Time-frequency learning machines

Paul Honeiné, Cédric Richard, *Member*, IEEE, Patrick Flandrin, *Fellow*, IEEE

***Abstract*— Over the last decade, the theory of reproducing kernels has made a major breakthrough in the field of pattern recognition. It has led to new algorithms, with improved performance and lower computational cost, for non-linear analysis in high dimensional feature spaces. Our paper is a further contribution which extends the framework of the so-called kernel learning machines to time-frequency analysis, showing that some specific reproducing kernels allow these algorithms to operate in the time-frequency domain. This link offers new perspectives in the field of non-stationary signal analysis, which can benefit from the developments of pattern recognition and Statistical Learning Theory.**

***Index Terms*— Time-frequency analysis, kernel machines, learning theory, support vector machines.**

I. Introduction

TIME-frequency and time-scale distributions provide a powerful tool for non-stationary signal analysis. Unlike conventional spectral methods, they reveal the time-varying spectral content of one-dimensional signals by mapping them into a two-dimensional time-frequency domain. Substantial theoretical work has been carried out in this direction and has yielded many different classes of time-frequency distributions, parametric or otherwise, in which optimal solutions for a given signal or task can be selected. As an example, distributions dedicated to signal analysis are studied in [1]–[3] whereas optimal distributions for signal detection are considered in [4], [5].

Since the pioneering work of Aronszajin [6], pattern recognition based on reproducing kernel Hilbert spaces (RKHS) has gained wide popularity. The most prominent recent developments include support vector machines (SVM) [7], kernel principal component analysis (KPCA) [8], kernel Fisher discriminant analysis (KFDA) [9], and its generalization to multiclass problems, kernel generalized discriminant analysis (KGDA) [10]. A key property behind such algorithms is that they can be expressed in terms of inner products only, involving pairs of input data. Replacing these inner products with a reproducing kernel provides an efficient way to implicitly map the data into a high, even infinite, dimensional RKHS and apply the original algorithm in this space. Because calculations are then carried out without making direct reference to the non-linear mapping of input vectors, this principle is commonly called the *kernel trick*. Kernel-based algorithms are computationally very efficient, and generally have their generalization performance guaranteed by Statistical Learning Theory [11], [12].

With the exception of [13], [14], there are very few works combining kernel learning machines and time-frequency analysis, although the interest in pattern recognition based on time-frequency representations remains strong. In [13], the authors solve a signal classification problem using a SVM and a reproducing kernel expressed in the time-frequency domain. Reproducing wavelet kernels are considered in [14] for non-parametric regression. Clearly, time-frequency analysis still has not fully benefited from the rapid development of kernel learning machines. In this paper, we show that some specific reproducing kernels allow any kernel learning machine to operate in the time-frequency domain. For the sake of simplicity, we begin by describing our approach applied to the Wigner distribution. Next we apply it to other time-frequency distributions. But before, let us briefly review the basics concepts of kernel learning machines.

P. Honeiné and C. Richard are with the laboratory ICD-M2S (FRE CNRS 2848), Université de technologie de Troyes, Troyes, France. P. Flandrin is with the Laboratoire de Physique (UMR CNRS 5672), ENS de Lyon, France

II. Kernel learning machines: a brief review of basic concepts

Most kernel learning machines are statistical learning algorithms that take advantage of the geometric and regularizing properties of RKHS, which are established by the kernel trick and the representer theorem [15], [16]. In this section, we briefly introduce these concepts through an example.

A. Example of kernel-based method: the KPCA algorithm

Problems commonly encountered in machine learning start with a training set $\mathcal{A}_n$ containing n instances $x_i \in \mathcal{X}$ and, in a supervised context, their labels or desired outputs $y_i \in \mathcal{Y}$. The objective of the exercise is usually related to feature extraction, density estimation or classification. Linear methods have played a crucial role in the development of machine learning because of their inherent simplicity from conceptual and implementational points of view. However, in many fields of current interest such as biological engineering and communications, it is necessary to deal with non-linear complex phenomena. A possible way to extend the scope of linear learning machines is to map the input data from $\mathcal{X}$ into a feature space $\mathcal{F}$ via a non-linear mapping $\phi(\cdot)$. The n instances $\phi(x_i)$ are then used as training samples. Clearly, this basic strategy may fail when $\mathcal{F}$ is a very high, or even infinite, dimensional space. As shown below with the KPCA algorithm, kernel learning machines overcome this limitation by using a powerful computational shortcut.

KPCA is a non-linear form of principal component analysis (PCA) which allows to extract features that are non-linearly related to the input variables. Consider a set of n data points $\phi(x_i)$ mapped into a feature space $\mathcal{F}$ and centered at the origin, that is, $\sum_{i=1}^{n} \phi(x_i) = 0$. PCA is performed in $\mathcal{F}$ by solving the eigenvalue problem $\Sigma\Phi = \mu\Phi$ with $\Phi \neq 0$, where $\Sigma = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\,\phi(x_i)^\top$ is the covariance matrix. If $\mathcal{F}$ is infinite-dimensional, note that $\phi(x_i)\,\phi(x_i)^\top$ may be seen

as the projection operator onto direction $\phi(x_i)$. The problem becomes $\sum_{i=1}^{n} \langle \phi(x_i), \Phi \rangle \, \phi(x_i) = n\mu \, \Phi$, where $\langle \cdot \, , \cdot \rangle$ is the inner product in $\mathcal{F}$. Observe that any solution Φ_k with $\mu_k \neq 0$ must lie in the span of $\phi(x_1) \ldots, \phi(x_n)$, and can then be expanded as follows

$$\Phi_k = \sum_{i=1}^{n} a_{i,k} \, \phi(x_i). \tag{1}$$

As detailed in [8], substituting this expression into the eigenvalue equation, and multiplying it from the right by $\phi(x_j)$, we obtain the following eigenvalue problem:[1] $K a_k = \lambda_k \, a_k$, where K is the n-by-n Gram matrix whose (i,j)-th entry is $\langle \phi(x_i), \phi(x_j) \rangle$, and $\lambda_k = n\mu_k$. The components of the k-th eigenvector a_k of K are the $a_{i,k}$'s defined in equation (1). Finally, the k-th principal component can be extracted from any $\phi(x)$ by projecting it onto Φ_k, namely,

$$\langle \phi(x), \Phi_k \rangle = \sum_{i=1}^{n} a_{i,k} \, \langle \phi(x), \phi(x_i) \rangle. \tag{2}$$

The most interesting characteristic of this algorithm is that the non-linear mapping $\phi(\cdot)$ only appears in the form of inner products $\langle \phi(x_i), \phi(x_j) \rangle$. Suppose we are given a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ with the property that there exists a map $\phi : \mathcal{X} \to \mathcal{F}$ such that for all $x_i, x_j \in \mathcal{X}$, we have $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. For a survey, see e.g. [6]. The so-called *kernel trick* consists of substituting each inner product $\langle \phi(x_i), \phi(x_j) \rangle$ by $\kappa(x_i, x_j)$. The power of this principle lies in that the inner products in the feature space $\mathcal{F}$ are computed without explicitly carrying out or even knowing the mapping $\phi(\cdot)$, which results in computationally efficient algorithms. Classic examples of valid kernels are the q-th degree polynomial kernel $\kappa(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^q$ and the Gaussian kernel $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$, where σ is the kernel bandwidth. Note that the feature space corresponding to the latter is infinite dimensional.

B. The kernel trick, the representer theorem

The kernel trick can be used to transform any linear data processing technique into a non-linear one, on the condition that the algorithm can be expressed in terms of inner products only, involving pairs of the input data. This is achieved by substituting each inner product $\langle x_i, x_j \rangle$ by a kernel $\kappa(x_i, x_j)$, leaving the algorithm unchanged and incurring essentially the same computational cost. In conjunction with the kernel trick, the representer theorem [15], [16] is a solid foundation of kernel-based methods such as SVM, KFDA, KGDA and KPCA. Consider the learning machine $\mathcal{L}(x) = \langle \phi(x), \Phi \rangle$ and the regularized risk functional

$$\sum_{i=1}^{n} V(\mathcal{L}(x_i), y_i) + \lambda \|\mathcal{L}\|^2, \tag{3}$$

with $V(\mathcal{L}(x), y)$ the cost of predicting $\mathcal{L}(x)$ when the desired output is y, and λ a positive parameter. The representer theorem states that, under very general conditions on the loss function V, any $\Phi \in \mathcal{F}$ minimizing the criterion (3) admits a representation of the form

$$\Phi = \sum_{i=1}^{n} a_i \, \phi(x_i). \tag{4}$$

This leads to the well-known kernel expansion $\mathcal{L}(x) = \sum_{i=1}^{n} a_i \, \kappa(x, x_i)$. As an example, note that expressions (1) and (2) arise as a direct consequence of this theorem when it is applied to the problem of PCA in feature space.

III. Time-frequency learning machines: general principles

The reason for time-frequency analysis is to give a mathematical core to the intuitive concept of time-varying Fourier spectrum for non-stationary signals. For a survey, see [17], [18] and references therein. Most of the parametric distributions of current interest belong to the Cohen class, which has proven useful in identifying non-stationarities in signals produced by a host of real-world applications. In this section, we investigate the use of kernel learning machines for pattern recognition in the time-frequency domain. To clarify the discussion, we will first focus on the Wigner distribution, which plays a central role in the Cohen class. This will be followed by an extension to other time-frequency distributions.

A. The Wigner distribution

Among the myriad of time-frequency representations that have been proposed, the Wigner distribution is considered fundamental in a number of ways. Its usefulness derives from the fact that it satisfies many desired mathematical properties such as the correct marginal conditions and the weak correct-support conditions [17], [18]. This distribution is also a suitable candidate for time-frequency-based detection since it is covariant to time shifts and frequency shifts and it satisfies the unitarity condition [4]. Let $\mathcal{X}$ be a subspace of $L_2(\mathbb{C})$, the space of finite-energy complex signals, equipped with the usual inner product $\langle x_i, x_j \rangle = \int_t x_i(t) \, x_j^*(t) \, dt$ and its corresponding norm. The Wigner distribution of any signal $x \in \mathcal{X}$ is defined as

$$W_x(t,f) = \int x(t + \tau/2) \, x^*(t - \tau/2) \, e^{-2j\pi f \tau} \, d\tau. \tag{5}$$

By applying conventional linear pattern recognition algorithms directly to time-frequency representations, we seek to determine a time-frequency pattern $\Phi(t,f)$ so that

$$\langle W_x, \Phi \rangle = \iint W_x(t,f) \, \Phi(t,f) \, dt \, df \tag{6}$$

optimizes a given criterion of the general form (3). The principal difficulty encountered in solving such problems is that they are typically very high dimensional, the size of Wigner distributions calculated from the training set being quadratic in the length of signals. This makes pattern recognition based on time-frequency representations time-consuming, if not impossible, even for reasonably-sized signals. With the kernel trick and the representer theorem, kernel learning

[1] The eigenvalue problem for non-centered data points $\phi(x_i)$ in feature space $\mathcal{F}$ is given by: $(K - 1_n K - K 1_n + 1_n K 1_n) a_k = \lambda_k \, a_k$, with $K(i,j) = \langle \phi(x_i), \phi(x_j) \rangle$ and $1_n(i,j) = \frac{1}{n}$ [9].

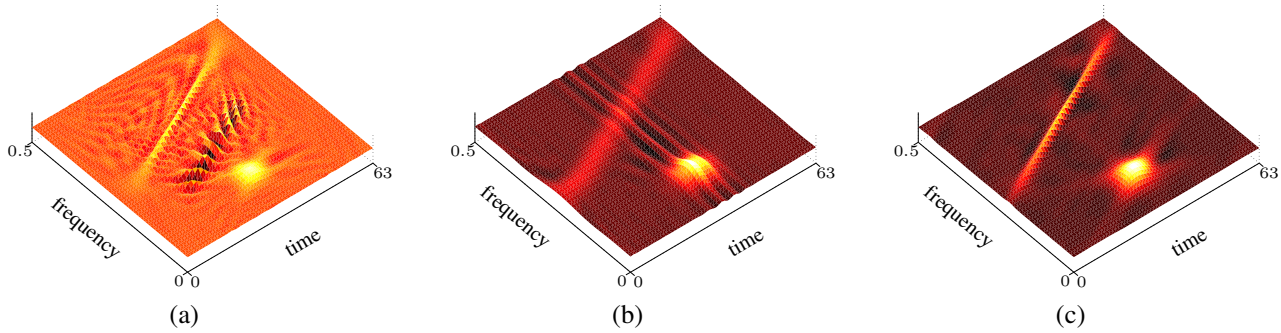

Fig. 1. First eigendistribution Φ_1 obtained using (a) Wigner-based KPCA, (b) Choï-Williams-based KPCA and (c) AOK algorithm applied to the Wigner-based KPCA result. The experiments were carried out using a collection of 1000 signals, each of length 64, consisting of a linear chirp and a Gaussian pulse in an additive white Gaussian noise with a standard deviation of 1.

machines eliminate this computational burden. It suffices to consider the following kernel

$$\kappa_W(x_i, x_j) = \langle W_{x_i}, W_{x_j} \rangle, \tag{7}$$

and note that W_{x_i} and W_{x_j} do not need to be computed since, by the unitarity of the Wigner distribution, we have

$$\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2. \tag{8}$$

Any kernel learning machine proposed in the literature can then be used with the kernel (8) to perform pattern recognition tasks in the time-frequency domain. The solution $\mathcal{L}(x) = \sum_{i=1}^{n} a_i \, |\langle x, x_i \rangle|^2$ guaranteed by the representer theorem allows for a time-frequency distribution interpretation, $\mathcal{L}(x) = \langle W_x, \Phi_W \rangle$, with

$$\Phi_W = \sum_{i=1}^{n} a_i \, W_{x_i}. \tag{9}$$

We should again emphasize that the coefficients a_i are estimated without calculating any Wigner distribution. The time-frequency features are obtained from a decomposition into time-domain ones, which is an interesting consequence of the kernel trick and the representer theorem that is worth mentioning. The time-frequency pattern Φ_W can be subsequently evaluated with (9), in an iterative manner, without suffering the drawback of storing and manipulating a large collection of Wigner distributions. The inherent sparsity of the coefficients a_i produced by most of the kernel learning machines, a typical example of which is SVM, may speed-up the calculation of Φ_W.

B. Example of time-frequency learning machine: the Wigner-based KPCA

We proceed now to illustrate the concept of time-frequency learning machines outlined above through an example involving KPCA, described in Section II-A, and the Wigner distribution. There have been so many contributions related to eigenvalue and singular value decompositions of time-frequency distributions that we can only mention one of the first papers [19] and recent applications [20], [21]. Consider a collection of n signals $x_1, \ldots, x_n$, each of length d. KPCA can be adapted to operate directly on their Wigner distributions by using the kernel $\kappa_W(x_i, x_j)$. The central step of the algorithm is to perform eigendecomposition of the Gram matrix $K_W - 1_n K_W - K_W 1_n + 1_n K_W 1_n$, with $K_W(i,j) = \kappa_W(x_i, x_j)$ and $1_n(i,j) = \frac{1}{n}$. Let the eigenvectors and eigenvalues be denoted by a_k and λ_k, respectively, with $\lambda_1 \geq \lambda_2 \geq \ldots$ The k-th principal component can be extracted from any signal x as follows

$$\mathcal{L}_k(x) = \langle W_x, \Phi_k \rangle = \sum_{i=1}^{n} a_{i,k} \, \kappa_W(x, x_i), \tag{10}$$

with $\Phi_k = \sum_{i=1}^{n} a_{i,k} \, W_{x_i}$, and $a_{i,k}$ the i-th component of a_k. As shown in [9], the normalization requirement $\|\Phi_k\| = 1$ leads to the condition $\lambda_k \|a_k\|^2 = 1$. We call Φ_k the k-th eigendistribution. We call Φ_k the k-th eigendistribution. Note, however, that it is not a valid time-frequency distribution in that no signal with time-frequency transform Φ_k necessarily exists. This approach is summarized in Table I.

To show that KPCA is a potentially useful tool in time-frequency signal processing, a set of 1000 noisy signals of length 64 was generated. Each signal was made up of a linear chirp with normalized frequency increasing from 0.2 Hz to 0.45 Hz, a Gaussian pulse centered at time index 32 and normalized frequency 0.1 Hz, and an additive zero-mean white Gaussian noise with a standard deviation of 1. This resulted in a signal-to-noise ratio of -5.1 dB in the time-frequency domain. Wigner-based KPCA was performed to determine the eigendistributions Φ_k. Their calculation was based on the discrete-time discrete-frequency Wigner distribution introduced in [23] since it satisfies most of properties of its continuous counterpart (5), in particular unitarity. The first eigendistribution Φ_1 represented in Figure 1(a) shows that significant information has been successfully extracted from noisy data. The two signal components can be clearly distinguished in it, as well as oscillating interferences that are characteristic of the Wigner distribution and often limit its expertise. This observation is corroborated by an increase in signal-to-noise ratio of 8.5 dB for Wigner distributions projected into the space spanned by the eigendistributions Φ_1 and Φ_2.

Applying standard PCA directly to the set of Wigner distributions would lead to the same result. However, this approach usually suffers from the high computational cost of calculating the d^2-by-d^2 covariance matrix of n time-frequency distributions, each of size d-by-d, and performing its eigendecomposi-

TABLE I
THE WIGNER-BASED KPCA ALGORITHM

Instructions	Complexity
1. Compute the Wigner distribution of each one of the n signals	$\mathcal{O}(d^2 \log d)$ per signal [22]
2. Compute the Gram matrix $K_W(i,j) = \kappa_W(x_i, x_j)$	$\mathcal{O}(dn^2)$
3. Perform eigendecomposition of $K_W - 1_n K_W - K_W 1_n + 1_n K_W 1_n$	$\mathcal{O}(n^3)$
4. Compute and normalize the eigendistributions Φ_k	$\mathcal{O}(d^2 n)$ per eigendistribution

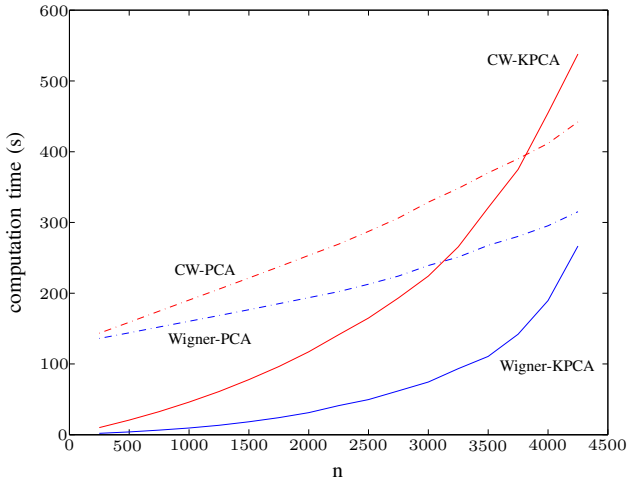

Fig. 2. Computation times of KPCA and PCA applied to time-frequency distributions, as a function of the number n of 64-sample signals. The Wigner distribution and the Choï-Williams distribution are considered. The former is a unitary distribution whereas the latter is not. All the codes were implemented in Matlab and run on a laptop PC with 1 GB RAM and Pentium M 1.60 GHz processor.

tion. The complexities of these computationally intensive steps are $\mathcal{O}(d^4 n)$ and $\mathcal{O}(d^6)$, respectively. They replace instructions 2. and 3. of the Wigner-based KPCA algorithm depicted in Table I, whose computational complexities are $\mathcal{O}(dn^2)$ and $\mathcal{O}(n^3)$, respectively. Figure 2 shows the computation time of both methods plotted as a function of n, with d fixed to 64. As expected, it is almost linear in n for PCA applied directly to Wigner distributions, and polynomial for Wigner-based KPCA. It can also be verified that the latter is computationally more efficient than PCA as long as n is less than d^2, a condition often satisfied in practice. These conclusions remain valid for standard pattern recognition methods that require either inversion or eigendecomposition of covariance matrices, such as FDA and GDA.

C. Application to other time-frequency distributions

Obviously, the concept of time-frequency learning machine is not limited only to the Wigner distribution. In this subsection, we illustrate it with other popular time-frequency distributions, linear and quadratic.

The short-time Fourier transform is probably the most common example of linear time-frequency distribution. Denoting by $w(t)$ an analysis window localized around the origin of the time-frequency domain, it is defined by

$$F_x(t,f) = \int x(\tau)\, w^*(\tau - t)\, e^{-2j\pi f\tau}\, d\tau, \qquad (11)$$

or, in an equivalent way, $F_x(t,f) = \langle x, w_{t,f}\rangle$ with $w_{t,f}(\tau) = w(\tau - t)\, e^{2j\pi f\tau}$. The kernel function $\kappa_F(x_i, x_j) = \langle F_{x_i}, F_{x_j}\rangle$, namely,

$$\kappa_F(x_i, x_j) = \|w\|^2 \langle x_i, x_j\rangle. \qquad (12)$$

can be used with any kernel learning machine proposed in the literature. The solution guaranteed by the representer theorem offers a time-frequency distribution interpretation: $\mathcal{L}(x) = \langle F_x, \Phi_F\rangle$ with $\Phi_F = \sum_{i=1}^n a_i\, F_{x_i}$.

The use of quadratic forms in non-stationary signal analysis is motivated by the need to collect information on the distribution of signal energy over time and frequency. Over the years, the Cohen class has received considerable attention because it contains all the distributions C_x that are covariant with respect to time-frequency shifts applied to the signal. These are taking the form

$$C_x(t,f) = \iint \Pi(t' - t, f' - f)\, W_x(t', f')\, dt'\, df' \qquad (13)$$

where Π is a weighting function. We can easily check that $\kappa_C(x_i, x_j) = \langle C_{x_i}, C_{x_j}\rangle$ is a valid kernel that can be used by any kernel learning machine. The solution can further be rewritten as $\mathcal{L}(x) = \langle \Phi_C, C_x\rangle$ with $\Phi_C = \sum_{i=1}^n a_i\, C_{x_i}$. The advantage of κ_C over κ_W is that the correlative form (13) can be exploited to improve the readability of Φ_C, that may be affected by the presence of troublesome oscillating interferences. Nevertheless, as can be seen on Figure 1(b) with an example of Choï-Williams-based KPCA, the same processing is simultaneously applied to interference terms and signal components, removing the former ones and spreading out the latter. The most popular quasi-interference-free distribution of the Cohen class is certainly the spectrogram. Formally defined as the squared magnitude of the short-time Fourier transform, the spectrogram is related to the kernel $\kappa_S(x_i, x_j) = \iint |\langle x_i, w_{t,f}\rangle\langle x_j, w_{t,f}\rangle|^2\, dt\, df$. Other examples of distributions include those that satisfy the unitary condition $\langle C_{x_i}, C_{x_j}\rangle = |\langle x_i, x_j\rangle|^2$, e.g., the Wigner distribution, the Page distribution and the Rihaczek distribution. Kernel learning machines $\mathcal{L}(x) = \langle \Phi_C, C_x\rangle$ based on unitary distributions share the same kernel (8), and then have the same performance. They differ by their time-frequency pattern Φ_C, which can be computed directly or using Φ_W as follows:

$$\Phi_C(t,f) = \iint \Pi(t' - t, f' - f)\, \Phi_W(t', f')\, dt'\, df'. \qquad (14)$$

In the general case of non-unitary distributions, the calculation of $\kappa_C(x_i, x_j)$ can be a time consuming part of training processes since it explicitly involves pairs of d-by-d time-frequency distributions. Computing the Gram matrix K_C thus

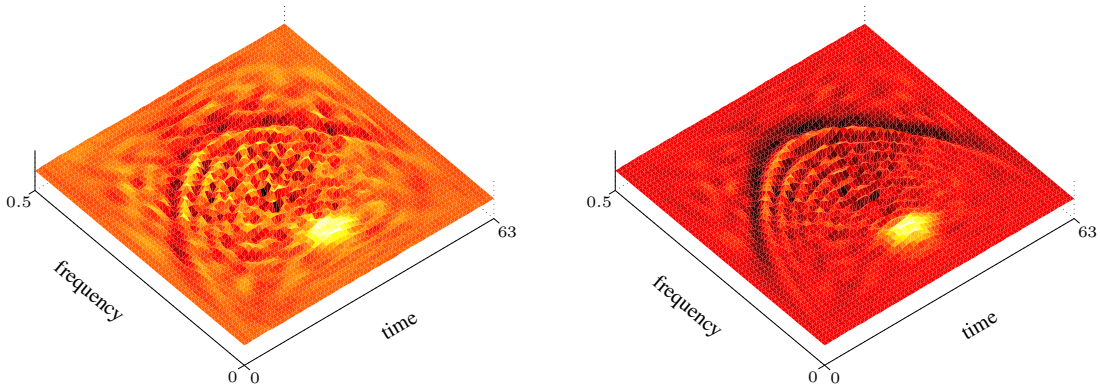

Fig. 3. Discriminant information extracted by (a) Wigner-based KFDA and (b) Wigner-based SVM. The experiments were carried out using two classes of 1000 signals, each of length 64, consisting of a frequency-modulated waves with a parabolic modulation and a Gaussian pulse, respectively, in an additive white Gaussian noise with a standard deviation of 2.2.

costs $\mathcal{O}(n^2 d^2)$. Definition (13) shows that extra computation cost of $\mathcal{O}(d^2)$ is also required to calculate time-frequency distributions other than the Wigner distribution, but the whole computation also takes time of order $\mathcal{O}(d^2 \log d)$ per distribution [22]. Computation times of Wigner-based and Choï-Williams-based KPCA plotted in Figure 2 corroborate this analysis. It can also be observed that Choï-Williams-based KPCA is computationally more efficient than PCA as long as n is strictly less than d^2, that is, as long as the size of Gram matrix is less than the size of covariance matrix.

In applications where computation time is a crucial factor, we suggest a simple heuristic procedure to derive rules of the form $\mathcal{L}(x) = \langle C_x, \Phi_C \rangle$. It consists of training the kernel learning machine with $\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2$. The time-frequency feature Φ_C is then obtained from Φ_W and equation (14). This strategy is clearly non-optimal when C_x does not satisfy the unitarity condition. However, it greatly improves computational efficiency and also simplifies the use of signal-dependent methods of designing Π. Figure 1(c) provides an example of time-frequency-based PCA followed by the AOK algorithm [2]. This result demonstrates a significant visual improvement over those given in Figures 1(a) and 1(b). A complete analysis of this heuristic falls beyond the scope of this paper and will be addressed in the future.

D. Signal classification with time-frequency learning machines

The last ten years have seen an explosion of research in supervised [9], [10], [24] and unsupervised [25] classification techniques based on kernels; see [26] for a recent survey. These include SVM, which map data into a high dimensional space where the classes of data are more readily separable, and maximize the distance – or margin – between the separating hyperplane and the closest points of each class [11]. SVM basically involve formulating the margin maximization problem into dot product form in order to use kernels, and solving a quadratic programming problem to estimate the parameters a_i of the test statistic

$$\mathcal{L}(x) = \sum_{i=1}^{n} a_i \, \kappa(x, x_i). \tag{15}$$

The excellent performance of SVM has inspired countless works in discriminant analysis. In particular, KFDA is a powerful method of obtaining non-linear Fisher discriminants. It also uses the kernel trick and the representer theorem to design, via an eigenvalue problem, kernel-based classifiers of the form (15) that maximize the Rayleigh coefficient of the between and the within class scatter matrices [9], [24]. KGDA is an extension to this approach for handling multiclass problems [10]. We are now going to illustrate the concept of time-frequency learning machines through supervised classification problems involving SVM, KFDA and KGDA.

The first example deals with a binary classification task involving two classes of 1000 signals of length 64. Each signal was consisting either of a Gaussian pulse centered at time index 32 and normalized frequency 0.1 Hz, or of a frequency-modulated wave with a parabolic modulation between 0.1 Hz and 0.4 Hz, corrupted by an additive zero-mean white Gaussian noise with a standard deviation of 2.2. Two time-frequency learning machines were evaluated, Wigner-based SVM and Wigner-based KFDA, obtained from SVM and KFDA with $\kappa_W(x_i, x_j)$. Preliminary experiments were conducted on a cross-validation set of 1000 signals to select the best settings for each algorithm. The regularization parameter c of SVM was set to 1, and KFDA was regularized by adding 10^{-3} to the diagonal elements of within class scatter matrix. Figure 3 provides, in both cases, the time-frequency pattern Φ that follows from the reformulation of (15) in terms of Wigner distributions. The signal components can be clearly distinguished, showing that discriminant information has been successfully extracted from training data. The positive orientation of the Gaussian pulse, and the negative orientation of the frequency-modulated wave and its interference terms, allow the test statistic (15) to increase the separability of the two competing classes of signals. Note that Figure 3(b) is more visually meaningful than Figure 3(a), which is corroborated by the error rate of the Wigner-based SVM and the Wigner-based

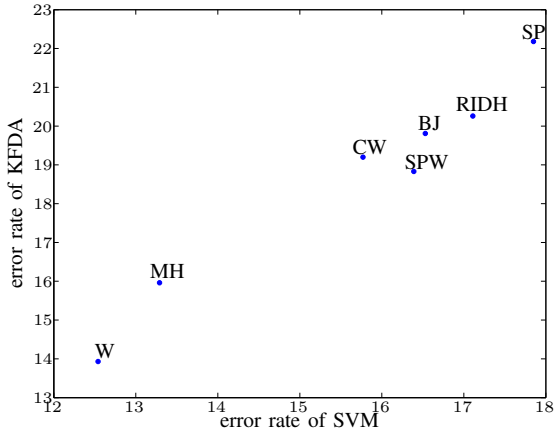

Fig. 4. Error rates of time-frequency-based SVM and KFDA using various distributions of the Cohen class: Wigner (W), smoothed pseudo-Wigner (SPW), Margenau-Hill (MH), Choï-Williams (CW), Born-Jordan (BJ), reduced-interference with Hanning window (RIDH) and spectrogram (SP). Same data as in Fig. 3.

KFDA estimated over a 10000-sample test set: 12.5% for the former and 13.9% for the latter. This illustrates the good generalization ability of SVM that makes them so popular. Our approach enables this technique to be adapted for use with time-frequency distributions. This experiment was extended to a wide variety of distributions of the Cohen class: smoothed pseudo-Wigner, Margenau-Hill, Choï-Williams, Born-Jordan, reduced-interference with Hanning window and spectrogram. Parameters of these distributions were set to the default values proposed by the Matlab toolbox TFTB[2]. Figure 4 shows that the best performance was obtained with the Wigner distribution, which means that the filtering process (13) caused some loss of relevant information in the other distributions. Wigner-based learning machines were also computationally the most efficient.

Given a m-class classification problem, KGDA provides $(m-1)$ test statistics of the form (15) that maximize the Rayleigh ratio of between and within class scatter matrices. The classification of new data is then achieved by comparing these test statistics to predetermined thresholds. As another application of time-frequency learning machines, we propose here to perform KGDA with kernel $\kappa_W(x_i, x_j)$. This example deals with a classification problem involving three classes of 300 signals, each of length 64. Each signal was consisting either of a Gaussian pulse centered at time index 32 and normalized frequency 0.25 Hz, of a 0.1 Hz sine wave, or of a frequency-modulated wave with a parabolic modulation between 0.15 Hz and 0.45 Hz. These signals were corrupted by an additive zero-mean white Gaussian noise with a standard deviation of 0.45. Figures 5(a) and 5(b) represent the time-frequency patterns extracted by the Wigner-based KGDA algorithm, denoted by Φ_a and Φ_b, respectively. Each signal can be easily distinguished in these distributions, meaning that all the discriminant information has been successfully collected. In particular, note that the task of Φ_a is to discriminate the parabolic modulation from the sine wave and the Gaussian pulse since the former is negatively oriented while the latter are positively oriented. The time-frequency pattern Φ_b makes it possible to discriminate between the parabolic modulation, which is almost absent, the sine wave and the Gaussian pulse, which are respectively negatively and positively oriented. This analysis is confirmed in Figure 5(c), where each training data x_i is represented in the $\Phi_a\Phi_b$ plane by a point whose coordinates are $\langle W_{x_i}, \Phi_a \rangle$ and $\langle W_{x_i}, \Phi_b \rangle$. The class of parabolic modulations is characterized by negative Φ_a-coordinates whereas the classes of sine waves and Gaussian pulses correspond to positive Φ_a-coordinates. Along the Φ_b-axis, the coordinates of the sine waves, parabolic chirps and gaussian pulses are negative, close to zero and positive, respectively.

[2]The Time-Frequency Toolbox (TFTB) is downloadable from http://tftb.nongnu.org/.

IV. Conclusion

The theory of reproducing kernels enabled the development of new learning algorithms for pattern recognition, whose formulation is independent of the representation space of data. Their success was largely influenced by the emerging field of Statistical Learning Theory, which simultaneously provided fundamental bounds on achievable performance. In this paper, we have focused our attention on the new concept of time-frequency learning machines. It takes advantage of this progress for implementing universal learning machines that extract time-frequency information from signals. We have illustrated the efficiency of these novel techniques for non-stationary signal analysis through unsupervised and supervised learning problems. Time-frequency learning machines can be used in many other applications, such as blind source separation [27] and filtering [28], where kernel-based methods have proved their efficiency. Their extension to higher order distributions also seems feasible.

In ongoing studies, we are investigating kernel-based methodologies that could be advantageously used to solve recurrent problems in the field of non-stationary signal analysis. For instance, we have recently proposed a method for selecting time-frequency distributions appropriate for given learning tasks [29]. It is based on a criterion that has recently emerged from the machine learning literature: the kernel-target alignment. Further work may contribute to strengthen these connections with the most recent methodological and theoretical developments of pattern recognition and Statistical Learning Theory, in order to offer new perspectives in the field of non-stationary signal analysis.

References

[1] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[2] D. Jones and R. Baraniuk, "An adaptive optimal-kernel time-frequency representation," *IEEE Transactions on Signal Processing*, vol. 43, no. 10, pp. 2361–2371, 1995.

[3] J. Gosme, C. Richard, and P. Gonçalvès, "Adaptive diffusion of time-frequency and time-scale representations: a review," *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4136– 4146, 2005.

[4] P. Flandrin, "A time-frequency formulation of optimum detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 9, pp. 1377–1384, 1988.

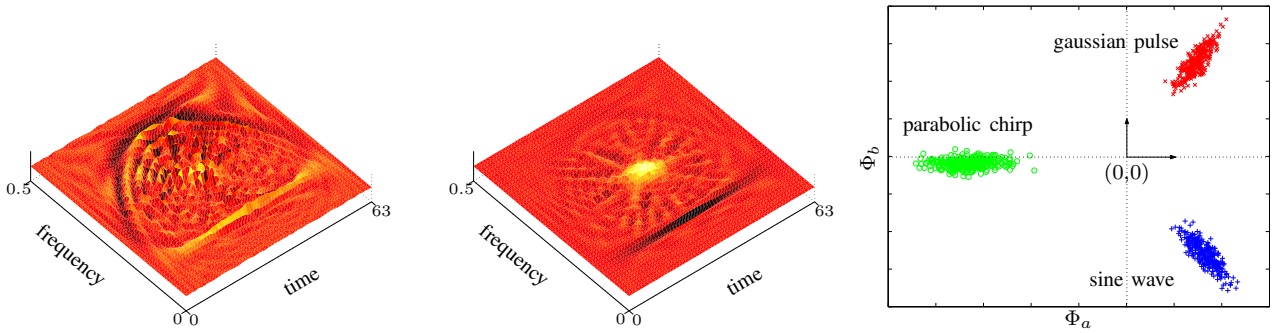

Fig. 5. The time-frequency patterns Φ_a and Φ_b, in (a) and (b) respectively, were extracted by Wigner-based GDA in a three-class classification problem involving sine waves, frequency-modulated waves with a parabolic modulation and Gaussian pulses embedded in noise. In (c), the training data are represented in the $\Phi_a\Phi_b$ plane.

[5] A. Sayeed and D. Jones, "Optimal detection using bilinear time-frequency and time-scale representations," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2872–2883, 1995.

[6] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[7] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[8] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[9] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. Müller, "Fisher discriminant analysis with kernels," in *Advances in neural networks for signal processing*, Y. H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. San Mateo, CA: Morgan Kaufmann, 1999, pp. 41–48.

[10] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[11] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[12] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2001.

[13] M. Davy, A. Gretton, A. Doucet, and P. Rayner, "Optimised support vector machines for nonstationary signal classification," *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 442–445, 2002.

[14] A. Rakotomamonjy, X. Mary, and S. Canu, "Non-parametric regression with wavelet kernels," *Applied Stochastic Models in Business and Industry*, vol. 21, no. 2, pp. 153–163, 2005.

[15] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, pp. 82–95, 1971.

[16] B. Schölkopf, R. Herbrich, and R. Williamson, "A generalized representer theorem," NeuroCOLT, Royal Holloway College, University of London, UK, Tech. Rep. NC2-TR-2000-81, 2000.

[17] B. Boashash, Ed., *Time-frequency signal analysis and applications*. Amsterdam: Elsevier, 2003.

[18] P. Flandrin, *Time-Frequency/Time-Scale Analysis*. San Diego, CA: Academic Press, 1998.

[19] N. Marinovitch, "The singular value decomposition of the wigner distribution and its applications," in *The Wigner distribution: theory and applications in signal processing*, W. Mecklenbräuker and F. Hlawatsch, Eds. Amsterdam: Elsevier, 1997.

[20] E. Bernat, W. Williams, and W. Gehring, "Decomposing ERP time-frequency energy using PCA," *Clinical Neurophysiology*, vol. 116, pp. 1314–1334, 2005.

[21] Z. H. Mamar, P. Chainais, and A. Aussem, "Probabilistic classifiers and time-scale representations: application to the monitoring of a tramway guiding system," in *Proc. European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2006.

[22] C. Richard and R. Lengellé, "Joint recursive implementation of time-frequency representations and their modified version by the reassignment method," *Signal Processing*, vol. 60, no. 2, pp. 163–179, 1997.

[23] E. Chassande-Mottin and A. Pai, "Discrete time and frequency Wigner-Ville distribution: Moyal's formula and aliasing," *IEEE Signal Processing Letters*, vol. 12, no. 7, pp. 508–511, 2005.

[24] F. Abdallah, C. Richard, and R. Lengellé, "An improved training algorithm for nonlinear kernel discriminants," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2798–2806, 2004.

[25] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K. R. Müller, Eds. Cambridge, MA: The MIT Press, 2000, pp. 582–588.

[26] J. Vert, K. Tsuda, and B. Schölkopf, "A primer on kernel methods," in *Kernel Methods in Computational Biology*, B. Schölkopf, K. Tsuda, and J. Vert, Eds. Cambridge, MA: MIT Press, 2004, pp. 35–70.

[27] F. Bach and M. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[28] Y. Engel, S. Mannor, and R. Meir, "Kernel recursive least squares," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

[29] P. Honeiné, C. Richard, P. Flandrin, and J.-B. Pothin, "Optimal selection of time-frequency representations for signal classification: a kernel-target alignment approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006.