ESTIMATION OF MINIMUM MEASURE SETS IN REPRODUCING KERNEL HILBERT SPACES AND APPLICATIONS.

Manuel Davy, Frederic Desobry, Stephane Canu

ESTIMATION OF MINIMUM MEASURE SETS IN REPRODUCING KERNEL HILBERT SPACES AND APPLICATIONS.

Manuel Davy
LAGIS/CNRS
Ecole Centrale Lille (France)
manuel.davy@ec-lille.fr

Frédéric Desobry[*]
Dept. of Engineering
Cambridge Uni. (UK).
fd238@eng.cam.ac.uk

Stéphane Canu
Laboratoire PSI
INSA Rouen (France).
scanu@insa-rouen.fr

ABSTRACT

Minimum measure sets (MMSs) summarize the information of a (single-class) dataset. In many situations, they can be preferred to estimated probability density functions (pdfs): they are strongly related to pdf level sets while being much easier to estimate in large dimensions.

The main contribution of this paper is a theoretical connection between MMSs and one class Support Vector Machines. This justifies the use of one-class SVMs in the following applications: novelty detection (we give explicit convergence rate) and change detection.

1. INTRODUCTION

Signal/Image processing decision algorithms often rely on the estimation of probability density functions (pdfs). Typical examples are in speech recognition, signal classification, or pattern (image) recognition. This may be performed by nonparametric techniques, such as Parzen windows, or by semi parametric techniques, such as mixtures of Gaussian. All these approaches, however, suffer from the curse of dimensionality problem, that is, the estimation becomes harder when the dimension of the space the pdf is defined on increases.

In most Signal/Image processing applications, however, solutions can be found without estimating a pdf. A typical example is that of data classification, where recent algorithms such as support vector machines (SVM) [1] are constructed without estimating densities. Another example is that of kernel change detection [2], where abrupt changes are detected without estimating a pdf as an intermediate step. In this paper, we propose an overview of an alternate method which can be used instead of pdf estimation in many problems. This approach proposes to use instead of the pdf a minimum measure set of this pdf. As will be shown below, minimum measure set (MMS) estimation is much easier than pdf estimation (especially in high dimension), and captures enough information about the data to enable accurate decisions.

Briefly, given a training set of vectors $\{x_1, \ldots, x_n\}$ in a space $\mathcal{X}$, MMS estimation consists of finding a subset C of $\mathcal{X}$ such that 1) C has minimum "volume" under some measure and 2) assuming the x_i's are distributed according to some probability measure P, and given some $\lambda \in [0; 1]$, the subset C verifies $P(C) = \lambda$ (see Section 2 for a rigorous definition).

This problem has been addressed in many ways in previous works. The most prominent strategies include that of Devroye and Wise [3], one class support vector machines [4], excess mass approaches and probability density functions (pdf) plug-in approaches. In this paper, we propose a unified view, where the MMS is sought in a class of subsets whose boundaries belong to a kernel space (*e.g.*, a Reproducing Kernel Hilbert Space – RKHS). More precisely, the boundary of such a subset C is $\{x \in \mathcal{X} | f(x) = 0\}$ where $f(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ belongs to some function space with kernel $k(\cdot, \cdot)$. The main contributions are i) a formal connection between one-class SVM and excess mass approaches (Section 3); ii) the derivation of a convergence rate for the probability of false alarms in one-class SVM novelty detection[1] (Section 5). Section 2 below recalls some fundamentals about MMSs, Section 4 points out some convergence study for the Lebesgue measure of the symmetric difference, and Section 6 proposes conclusions and future work directions.

2. MINIMUM MEASURE SETS

Let P a probability measure and Q a measure over the measurable space $\left(\mathcal{X}, \mathcal{B}(\mathcal{X})\right)$ such that Q dominates P. Let $\mathcal{C} \subset \mathcal{B}(\mathcal{X})$ a collection of Q-measurable subsets of $\mathcal{X}$. We assume that Q is known; P is supposedly unknown but a learning set $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} P$ is available. Let $\lambda \in [0; 1]$; we define the minimum Q-measure set of P in $\mathcal{C}$ as the set $C(\lambda)$ such that:

$$\begin{cases} P(C(\lambda)) = \lambda \\ Q(C(\lambda)) = \arg \inf_{C \in \mathcal{C}} \{Q(C); P(C) = \lambda\} \end{cases} \quad (1)$$

In the following, we assume that P admits a density p with respect to the measure Q[2]. Then, there exists $p_\lambda \in [0; \sup_{x \in \mathcal{X}} |p(x)|]$ such that MMS in Eq. (1) can also be defined as:

$$C(\lambda) = \{x \in \mathcal{X} : p(x) \geq p_\lambda\} \quad (2)$$

where p_λ depends on both λ and Q. Eq. (1) and Eq. (2) are equivalent definitions of $C(\lambda)$, which implies that one can use either the parametrization that uses either λ or p_λ. However, many reasons make the definition in Eq. (1) more suitable. In particular: 1) as the learning set size n tends to infinity, λ is asymptotically the ratio of learning samples that actually *fall* in $C(\lambda)$; 2) Any decision problem involving the comparison of two or more MMSs requires that they are defined for some fixed λ and not for fixed p_λ (the comparison does not have any sense otherwise); 3) λ has a direct interpretation in terms of distribution quantiles. This is needed when dealing with

[*]Corresponding author, acknowledges support from the Boeing company under the University of Cambridge grant RG40562.

[1]A similar rate can be obtained for one-class SVM based change detection.

[2]This assumption is aimed at making the presentation clearer, but it is not formally needed for most of the material presented in this paper to be true.

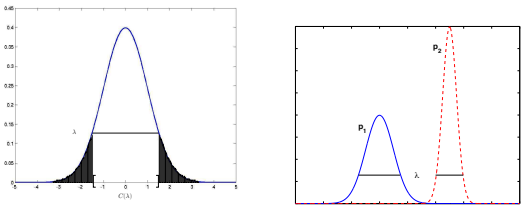

Fig. 1: When considering only one pdf, defining the MMS with the density level p_λ or its P-measure λ is equivalent, and $(1-\lambda)$ is the P-measure of the tails of the pdf (left). However, the comparison of two pdfs p_1 and p_2 has to be made for constant λ and not constant p_λ (right).

applications such as outliers detection or change detection. In the following, though, we use the definition in Eq. (2) because it enables clear and direct algorithms derivations. We will show that, *in fine*, the definition actually used in algorithms is that of Eq (1). In practice, the probability measure P is unknown, and an estimate, denoted $\widehat{C}_n(\lambda)$, is learned from $\{x_1, \ldots, x_n\}$. Here, we consider a nonparametric estimate and it has to: 1) be (strongly) consistent; 2) achieve relatively fast rates of convergence; 3) lead to a practicable and computationally cheap algorithm. Moreover, as many applications of MMSs involve decision, we also require that similarities between MMSs estimates can be calculated in a computationally tractable way, though $\mathcal{X}$ may have large dimension. These requirements are somehow hard to meet jointly. Methods found in the statistics literature either fail at providing a computationally tractable, practicable algorithm, at having good rates of convergence or at being suited to high dimensional data. In the next Section, we show that a RKHS together with excess mass methods make such good solutions possible.

3. A (REPRODUCING) KERNEL-BASED NONPARAMETRIC ESTIMATION PROCEDURE

The main result in [5] is that a fast rate of convergence for estimating MMSs can be achieved if $\mathcal{C}$ is a *poor* class of sets, such as VC or Glivenko-Cantelli. However, no practicable algorithm is provided. On the other hand, so-called *RKHS methods* [6, 7] provide a mean of exploring efficiently such classes of sets, through a representer theorem. We first focus on the *excess mass* approach, then we embed it into a RKHS.

3.1. Excess mass

The set $C(\lambda)$ is the set in $\mathcal{C}$ that maximizes the *excess mass* $m(C)$ of a set C , defined as:

$$m(C) \;=\; P(C) - p_\lambda Q(C) \;=\; \int_C (p(x) - p_\lambda) dQ(x) \qquad (3)$$

Indeed, for this set, the part of the integration set C in Eq. (3) for which $p(x) - p_\lambda$ is negative is reduced to its minimum, thus ensuring that $p(x) \geq p_\lambda$ for $x \in C(\lambda)$, see Eq. (2). For practical estimation, we define an empirical counterpart $m_n(C)$ to the true $m(C)$ and let:

$$\widehat{C}_n(\lambda) = \arg\max_{C \in \mathcal{C}} m_n(C) \qquad (4)$$

The simplest empirical counterpart for $m(C)$ surely is (where $1_{x_i \in C} = 1$ whenever $x_i \in C$ and $1_{x_i \in C} = 0$ otherwise and the term $\frac{1}{n}\sum_{i=1}^n 1_{x_i \in C}$ is called the *empirical measure*):

$$m_n(C) = \frac{1}{n}\sum_{i=1}^n 1_{x_i \in C} - p_\lambda Q_n(C) \qquad (5)$$

Finding the MMS $C(\lambda)$ comes down to solving one of the following equivalent problems:

$$\begin{array}{rcl} \max\limits_{C \in \mathcal{C}} m_n(C) & \Leftrightarrow & \min\limits_{C \in \mathcal{C}} \left(p_\lambda Q_n(C) - \frac{1}{n}\sum_{i=1}^n 1_{\{x_i \in C\}} \right) \\ & \Leftrightarrow & \min\limits_{C \in \mathcal{C}} \left(p_\lambda Q_n(C) + \frac{1}{n}\sum_{i=1}^n 1_{\{x_i \notin C\}} \right) \end{array} \qquad (6)$$

This problem is, however, ill-posed and cannot be solved easily in practice. The following section shows that RKHS approaches help solve the problem.

3.2. The kernel approach and connection with 1-class Support Vector Machine

The problem in Eq. (6) can be solved using kernels. Such methods have already proven useful in many machine learning problems, mostly because they yield linear interpretation of nonlinear problems, and because they enable easy evaluation of functions complexity (via the induced norm). Let $\mathcal{H}$ be a RKHS with reproducing kernel[3] $k(\cdot,\cdot)$. We now define $\mathcal{C}$ as the collection of sets $\{x \in \mathcal{X}; f(x) \geq \rho\}$ for $f \in \mathcal{H}$ and $\rho \geq 0$, and we use the shorthand $\{f \geq \rho\}$ to denote the set $\{x \in \mathcal{X}$ such that $f(x) \geq \rho\}$. In practice, we need to estimate f and ρ from the learning set $\{x_1, \ldots, x_n\}$ and we denote such estimates $\widehat{f}_n$ and $\widehat{\rho}_n$. In addition to propose a choice for $\mathcal{C}$, we also modify the criterion in Eq. (6) by changing $1_{\{x_i \notin C\}}$ into $(\rho_n - f_n(x_i))\, 1_{\{f_n(x_i) < \rho_n\}}$(this is the standard hinge loss which is used in, *e.g.*, SVMs). Similar smoothing also arises in kernel density estimation (see, *e.g.*, [8, Chapter 9]). The modified excess mass problem writes:

$$\begin{array}{l} \max\limits_{C \in \mathcal{C}} m_n(C) \Leftrightarrow \min\limits_{f_n \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n (\rho_n - f_n(x_i))\, 1_{\{f_n(x_i) < \rho_n\}} \\ + p_\lambda Q_n(\{f_n \geq \rho_n\}) \end{array} \qquad (7)$$

The r.h.s. of Eq. (7) appears to be a classical regularization criterion, where the term $\frac{1}{n}\sum_{i=1}^n (\rho_n - f_n(x_i))\, 1_{\{f_n(x_i) < \rho_n\}}$ is a hinge loss and the term $p_\lambda Q_n(\{f_n \geq \rho_n\})$ is a regularizer, independent of the learning set. The regularization parameter is the density level p_λ. We now state our main result.

Proposition 3.1 ***(Choice of the measure Q).*** *For any RKHS $\mathcal{H}$, there exists a measure Q such that minimizing the induced norm $\|f_n\|^2_{\mathcal{H}}$ in the RKHS $\mathcal{H}$ implies minimizing the regularizer in* (7), i.e. *1-class SVMs implement an excess mass approach.*

Proof (sketch): Let $T : L^2(\mathcal{X}) \to \mathcal{H} = \mathrm{Im(S)} \subset \mathbb{R}^{\mathcal{X}}$, $g \mapsto f = Tg$ and $\mathcal{H}$ be dense in $L^2(\mathcal{H})$. We have (see, *e.g.*, [7]): $\langle f, f\rangle_H = \langle Tg, Tg\rangle_H = \langle g, g\rangle_{L^2(x)}$. Let $1_{\mathcal{H}}$ denote the function in $\mathcal{H}$ which is the closest to the unit constant function, in the sense of the $L^2(\mathcal{X})$ norm. Then,

$$q(\cdot) = (T^{-1} 1_{\mathcal{H}})^2(\cdot) \qquad (8)$$

is such that the optimum of 1-class SVM criterion is the minimizer of Eq. (7), as, with $C_n(\lambda) = \{x : f_n(x) \geq \rho_n\}$, we have: $Q_n(C_n(\lambda)) \propto \int_{C_n(\lambda)} (T^{-1} f_n)^2(x) dx$. Hence, $Q_n(C_n(\lambda)) \lesssim \int_{\mathcal{X}} (T^{-1} f_n)^2(x) dx = \langle T^{-1} f_n, T^{-1} f_n\rangle_{L^2(\mathcal{X})} = \|f_n\|^2_{\mathcal{H}}$, which proves Proposition 3.1.

[3] We do not discuss in here the conditions on k needed to ensure the convergence of $\widehat{f}_n$.

Alternate proof consists in: Let $\mu = 1_{\{f_n > \rho_n\}}$ and assume the existence of $\Gamma_t(.) \in L^2(\mu)$ such that $k(t,\tau) =< \Gamma_t(.), \Gamma_\tau(.) >_{L^2(\mu)}$. With $P : L^2(\mu) \to H$ (built as a bijection, $g_n = P^{-1} f_n$ and $q(x) = g_n(x)^2 \mu(x)$, one has $Q(f_n > \rho_n) = \|g_n\|^2_{L^2(\mu)} = \|f_n\|^2_H$, which yields the equivalence (with stronger assumptions). One can easily check in both cases that $Q = \lim_{n\to\infty} Q_n$ is a measure. □

We recognize the standard one-class SVM in Eq. (7), as it writes:

$$\min_{f_n \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\rho_n - f_n(x_i)) \, 1_{\{f_n < \rho\}} + p_\lambda \|f_n\|^2_{\mathcal{H}} \tag{9}$$

Proposition 3.1 yields interesting interpretations: 1) the control over the richness of classes of functions obtained by minimizing $\|\cdot\|^2_{\mathcal{H}}$ expresses explicitly as the minimization of a certain measure of subsets of $\mathcal{X}$; 2) With $\mathcal{H}$ dense in $L^2(\mathcal{X})$, the f_n's are approximations for the indicator function of $C(\lambda)$ and the oscillating effect and the rationale for choosing ρ shortly lower than the density level in [4] are explained.

3.3. Parameter tuning

Proposition 3.1 shows that one-class SVMs are special instances of excess mass based kernel MMS estimation. More importantly, it also yields a representer theorem for excess mass estimation: the minimum measure set $\widehat{C}_n(\lambda) = \{x \in \mathcal{X}; \widehat{f}_n(x) \geq \rho_n\}$ is such that:

$$\widehat{f}_n(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \text{ with } \alpha_1, \ldots, \alpha_n \text{ in } \mathbb{R} \tag{10}$$

The hard issue in the above setting is the tuning of p_λ (see Section 2), because it may not be easily expressed as a practicable function of λ, and because it is a level of the density p, defined *wrt the measure* Q. The ν-SVM solution [4] consists of a re-writing the criterion in Eq. (9):

$$\min_{f_n \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} (\rho_n - f_n(x_i)) \, 1_{\{f_n(x_i) < \rho_n\}} + \frac{1}{2} \|f_n\|^2_{\mathcal{H}} - \nu \rho_n \right) \tag{11}$$

This modification enables to come back to the initial MMS estimation settings, Eq. (1). It can easily be shown that $\#\{x_i : f_n(x_i) < \rho_n\} \leq \nu n \leq \#\{x_i : f_n(x_i) \leq \rho_n\}$. Moreover, since almost surely with $\{x_1, \ldots, x_n\}$, , $\lim_{n\to\infty} P(\{f_n = \rho_n\}) = 0$, we have that $\nu \overset{a.s.}{=} \lim_{n\to\infty} P(\{f \leq \rho\})$. Hence, $\lambda = 1 - \nu$ which settles the connexion and shows the practicability of the approach.

In this Section, the interpretation of one-class SVMs as an excess mass approach in RKHS for estimating MMS is established formally. This connection enables the direct use of, *e.g.*, the pioneering work by Polonik [5] in order to derive theoretical studies of one-class SVMs convergence properties. Moreover, it justifies the use of one-class SVMs in many applications such as for novelty detection and change detection in Section 5. Under the assumption of a Gaussian kernel, a rate of convergence of 1-class SVM is obtained in forthcoming [9], which yields similar conclusions in the proposed framework.

4. CONVERGENCE ISSUES

For the sake of clarity, we state the rate of convergence obtained in [9] for the special case of a Gaussian kernel (See also work by Steinwart and coauthors). Let $k(\cdot,\cdot)$ be a Gaussian kernel with width parameter σ_n (which decay rate is chosen specifically), let d denote the dimension of $\mathcal{X}$, and suppose that for some $0 \leq \beta \leq 1$, $c_1 > 0$ and for $\delta \geq 0$, the pdf p satisfies:

$$\sup_{\|x - x'\| \leq \delta} |p(x) - p(x')| \leq c_1 \delta^\beta \tag{12}$$

Then, for any $\epsilon \geq 0$:

$$\|f - \widehat{f}_n\|^2_{L^2(\mathcal{X})} = O_P\left(\left(\frac{1}{n} \right)^{\frac{2\beta}{4\beta + (2+\beta)d} - \epsilon} \right) \tag{13}$$

In the following, we need to express this rate of convergence in terms of the sets $\widehat{C}_n(\lambda)$ and $C(\lambda)$ rather than the functions $\widehat{f_n}$ and f. We measure this rate with the symmetric difference:

$$d_\Delta(\widehat{C}_n(\lambda) \Delta C(\lambda)) = \text{Leb}\left(\widehat{C}_n(\lambda) \backslash C(\lambda) \right) + \text{Leb}\left(C(\lambda) \backslash \widehat{C}_n(\lambda) \right) \tag{14}$$

The first term in the right hand side writes (a similar reasoning holds for the second term):

$$\begin{array}{rcl} \text{Leb}(\widehat{C}_n(\lambda) \backslash C(\lambda)) & = & \text{Leb}(\{\widehat{f}_n \geq \rho_n\} \backslash \{f \geq \rho\}) \\ & \leq & \text{Leb}\left(\left\{ |f(x) - \rho| \leq c_2 n^{-\frac{2\beta}{4\beta + (2+\beta)d}} \right\} \right) \text{ with } c_2 > 0 \end{array} \tag{15}$$

Then, under Polonik's smoothness assumption:

$$\exists \gamma \text{ such that } \sup_\lambda \text{Leb}(\{x \in \mathcal{X}; |f(x) - \rho| \leq \eta\}) \lesssim \eta^\gamma \tag{16}$$

which yields:

$$\text{Leb}(\widehat{C}_n(\lambda) \backslash C(\lambda)) \lesssim n^{-\frac{2\gamma\beta}{4\beta + (2+\beta)d}} \tag{17}$$

In particular, if the Lebesgue density of P is regular ($\gamma = 1$), then:

$$d_\Delta(\widehat{C}_n(\lambda \Delta C(\lambda))) = O_P\left(n^{-\frac{2\beta}{4\beta + (2+\beta)d}} \right) \tag{18}$$

In the following, we use the rate obtained in Eq. (18) to justify the use of 1-class SVM for novelty detection. Similar development can be derived for change detection but will be omitted here due to the lack of space.

5. APPLICATIONS AND EXPERIMENTS

5.1. Novelty detection

Novelty detection consists of deciding whether new sample x is *novel* or not, based on the learning set $\{x_1, \ldots, x_n\}$, which yields the following hypothesis test, for a given λ:

$$\left\{ \begin{array}{ll} \text{Hypothesis } H_0 : & \text{Sample } x \text{ is not novel, } \textit{i.e.} \; x \in \widehat{C}_n(\lambda); \\ \text{Hypothesis } H_1 : & \text{Sample } x \text{ is novel, } \textit{i.e.} \; x \notin \widehat{C}_n(\lambda). \end{array} \right. \tag{19}$$

with H_0 the null hypothesis. This test has proved to yield solid performance in applications (industrial [10], audio [11]). A key quantity in the analysis of nonparametric detection tests is the *probability of false alarms*, denoted p^{fa}, and its empirical counterpart denoted p_n^{fa}:

$$p_n^{fa} = \mathbb{P}\left\{ x \notin \widehat{C}_n(\lambda) | x \in C(\lambda), x_1, \ldots, x_n \right\} \tag{20}$$

We study the asymptotic behavior of the probability of false alarm, and are interested in deriving a *central limit theorem*-like result for the convergence of p_n^{fa} to 0.

Proposition 5.1 *Under the assumptions of Section 4:*

$$\mathbb{P}\{|p_n^{fa} - E_n[p_n^{fa}]| \geq \epsilon\} \lesssim \exp(-\epsilon^2 n^{-\frac{2\beta}{4\beta + (2+\beta)d}}) \tag{21}$$

Proof (sketch): p_n^{fa} verifies the bounded difference property, as:

$$\begin{aligned} b_n &\equiv \sup_{x_i' \in \mathcal{X}} |p_n^{fa}(x_1, \ldots, x_n) \\ &\quad - p_n^{fa}(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \\ &\leq 2 \sup_{x \in \mathcal{X}} |p(x)| d_\Delta(\widehat{C}_n(\lambda) \Delta C(\lambda)) \lesssim n^{-\frac{2\beta}{4\beta+(2+\beta)d}} \end{aligned} \tag{22}$$

McDiarmid inequality then yields the deviation bound of Eq. (21).□

The proof techniques employed are similar to those in [12] where similar convergence rates are derived for $\widehat{C}_n(\lambda = 1)$ a nonparametric estimate for the support of P made of union of balls centered on the x_i's ($i = 1, \ldots, n$). In our case, however, rates are faster and obtained for any level λ. Up to our knowledge, the above convergence rate is the first result of this type for kernel-based novelty detection.

5.2. Change detection

The application we address in this Section is change detection, which framework is similar to that of Section 5.1. Here, however, we test T samples instead of one. More formally, let $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} P_0$ and $x_{n+1}, \ldots, x_{n+T} \overset{\text{i.i.d.}}{\sim} P_1$. We do not consider the classic detection problem of deciding whether $P_0 = P_1$ or not; instead, we concentrate on the related detection problem based on the comparison, for a given λ, of $C_0(\lambda)$ and $C_1(\lambda)$, as in [3, Section 4]. We therefore implement the following hypothesis test, for threshold t:

$$\begin{cases} \text{Hypothesis } H_0: & d_\Delta(\widehat{C}_{0,n}(\lambda), \widehat{C}_{1,T}(\lambda)) < t; \\ \text{Hypothesis } H_1: & d_\Delta(\widehat{C}_{0,n}(\lambda), \widehat{C}_{1,T}(\lambda)) \geq t. \end{cases} \tag{23}$$

with H_0 the null hypothesis and where $\widehat{C}_{0,n}(\lambda)$ (resp. $\widehat{C}_{1,T}(\lambda)$) is the estimate of the Q-MMS with P_0-measure (resp. P_1-measure) λ based on the learning set $\{x_1, \ldots, x_n\}$ (resp. $\{x_{n+1}, \ldots, x_{n+T}\}$).

The above procedure enables change detection via the tuning of λ. The test replaces the comparison of the unknown measures P_0 and P_1 with a comparison of MMSs. A slightly modified version of this test was successfully applied to music segmentation of complex audio signals, see [2].

In the remainder of this Section, we compare the performance of kernel-based MMS (1-class SVM) change detection to a particle filter based Generalized Likelihood Ratio (GLR) approach. The time-series we consider is defined by the popular toy nonlinear model:

$$\begin{aligned} x_i &= \tfrac{1}{2} x_{i-1} + a_1 \tfrac{x_{i-1}}{1+x_{i-1}^2} + 8\cos(1.2(i-1)) + \omega_{i-1} \\ y_i &= a_2 x_i^2 + v_{i-1} \end{aligned} \tag{24}$$

with $\omega_i \sim \mathcal{N}(0, 0.1)$ and $v_i \sim \mathcal{N}(0, 1)$. Both approaches are used to detect changes in the time-series for 500 realizations of ω_i and v_i; in the first 250 signals, there is a change at time instant T in the model parameters, with a_1 jumping from 25 to 12.5, and a_2 from 0.05 to 0.1035. The other 250 signals are kept with $a_1 = 25$ and $a_2 = 0.05$.

The two approaches are tuned as follows. The MMS method uses settings described in see [2] where the space $\mathcal{X}$ derives from the Gaussian window (length 51 points) Spectrogram of $\{y_i\}$, for $i = 1, 2, \ldots$. 51. Each x_i is sub-image made of 25 consecutive spectrogram columns; the learning sets sizes are $n = T = 10$. The ν 1-class SVM parameters are $\nu = 0.5$ and $\sigma = 25$ for kernel width. The particle filter GLR is given the correct model and tuned as in [13]. Figure 2 plots the ROC curves (true alarms vs. false alarms). Both methods have good performance, as confirmed by the estimated change time instants histograms.

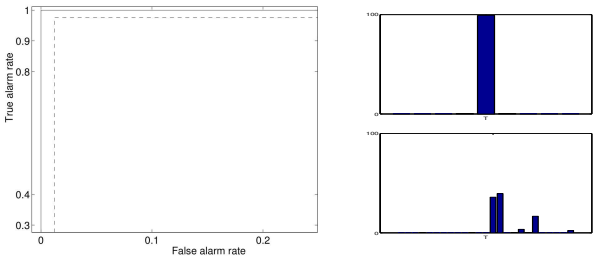

Fig. 2: ROC curves (left) and histograms of estimated change time instants for MMS-based approach (solid, and top right) and particle filtering based GLR (dash, and bottom right). MMS-based approach and the particle filter GLR both perform very accurately.

6. CONCLUSION AND PERSPECTIVES

In this paper, we recall fundamentals concerning MMSs and present results including estimation, connection with 1-class SVMs, theoretical justification for MMS novelty detection. The application to change detection, though implementing a suboptimal hypothesis test, yields better performance on a highly nonlinear time series than a particle filter GLR using the correct model. Short-term perspectives include the application of MMSs to the problem of defining a kernel between finite sets of vectors as well as the improvement of first good results obtained for change detection in music signals.

7. REFERENCES

[1] B. Schölkopf and A.J. Smola, *Learning with Kernels*, The MIT Press, 2002.

[2] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Transactions on Signal Processing*, August 2005.

[3] L. Devroye and G.L. Wise, "Detection of abnormal behavior via nonparametric estimation of the support," *SIAM Journal on Applied Mathematics*, vol. 38, no. 3, pp. 480–488, June 1980.

[4] B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, and J.C. Platt, "Support vector method for novelty detection," in *NIPS*, 2000, vol. 12, pp. 582–588.

[5] W. Polonik, "Measuring mass concentration and estimating density contour clusters - an excess mass approach," *Ann. Statist.*, vol. 23, pp. 855–881, 1995.

[6] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel hilbert spaces in probability and statistics*, Kluwer academic press, 2004.

[7] S. Canu, X. Mary, and A. Rakotomamonjy, "Functional learning through kernel," in *Advances in Learning Theory: Methods, Models and Applications. NATO Science Series III: Computer and Systems Sciences*, 2003.

[8] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, 2001.

[9] R. Vert and J.P. Vert, "Consistency of one-class svm and related algorithms," in *NIPS*, 2005.

[10] P. Hayton, B. Scholkopf, L. Tarassenko, and P. Anuzis, "Support vector novelty detection applied to jet engine vibration spectra," in *NIPS*, 2000.

[11] M. Davy, F. Desobry, and A. Gretton, "An online support vector machine for abnormal event detection," *Signal Processing*, 2005, to appear.

[12] A. Baillo, A. Cuevas, and A. Justel, "Set estimation and nonparametric detection," *The Canadian Journal of Statistics*, vol. 28, no. 4, pp. 765–782, 2000.

[13] V. Kadirkamanathan, P. Li, M.H. Jaward, and S.G. Fabri, "An smc filtering approach to fault detection and isolation in nonlinear systems," in *IEEE CDC*, 2000.