distribution. Figure 2 shows the histogram, the estimated density curve, the naive estimations with sparsity measure M_1 (five upper lines) and the One-Class SVM (five lower lines) estimations of the 50% highest density region. Again, the parameters have been chosen as described at the beginning of Section 4, and lines are drawn for each method in increasing order in the h parameter, starting from the bottom. Once more, the naive method using measures M_2 and M_3 provide similar intervals to those obtained using measure M_1, and are not shown for space reasons. Regarding the quality of the results, note that the 50% densest region corresponds to points from the normal distribution. All the naive estimations (upper lines) match the correct region, while the One-Class SVM (lower lines) spreads part of the points in the uniform zone. However, all points in the uniform zone have lower density than those found by the naive procedure.

Increasing the Data Dimension. In this experiment we want to evaluate whether the performance of the Naive method and One-Class SVM algorithms degrades as the data dimension increases. To this end, we have generated 20 data sets with increasing dimension from 2 to 200. Each data set contains 2000 points from a multivariate normal distribution $N(0, I_d)$, where I_d is the identity matrix in $\mathbb{R}^d$. Detailed results are not shown for space reasons. We will only show the conclusions. Since the data distribution is known, we can retrieve the true outliers, that is, the true points outside the support corresponding to any percentage specified in advance. For each dimension and each method, we have determined, from the points retrieved as outliers, the proportion of true ones.

As the data dimension increases, the performance of One-Class SVM degrades: it tends to retrieve as outliers an increasing number of points. The best results for One-Class SVM are obtained for the largest magnitudes of the parameter c (only when convergence for the optimization problem within was achieved).

Regarding the naive method, robustness with regard to the parameter choice is observed. Dimension barely affects the performance of our method, and results are consistently better than those obtained with One-Class SVM. For instance, for a percentage of outliers equal to 1%, the best result for One-Class SVM is 15%, against 100% using our method (for all the sparsity measures considered). For a percentage of outliers equal to 5%, the best result for One-Class SVM is 68%, against 99% using the naive method.

4.2 A Practical Example: Outlier Detection in Handwritten Digit Recognition

The database used next contains nearly 4000 instances of handwritten digits from Alpaydin and Kaynak [1]. Each digit is represented by a vector in $\mathbb{R}^{64}$ constructed from a 32×32 bitmap image. The calligraphy of the digits in the database seems to be easily perceivable, which is supported by the high success rate of various classifiers. In particular, for each digit, nearest neighbour classifiers accuracy is always over 97% [1].

In the present case there is a nice interpretation for points outside the sets $S_\alpha(f)$ (the support of the data set, see Section 1). The outlying points should correspond to 'badly'-written characters. In order to check out this behaviour, 10 apparent outliers (shown in Figure 3) have been added to the database. We will consider the whole database as a sample from a multivariate distribution, and we will verify if the proposed algorithm is able to detect this outlying instances. Note that there is an added difficulty in this case, namely, the underlying distribution is multimodal in a high dimensional environment.

Figure 4 shows the outliers obtained by the naive algorithm when the support for the 99.5% percentile is calculated. Using this support percentage exactly 20

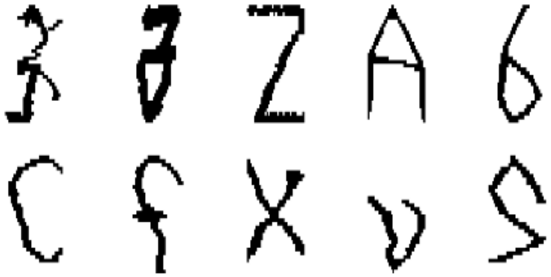

Fig. 3. Ten (apparent) outliers added to the original Alpaydin & Kaynak handwritten digits database

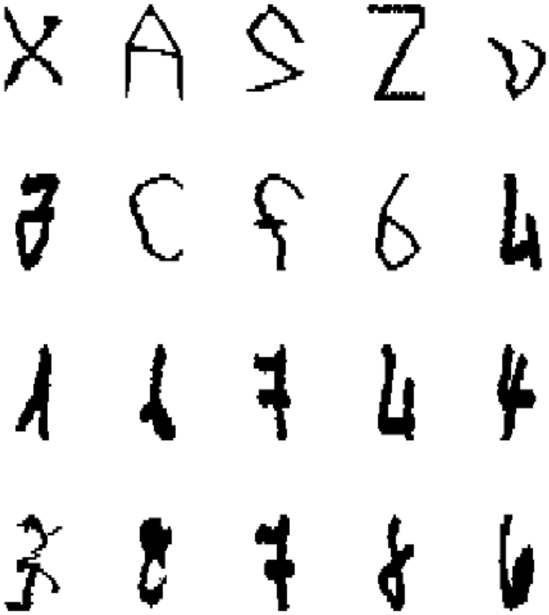

Fig. 4. The outlying digits found by the naive algorithm, ordered left–right and up–down using the sparsity measure $M(x, s_n) = d(x, x^{(k)})$

outliers are to be retrieved. We expect to detect the 10 outliers we have included, and we are curious about the aspect of the 10 other most outlying digits in the database. In Figure 4, the digits retrieved as outliers by the naive method using the sparsity measure M_1 are shown in decreasing order (left-right and up-down). Here $k = 1$, using $k = n^{4/(d+4)}$, where d is the space dimension. This value is known to be proportional to the (asymptotically) optimal value [11] for density estimation tasks. Nine of the ten added outliers are detected as the most outlying points. The remaining eleven outliers include the other added instance (similar to a ‘3’), and ten more figures whose calligraphy seems to be different from representative digits within each class. Similar results are obtained for sparsity measures M_2 and M_3.

Using a One-Class SVM with exponential kernel (trying a wide range of values for the c parameter, including those proposed in [9,10]) none of the ten added outliers was detected.

5 Conclusions

In this paper a new method to estimate minimum volume sets of the form $S_\alpha(f) = \{x|f(x) \geq \alpha\}$, has been proposed. Our proposal introduces the use of neighbourhood measures. These measures asymptotically preserve the order induced by the density function f. In this way we avoid the complexity of solving a pure density estimation problem. Regarding computational results, the naive method performs consistently better than One-Class SVM in all the tested problems (the ones shown here and many others omitted for space reasons). The advantage that the naive method has over the One-Class SVM is due to Theorem 1 which guarantees that it asymptotically finds the desired α-level sets. The suboptimal performance of One-Class SVM arises from the fact that its decision function is not based on sparsity or concentration measures and that there are no results of the nature of Theorem 1 for One-Class SVM. In particular, we have shown that the neither the exponential kernel nor polynomial kernels come from neighbourhood measures (and therefore Theorem 1 does not hold for these kernels).

Acknowledgments. This work was partially supported by spanish grants SEJ2004-03303, 06/HSE/0181/2004, TIC2003-05982-C05-05 (MCyT) and PPR-2004-05 (URJC).

References

1. E. Alpaydin and C. Kaynak. *Cascading Classifiers.* Kybernetika, 34(4):369-374, 1998.
2. L. Devroye. *Recursive estimation of the mode of a multivariate density.* The Canadian Journal of Statistics, 7(2):159-167, 1979.
3. L. Devroye and A. Krzyzak. *On the Hilbert kernel density estimate.* Statistics and Probability Letters, 44:299-308, 1999.
4. J.M. Moguerza and A. Muñoz. *Solving the One-Class Problem using Neighbourhood Measures.* LNCS 3138:680-688, Springer, 2004.
5. J.M. Moguerza, A. Muñoz and M. Martin-Merino. *Detecting the Number of Clusters Using a Support Vector Machine Approach.* LNCS 2415:763-768, Springer, 2002.
6. A. Muñoz and J.M. Moguerza. *One-Class Support Vector Machines and density estimation: the precise relation.* LNCS 3287:216-223, Springer, 2004.
7. A. Muñoz and J. Muruzabal. *Self-Organizing Maps for Outlier Detection.* Neurocomputing, 18:33-60, 1998.

8. G. Rätsch, S. Mika, B. Schölkopf and K.R. Müller. *Constructing Boosting Algorithms from SVMs: an Application to One-Class Classification.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(9):1184-1199, 2002.
9. B. Schölkopf, C. Burges and V. Vapnik. *Extracting Support Data for a given Task.* Proc. of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1995.
10. B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola and R.C. Williamson. *Estimating the Support of a High Dimensional Distribution.* Neural Computation, 13(7):1443-1471, 2001.
11. B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1990.
12. D.M.J. Tax and R.P.W. Duin. *Support Vector Domain Description.* Pattern Recognition Letters, 20:1991-1999, 1999.