SV Estimation of a Distribution's Support

Bernhard Schölkopf[§,⋆], Robert C. Williamson[⋆], Alex Smola[§,⋆], John Shawe-Taylor[†]
[§] GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
[⋆] Department of Engineering, Australian National University, Canberra 0200, Australia
[†] Royal Holloway College, University of London, Egham, UK
bs, smola@first.gmd.de, Bob.Williamson@anu.edu.au, john@dcs.rhbnc.ac.uk

Abstract

Suppose you are given some dataset drawn from an underlying probability distribution P and you want to estimate a subset S of input space such that the probability that a test point drawn from P lies outside of S is bounded by some a priori specified $0 < \nu \leq 1$.

We propose an algorithm which approaches this problem by trying to estimate a function f which is positive on S and negative on the complement. The functional form of f is given by a kernel expansion in terms of a potentially small subset of the training data; it is regularized by controlling the length of the weight vector in an associated feature space.

The algorithm is a natural extension of the support vector algorithm to the case of unlabelled data.

1 INTRODUCTION

During recent years, a new set of kernel techniques for supervised learning has been developed [28, 20]. Specifically, support vector (SV) algorithms for pattern recognition, regression estimation and solution of inverse problems have received considerable attention.

There have been a few attempts to transfer the idea of using kernels to compute inner products in feature spaces to the domain of unsupervised learning. The problems in that domain are, however, less precisely specified. Generally, they can be characterized as estimating *functions* of the data which tell you something interesting about the underlying distributions. For instance, kernel PCA can be characterized as computing functions which on the training data produce unit variance outputs while having minimum norm in feature space [21]. Another kernel-based unsupervised learning technique, regularized principal manifolds [25], computes functions which give a mapping onto a lower-dimensional manifold minimizing a regularized quantization error. Clustering algorithms are further examples of unsupervised learning techniques which can be kernelized [21]. An extreme point of view is that unsupervised learning is about estimating densities. Clearly, knowledge of the density of P would then allow us to solve whatever problem can be solved on the basis of the data.

The present paper addresses an easier problem: it proposes an algorithm which computes a binary function which is supposed to capture regions in input space where the probability density lives (its support), i.e. a function such that most of the data will live in the region where the function is nonzero. In doing so, it is in line with Vapnik's principle never to solve a problem which is more general than the one we actually need to solve. Moreover, it is applicable also in cases where the density of the data's distribution is not well-defined, e.g. if there are singular components.

2 PREVIOUS WORK

Part of the motivation for the present work was the paper of Ben-David and Lindenbaum [1]. It turns out that there is a considerable amount of prior work in the statistical literature, and in this section we briefly summarise that. We do not attempt a detailed comparison of the proof techniques of the specific results achieved, but confine ourselves to placing the previous work in context.

In order to summarize the methods, it is convenient to introduce the following definition of a (multi-dimensional) quantile function (introduced by Einmahl and Mason [8]). Let $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ be i.i.d. random variables in $\mathcal{X}$ with distribution F. Let $\mathcal{C}$ be a class of measurable subsets of $\mathcal{X}$ and let λ be a real-valued function defined on $\mathcal{C}$. The *quantile function* with respect to $(F, \lambda, \mathcal{C})$ is

$$U(\alpha) = \inf\{\lambda(C) \colon F(C) \geq \alpha, C \in \mathcal{C}\} \quad 0 < \alpha \leq 1$$

If F_ℓ is the empirical distribution ($F_\ell(C) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathbf{1}_C(\mathbf{x}_i)$), the *empirical quantile function* is

$$U_\ell(\alpha) = \inf\{\lambda(C) \colon F_\ell(C) \geq \alpha, C \in \mathcal{C}\} \quad 0 < \alpha \leq 1.$$

We denote by $C(\alpha)$ and $C_\ell(\alpha)$ the (not necessarily unique) $C \in \mathcal{C}$ that attains the infimum (when it is achievable). The most common choice of λ is Lebesgue measure, in which case $C(\alpha)$ is the minimum volume $C \in \mathcal{C}$ that contains at least a fraction α of the probability mass. We will assume λ is Lebesgue measure from here on. Estimators of the form $C_\ell(\alpha)$ are called *minimum volume estimators*.

Estimating the support of a density. Observe that $C(1)$ is the *support* of the density f corresponding to F, assuming it exists. (Note that $C(1)$ is well defined even when f does not exist.) The problem of estimating $C(1)$ appears to have first been studied by Geffroy [10] who considered $\mathcal{X} = \mathbb{R}^2$ with piecewise constant estimators. There have been a number of works studying a natural nonparametric estimator of $C(1)$ (e.g. [3, 7, 5]; see [9] for further references). The nonparametric estimator is simply

$$\hat{C}_\ell = \bigcup_{i=1}^{\ell} B(\mathbf{x}_i, \epsilon_n) \tag{1}$$

where $B(\mathbf{x}, \epsilon)$ is the $l_2(\mathcal{X})$ ball of radius ϵ centered at $\mathbf{x}$ and $(\epsilon_n)_n$ is an appropriately chosen decreasing sequence. Devroye and Wise [7] showed the asymptotic consistency of (1) with respect to the symmetric difference between $C(1)$ and $\hat{C}_\ell$. Cuevas [5] did the same, but for Hausdorff distance. Cuevas and Fraiman [6] studied the asymptotic consistency of a *plug-in* estimator of $C(1)$: $\hat{C}^{\mathrm{p}} = \{\hat{f}_\ell > 0\}$ where $\hat{f}_\ell$ is a kernel density estimator. In order to avoid problems with $\hat{C}^{\mathrm{p}}$ they actually analyzed $\hat{C}^{\mathrm{p}}_\beta := \{\hat{f}_\ell > \beta_\ell\}$ where $(\beta_\ell)_\ell$ is an appropriately chosen sequence. Clearly for a given distribution, α is related to β, but this connection can not be readily exploited by this type of estimator.

The most recent work relating to the estimation of $C(1)$ is by Gayraud [9] who has made an asymptotic minimax study of estimators of *functionals* of $C(1)$. Two examples are $\operatorname{vol} C(1)$ or the center of $C(1)$. (See also [12, Chapter 8].)

Estimating high probability regions ($\alpha \neq 1$). Turning to the case where $\alpha < 1$, it seems the first work was reported by Sager [18]and then Hartigan [11] who considered $\mathcal{X} = \mathbb{R}^2$ with $\mathcal{C}$ being the class of closed convex sets in $\mathcal{X}$.(They actually consider density contour clusters; see below for a definition.) Nolan [14] considered higher dimensions with $\mathcal{C}$ being the class of ellipsoids.

Tsybakov [27] has studied an estimator based on piecewise polynomial approximation of $C(\alpha)$ and has shown it attains the asymptotically minimax rate for certain classes of densities f.

Polonik [17] has studied the estimation of $C(\alpha)$ by $C_\ell(\alpha)$. He derived asymptotic rates of convergence in terms of various measures of richness of $\mathcal{C}$. He considers both VC classes and classes with a log ϵ-covering number with bracketing of order $O(\epsilon^{-r})$ for $r > 0$. He also summarizes a number of other previous works on minimum volume estimators which we have not mentioned here.

Polonik [16] has also studied the use of the "excess mass approach" [13] to construct an estimator of "generalized α-clusters" which are related to $C(\alpha)$.

Define the *excess mass over* $\mathcal{C}$ *at level* α as

$$E_{\mathcal{C}}(\alpha) = \sup\{H_\alpha(C) \colon C \in \mathcal{C}\}$$

where $H_\alpha(\cdot) = (F - \alpha\lambda)(\cdot)$ and again λ denotes Lebesgue measure. Any set $\Gamma_{\mathcal{C}}(\alpha) \in \mathcal{C}$ such that

$$E_{\mathcal{C}}(\alpha) = H_\alpha(\Gamma_{\mathcal{C}}(\alpha))$$

is called a *generalized α-cluster in* $\mathcal{C}$. Replace F by F_ℓ in these definitions to obtain their empirical counterparts $E_{\ell,\mathcal{C}}(\alpha)$ and $\Gamma_{\ell,\mathcal{C}}(\alpha)$. In other words, his estimator is

$$\Gamma_{\ell,\mathcal{C}}(\alpha) = \arg\max\{(F_\ell - \alpha\lambda)(C) : C \in \mathcal{C}\}$$

where the $\max$ is not necessarily unique. Now whilst $\Gamma_{\ell,\mathcal{C}}(\alpha)$ is clearly different to $C_\ell(\alpha)$, it is related to it in that it attempts to find small regions with as much excess mass (which is similar to finding small regions with a given amount of probability mass). Actually $\Gamma_{\ell,\mathcal{C}}(\alpha)$ is more closely related to the determination of *density* contour clusters at level α:

$$c_f(\alpha) := \{x \colon f(x) \geq \alpha\}.$$

Simultaneously, and independently, Ben-David and Lindenbaum [1] studied the problem of estimating $c_f(\alpha)$. They too made use of VC classes but stated their results in a stronger form which is meaningful for finite sample sizes.

Finally we point out a curious connection between minimum volume sets of a distribution and its differential entropy in the case that $\mathcal{X}$ is one dimensional. Suppose X is a one dimensional random variable with density f. Let $S = C(0)$ be the support of f and define the *differential entropy* of X by

$$h(X) = -\int_S f(x) \log f(x) dx.$$

For $\epsilon > 0$ and $n \in \mathbb{N}$, define the *typical set* $A_\epsilon^{(n)}$ with respect to f by

$$A_\epsilon^{(n)} = \{(x_1, \ldots, x_n) \in S^n \colon | -\tfrac{1}{n} \log f(x_1, \ldots, x_n) - h(X)| \leq \epsilon\},$$

where $f(x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i)$.

If $(a_n)_n$ and $(b_n)_n$ are sequences, the notation $a_n \doteq b_n$ means $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Cover and Thomas [4, p.227] show that for all $\epsilon, \delta < \frac{1}{2}$, then

$$\operatorname{vol} A_\epsilon^{(n)} \doteq \operatorname{vol} C^n(1-\delta) \doteq 2^{nh}.$$

They point out that this result "indicates that the volume of the smallest set that contains most of the probability is approximately 2^{nh}. This is a n-dimensional volume, so the corresponding side length is $(2^{nh})^{1/n} = 2^h$. This provides an interpretation of differential entropy."

Applications. A number of applications have been suggested for these techniques. They include problems in marketing [1], condition monitoring of machines [7], Econometrics and generalized nonlinear principal curves [27, 12], regression and spectral analysis [17], tests for multimodality and clustering [16] and others [13].

Polonik [15] has shown how one can use estimators of $C(\alpha)$ to construct density estimators. The point of doing this is that it allows one to encode a range of prior assumptions about the true density f that would be impossible to do within the traditional density estimation framework. He has shown asymptotic consistency and rates of convergence for densities belonging to VC-classes or with a known rate of growth of metric entropy with bracketing.

Relationship with the Present Work. The present paper describes an algorithm which finds regions close to $C(\alpha)$. We do not try and find *the* minimum volume such region. On the other hand, our algorithm has tractable computational complexity, even in several variables. Our theory, which uses very similar tools to those used by Polonik, gives results that are of more use in a finite sample size setting.

3 ALGORITHMS

We first introduce terminology and notation conventions. We consider training data

$$\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{X}, \tag{2}$$

where $\ell \in \mathbb{N}$ is the number of observations, and $\mathcal{X}$ is some set. In this paper, we assume it to be a compact subset of $\mathbb{R}^N$. Let Φ be a feature map $\mathcal{X} \to F$, i.e. a map into a dot product space F such that the dot product in the image of Φ can be computed by evaluating some simple kernel [28, 20]

$$k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})), \tag{3}$$

such as the Gaussian kernel

$$k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/c}. \tag{4}$$

Indices i and j are understood to range over $1, \ldots, \ell$ (in compact notation: $i, j \in [\ell]$). Bold face greek letters denote ℓ-dimensional vectors comprising the corresponding letters typeset in normal face.

To separate the data set from the origin, we solve the following quadratic program:

$$\begin{aligned} \min_{\mathbf{w}\in F, \boldsymbol{\xi}\in\mathbb{R}^\ell, \rho\in\mathbb{R}} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + \tfrac{1}{\ell}\textstyle\sum_i \xi_i - \nu\rho \\ \text{subject to} \quad & (\mathbf{w} \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \;\; \xi_i \geq 0. \end{aligned} \tag{5}$$

Here, $\nu \in (0, 1]$ is a parameter whose meaning will become clear later.

Since nonzero slack variables ξ_i are penalized in the objective function, we can expect that if $\mathbf{w}$ and ρ solve this problem, then the decision function

$$f(\mathbf{x}) = \operatorname{sgn}((\mathbf{w} \cdot \Phi(\mathbf{x})) - \rho) \tag{6}$$

will be positive for most examples $\mathbf{x}_i$ contained in the training set, while the SV type regularization term $\|\mathbf{w}\|$ will still be small. The actual trade-off between these two goals is controlled by ν.

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$L(\mathbf{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\ell}\sum_i \xi_i - \nu\rho - \sum_i \alpha_i((\mathbf{w} \cdot \Phi(\mathbf{x}_i)) - \rho + \xi_i) - \sum_i \beta_i \xi_i, \tag{7}$$

and set the derivatives by the primal variables $\mathbf{w}, \boldsymbol{\xi}, \rho$ equal to zero, yielding

$$\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i), \tag{8}$$

$$\alpha_i = \frac{1}{\ell} - \beta_i \leq \frac{1}{\ell}, \quad \sum_i \alpha_i = \nu. \tag{9}$$

In (8), all patterns with $\alpha_i > 0$ are called Support Vectors. Together with (3), the SV expansion transforms the decision function (6) into a kernel expansion

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho\right). \tag{10}$$

Substituting (8) - (9) into L (7), and using (3), we obtain the dual problem:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \quad \text{subject to } 0 \leq \alpha_i \leq \frac{1}{\ell}, \quad \sum_i \alpha_i = \nu. \tag{11}$$

One can show that at the optimum, the two inequality constraints (5) become equalities if α_i and β_i are nonzero, i.e. if $0 < \alpha_i < 1/\ell$. Therefore, we can recover ρ by exploiting that for any such α_i, the corresponding pattern $\mathbf{x}_i$ satisfies

$$\rho = (\mathbf{w} \cdot \mathbf{x}_i) = \sum_j \alpha_j k(\mathbf{x}_j, \mathbf{x}_i). \tag{12}$$

The quadratic program (11) can be solved with standard QP tools. Alternatively, one can used a kernel version of the iterative algorithm proposed by [30] for the linear case.

To conclude this section, we note that one can also use *balls* to describe the data in feature space, close in spirit to algorithms used in [19, 26]. Again, we try to put "most of" the data into a small ball by solving, for $\nu \in (0, 1]$,

$$\begin{aligned} \min_{R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell, c \in F} \quad & \nu \cdot R^2 + \tfrac{1}{\ell} \textstyle\sum_i \xi_i \\ \text{subject to} \quad & \|\Phi(\mathbf{x}_i) - c\|^2 \leq R^2 + \xi_i, \;\; \xi_i \geq 0 \quad \text{for } i \in [\ell]. \end{aligned} \tag{13}$$

This leads to the dual

$$\min_{\boldsymbol{\alpha}} \quad \textstyle\sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) - \nu \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) \tag{14}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \tfrac{1}{\ell}, \;\; \textstyle\sum_i \alpha_i = \nu. \tag{15}$$

For kernels $k(\mathbf{x}, \mathbf{y})$ which only depend on $\mathbf{x} - \mathbf{y}$, $k(\mathbf{x}, \mathbf{x})$ is constant. In this case, the equality constraint implies that the linear term in the dual target function is constant, and the problem (14–15) turns out to be equivalent to (11).

4 THEORY

In this section, we analyse the algorithm theoretically, starting with the uniqueness of the hyperplane (Proposition 2). We then describe the connection to pattern recognition (Proposition 3), and show that the parameter ν characterizes the fractions of SVs and outliers (Proposition 4). Following that, we give a robustness result for the soft margin (Proposition 5) and finally we briefly state error bounds (Theorem 9).

Definition 1 *A data set*

$$\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_\ell) \tag{16}$$

is called separable if there exists some $\mathbf{w} \in F$ *such that* $(\mathbf{w} \cdot \Phi(\mathbf{x}_i)) > 0$ *for* $i \in [\ell]$.

Proposition 2 *If the data set (16) is separable, then there exists a unique* supporting hyperplane *with the properties that (1) it is closer to the origin than all data, and (2)* *its distance to the origin is maximal among all such hyperplanes. For any $\rho > 0$, it is given by*

$$\min_{\mathbf{w} \in F} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } (\mathbf{w} \cdot \Phi(\mathbf{x}_i)) \geq \rho. \tag{17}$$

Proof Due to the separability, the convex hull of the data does not contain the origin. The existence and uniqueness of the hyperplane then follows from the supporting hyperplane theorem [e.g. 2].

Moreover, separability implies that there actually exists some $\rho > 0$ and $\mathbf{w} \in F$ such that $(\mathbf{w} \cdot \Phi(\mathbf{x}_i)) \geq \rho$ for $i \in [\ell]$ (by rescaling $\mathbf{w}$, this can be seen to work for arbitrarily large ρ). By Cauchy-Schwartz, the distance of the hyperplane $\{\mathbf{z} \in F : (\mathbf{w} \cdot \mathbf{z}) = \rho\}$ to the origin is $\rho/\|\mathbf{w}\|$. Therefore the optimal hyperplane is obtained by minimizing $\|\mathbf{w}\|$ subject to these constraints, i.e. by the solution of (17). ■

The following result elucidates the relationship between single-class classification and binary classification.

Proposition 3 *(i) Suppose $(\mathbf{w}, \rho)$ parametrizes the supporting hyperplane for the data (16). Then $(\mathbf{w}, 0)$ parametrizes the optimal separating hyperplane (passing through the origin [28]) for the labelled data set $\{(\Phi(\mathbf{x}_1), 1), \ldots, (\Phi(\mathbf{x}_\ell), 1), (-\Phi(\mathbf{x}_1), -1), \ldots, (-\Phi(\mathbf{x}_\ell), -1)\}$.*
(ii) Suppose $(\mathbf{w}, 0)$ parametrizes the optimal separating hyperplane passing through the origin for a labelled data set $\{(\Phi(\mathbf{x}_1), y_1), \ldots, (\Phi(\mathbf{x}_\ell), y_\ell)\}$, $(y_i \in \{\pm 1\}$ for $i \in [\ell])$. Suppose, moreover, that $\mathbf{w}$ is aligned such that $(\mathbf{w} \cdot \Phi(\mathbf{x}_i))$ is positive whenever $y_i = 1$, and that $\rho/\|\mathbf{w}\|$ is the margin of the optimal hyperplane. Then $(\mathbf{w}, \rho)$ parametrizes the supporting hyperplane for the unlabelled data set $\{y_1\Phi(\mathbf{x}_1), \ldots, y_\ell\Phi(\mathbf{x}_\ell)\}$.

Proof *Ad (i).* Observe that $(-\mathbf{w}, \rho)$ parametrizes the supporting hyperplane for the data set reflected through the origin, and that it is parallel to the one given by $(\mathbf{w}, \rho)$. This provides an optimal separation of the two sets, with distance 2ρ, and a separating hyperplane $(\mathbf{w}, 0)$.
Ad (ii). By assumption, we have $y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i)) \geq \rho$ [cf. 28], hence $(\mathbf{w} \cdot y_i\Phi(\mathbf{x}_i)) \geq \rho$ for $i \in [\ell]$ ■

Note that this relationship holds true also if we consider nonseparable problems. In that case, *margin errors* in binary classification (i.e. points which are either on the wrong side of the separating hyperplane or which fall inside the margin) translate into *outliers* in single-class classification, i.e. into points which fall on the wrong side of the hyperplane. Proposition 3 then holds, cum grano salis, for the training sets with margin errors and outliers, respectively, removed.

The utility of Proposition 3 lies in the fact that it allows us to recycle certain results proven for binary classification [22] for use in the single-class scenario. The following, explaining the significance of the parameter ν, is such a case.

Proposition 4 *Assume the solution of (5) satisfies $\rho \neq 0$. The following statements hold:*
(i) ν is an upper bound on the fraction of outliers.

(ii) ν is a lower bound on the fraction of SVs.
(iii) Suppose the data (16) were generated independently from a distribution $P(\mathbf{x})$ which does not contain discrete components. Suppose, moreover, that the kernel is analytic and non-constant. With probability 1, asymptotically, ν equals both the fraction of SVs and the fraction of outliers.

Parts (i) and (ii) follow directly from Proposition 3 and the fact that outliers are dealt with in exactly the same way as margin errors in the optimization problem for the binary classification case [22]. (A direct proof would simply exploit the constraints that (9) imposes on the fraction of patterns that can have $\alpha_i = 1/\ell$, upper bounding the fraction of outliers, and on the fraction of patterns that must have $\alpha_i > 0$, the SVs.) Part (iii) can be proven by a uniform convergence argument showing that since the covering numbers of kernel expansions regularized by a norm in some feature space are well-behaved, the fraction of points which lie exactly on the hyperplane is asymptotically negligible [cf. 22].

Proposition 5 (Resistance) *Local movements of outliers parallel to $\mathbf{w}$ do not change the hyperplane.*

Proof Suppose $\mathbf{z}_o := \Phi(\mathbf{x}_o)$ is an outlier, i.e. $\xi_o > 0$, hence by the KKT conditions [e.g. 2] $\alpha_o = 1/\ell$. Transforming it into $\mathbf{z}'_o := \mathbf{z}_o + \delta \cdot \mathbf{w}$, where $|\delta| < \xi_o/\|\mathbf{w}\|$, leads to a slack which is still nonzero, i.e. $\xi'_o > 0$, hence we still have $\alpha_o = 1/\ell$. Therefore, $\boldsymbol{\alpha}' = \boldsymbol{\alpha}$ is still feasible, as is the primal solution $(\mathbf{w}', \boldsymbol{\xi}', \rho')$. Here, we use $\xi'_i = \xi_i$ for $i \neq o$, $\mathbf{w}' = \mathbf{w} + \delta \cdot \alpha_o \mathbf{w}$, and ρ' as computed from (12). Finally, the KKT conditions are still satisfied, as still $\alpha'_o = 1/\ell$. Thus [2, e.g.], $\boldsymbol{\alpha}$ is still the optimal solution. ∎

We now move on to the subject of generalization. Our goal is to bound the probability that a novel point drawn from the same underlying distribution lies outside of the estimated region by a certain margin. We start by introducing a common tool for measuring the capacity of a class $\mathcal{F}$ of functions that map $\mathcal{X}$ to $\mathbb{R}$.

Definition 6 *Let (X, d) be a pseudo-metric space, let A be a subset of X and $\epsilon > 0$. A set $B \subseteq X$ is an* ϵ-cover *for A if, for every $a \in A$, there exists $b \in B$ such that $d(a, b) \leq \epsilon$. The* ϵ-covering number *of A, $\mathcal{N}_d(\epsilon, A)$, is the minimal cardinality of an ϵ-cover for A (if there is no such finite cover then it is defined to be ∞). The cover is called proper if $B \subseteq A$.*

Note that we have used less than or equal to in the definition of a cover. This is somewhat unconventional, but will not change the bounds we use. It does, however, prove technically useful in the proofs. The idea is that B should be finite but approximate all of A with respect to the pseudometric d. We will use the l^∞ distance over a finite sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_\ell)$ for the pseudo-metric in the space of functions,

$$d_{\mathbf{X}}(f, g) = \max_{i \in [\ell]} |f(\mathbf{x}_i) - g(\mathbf{x}_i)|. \tag{18}$$

Let $\mathcal{N}(\epsilon, \mathcal{F}, \ell) = \max_{\mathbf{X} \in \mathcal{X}^\ell} \mathcal{N}_{d_{\mathbf{X}}}(\epsilon, \mathcal{F})$. Below, logarithms are to base 2.

Theorem 7 *Consider any distribution P on $\mathcal{X}$ and any $\theta \in \mathbb{R}$. Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ are generated i.i.d. from P. Then with probability $1 - \delta$ over such an ℓ-sample, if we find $f \in \mathcal{F}$ such that $f(\mathbf{x}_i) \geq \theta + \gamma$ for all $i \in [\ell]$*

$$P\{\mathbf{x} : f(\mathbf{x}) < \theta - \gamma\} \leq \tfrac{2}{\ell}(k + \log \tfrac{2\ell}{\delta}),$$

where $k = \lceil \log \mathcal{N}(\gamma, \mathcal{F}, 2\ell) \rceil$.

The basis of the proof is [23, Lemma 3.9].

We now consider the possibility that for a small number of points $f(\mathbf{x}_i)$ fails to exceed $\theta + \gamma$. This corresponds to having a non-zero slack variable ξ_i in the algorithm, where we take $\theta + \gamma = \rho/\|\mathbf{w}\|$ and use the class of linear functions in feature space in the application of the theorem. There are well-known bounds for the log covering numbers of this class. We first introduce notation for the size of the shortfall in $f(\mathbf{x})$.

Definition 8 *Let f be a real valued function on a space $\mathcal{X}$. Fix $\theta \in \mathbb{R}$. For $\mathbf{x} \in \mathcal{X}$ we denote by $d(\mathbf{x}, f, \gamma)$ the quantity $d(\mathbf{x}, f, \gamma) = \max\{0, \theta + \gamma - f(\mathbf{x})\}$. Similarly for a training sequence $\mathbf{X}$, we define*

$$D(\mathbf{X}, f, \gamma) = \sqrt{\sum_{\mathbf{x} \in \mathbf{X}} d(\mathbf{x}, f, \gamma)^2}.$$

Theorem 9 *Fix $\theta \in \mathbb{R}$. Consider a fixed but unknown probability distribution P on the input space $\mathcal{X}$ and a class of real valued functions $\mathcal{F}$ with range $[a, b]$. Then with probability $1 - \delta$ over randomly drawn training sequences $\mathbf{X}$ of size ℓ, for all $\gamma > 0$ and any $f \in \mathcal{F}$,*

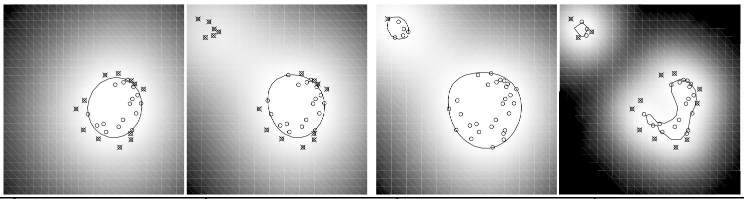
$$P\{\mathbf{x}\colon f(\mathbf{x}) < \theta - \gamma \text{ and } \mathbf{x} \notin \mathbf{X}\} \leq \tfrac{2}{\ell}(k + \log \tfrac{4\ell}{\delta}),$$

where $k = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2\ell) + \frac{64 D(\mathbf{X}, f, \gamma)^2}{\gamma^2} \log\left(\frac{e\ell(b-a)\gamma}{8 D(\mathbf{X}, f, \gamma)^2}\right) \log\left(\frac{32\ell(b-a)^2}{\gamma^2}\right) \right\rceil$.

The proof is based on similar proofs for the classification case in [24, Theorem 3]. The theorem bounds the probability of a new training point falling in the region for which $f(\mathbf{x})$ has value less than $\theta - \gamma$, this being the complement of the estimate for the support of the distribution. In the algorithm described in this paper, one would use the hyperplane shifted by 2γ towards the origin to define the region. Note that there is no restriction placed on the class of functions though these functions could be probability density functions.

The choice of γ gives a trade-off between the size of the region over which the bound holds (increasing γ increases the size of the region) and the size of the probability ϵ (increasing γ decreases the size of the log covering numbers).

The result shows that we can bound the probability of points falling outside the region of estimated support by a quantity involving the ratio of the log covering numbers (which can be bounded by the fat shattering dimension at scale proportional to γ) and the number of training examples, plus a factor involving the 2-norm of the slack variables. A similar bound can also be made in terms of the 1-norm of the slack variables, which is the form which appears in the optimization performed by the algorithm.



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|\mathbf{w}\|$	0.84	0.70	0.62	0.48

Figure 1: *First two pictures:* A single-class SVM applied to two toy problems; $\nu = c = 0.5$, domain: $[-1, 1]^2$. Note how in both cases, at least a fraction of ν of all examples is in the estimated region (cf. table). The large value of ν causes the additional data points in the upper left corner to have almost no influence on the decision function. For smaller values of ν, such as 0.1 *(third picture)*, the points cannot be ignored anymore. Alternatively, one can force the algorithm to take these 'outliers' into account by changing the kernel width (4): in the *fourth picture*, using $c = 0.1, \nu = 0.5$, the data is effectively analyzed on a different length scale which leads the algorithm to consider the outliers as meaningful points.

The result is significantly stronger than related results given in Ben-David and Lindenbaum [1], since their bound involves the square root of the ratio of the Pollard dimension (the fat shattering dimension when γ tends to 0) and the number of training examples.

5 EXPERIMENTS

We briefly illustrate the approach on artificial and real-world data. Figure 1 displays 2-D toy examples, and shows how the parameter settings influence the solution. Figure 2 shows a plot of the outputs $(\mathbf{w} \cdot \Phi(\mathbf{x}))$ on training and test set of the US postal service database of handwritten digits. The database contains 9298 digit images of size 16×16, 2007 of which are used as a test set. We fed our algorithm, using a Gaussian kernel of width 0.5 (a typical value for SVMs, cf. [19]), with the training instances of digit 0 only. Testing was done on both digit 0 and on all other digits. As shown in figure 2, $\nu = 0.5$ leads to *zero* false positives (i.e. even though the learning machines has not seen any non-0s during training, it correctly identifies all non-0s as such), while still recognizing 44% of the digits 0 in the test set. Higher recognition rates can be achieved using smaller values of ν: for 0.05, we get 91% correct recognition of digits 0 in the test set, while the false positive rate is still fairly moderate at 7%. While these first results are necessarily incomplete, they still convey an idea of the potential utility of the proposed algorithm.

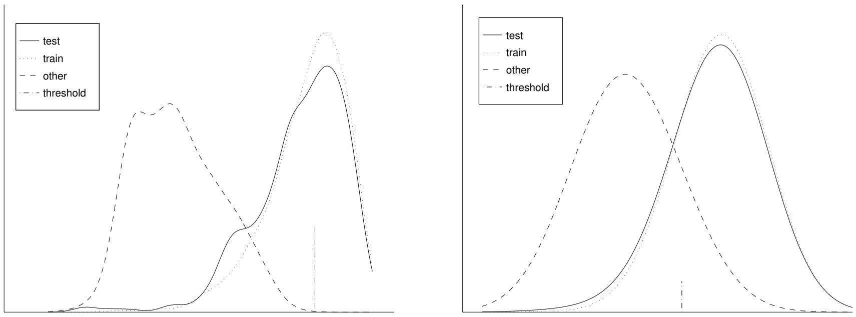


Figure 2: Experiments on the US postal service OCR dataset. Recognizer for digit 0; output histogram (Parzen windows density estimate)for the exemplars of 0 in the training/test set, and on test exemplars of other digits. For $\nu = 50\%$ *(left)*, we get 50% SVs and 49% outliers (consistent with Proposition 4); for $\nu = 5\%$ *(right)*, we get 6% and 4%, respectively. The threshold ρ is marked in the graphs.

6 DISCUSSION

The main inspiration for our approach stems from the earliest work of Vapnik and collaborators [30, 29]. In the sixties, they proposed an algorithm for characterizing a set of unlabelled data points by separating it from the origin using a hyperplane. However, they quickly moved on to two-class classification problems, both in terms of algorithms and in terms of the theoretical development of statistical learning theory which originated in those days.

From an algorithmic point of view, we can identify two weaknesses of the original approach which may have caused research in this direction to stop for more than three decades. Firstly, the original algorithm in [29] was limited to linear decision rules in input space, secondly, there was no way of dealing with outliers. In conjunction, these restrictions are indeed severe — a generic dataset need not be separable from the origin by a hyperplane.

The two modifications that we have incorporated dispose of these shortcomings. First, the kernel trick allows for a much larger class of functions by nonlinearly mapping into a high-dimensional feature space, and thereby increases the chances of a separation from the origin being possible. In particular, using a Gaussian kernel (4), such a separation always exists: to see this, note that $k > 0$, thus all dot products between mapped patterns are positive, implying that all patterns lie inside the same orthant. Moreover, since $k(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x}$, they all have unit length. Hence they are separable. The second modification directly allows for the possibility of outliers. We have incorporated this 'softness' of the decision rule using the ν-trick [22] and thus obtained a direct handle on the fraction of outliers.

We believe that our approach, proposing a concrete algorithm with well-behaved computational complexity (convex quadratic programming) for a problem that so far has mainly been studied from a theoretical point of view [e.g. 17, 4], has abundant practical applications, such as condition monitoring [7], marketing [1], or information retrieval. To turn the algorithm into an easy-to-use black-box method for practicioners, questions like the selection of kernel parameters (such as the width of a Gaussian kernel) have to be tackled. It is our expectation that the theoretical results which we have only briefly outlined in this paper will provide a solid foundation for this formidable task.

xxx open issues: choice of the origin

Additional possible future work includes the use of parametric noise models and the formulation of linear programming variants.

Acknowledgement. This work was supported by the ARC and the DFG (# Ja 379/9-1). Thanks to C. Schnörr and N. Oliver for helpful discussions.

References

[1] S. Ben-David and M. Lindenbaum. Learning distributions by their density levels: A paradigm for learning without a teacher. *Journal of Computer and System Sciences*, 55:171–182, 1997.

[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

[3] J. Chevalier. Estimation du support et du contour du support d'une loi de probabilité. *Annales de l'Institut Henri Poincaré. Section B. Calcul des Probabilités et Statistique. Nouvelle Série*, 12(4):339–364, 1976.

[4] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[5] A. Cuevas. On pattern analysis in the non-convex case. *Kybernetes*, 19(6):26–33, 1990.

[6] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6):2300–2312, 1997.

[7] L. Devroye and G.L. Wise. Detection of abnormal behaviour via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.

[8] J.H.J. Einmal and D.M. Mason. Generalized quantile processes. *The Annals of Statistics*, 20(2):1062–1078, 1992.

[9] G. Gayraud. Estimation of functional of density support. *Mathematical Methods of Statistics*, 6(1):26–46, 1997.

[10] J. Geffroy. Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, 13:191–210, 1964.

[11] J.A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82:267–270, 1987.

[12] A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction*. Springer, New York, 1993.

[13] D.W. Müller. The excess mass approach in statistics. Beiträge zur Statistik, Universität Heidelberg, 1992.

[14] D. Nolan. The excess mass ellipsoid. *Journal of Multivariate Analysis*, 39:348–371, 1991.

[15] W. Polonik. Density estimation under qualitative assumptions in higher dimensions. *Journal of Multivariate Analysis*, 55(1):61–81, 1995.

[16] W. Polonik. Measuring mass concentrations and estimating density contour clusters — an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995.

[17] W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69:1–24, 1997.

[18] Thomas W. Sager. An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association*, 74(366):329–339, 1977.

[19] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings, First International Conference on Knowledge Discovery & Data Mining*. AAAI Press, Menlo Park, CA, 1995.

[20] B. Schölkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999.

[21] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999. 327 – 352.

[22] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *To appear in: Neural Computation*, 1999. Also: NeuroColt2-TR 1998-031.

[23] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44(5):1926–1940, 1998.

[24] J. Shawe-Taylor and N. Cristianini. Margin distribution bounds on generalization. In Paul Fischer and Hans Ulrich Simon, editors, *Computational Learning Theory, 4th European Conference, EuroCOLT'99*. Springer, 1999.

[25] A. Smola, R. C. Williamson, S. Mika, and B. Schölkopf. Regularized principal manifolds. In *Computational Learning Theory: 4th European Conference*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 214 – 229. Springer, 1999.

[26] D.M.J. Tax and R.P.W. Duin. Data domain description by support vectors. In M. Verleysen, editor, *Proceedings ESANN*, pages 251 – 256, Brussels, 1999. D Facto.

[27] A.B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

[28] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[29] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).

[30] V. Vapnik and A. Lerner. Pattern recognition using generalized portraits. *Avtomatika i Telemekhanika*, 24:774 – 780, 1963.