Nonstationary signal analysis with kernel machines

Paul Honeine*[(1)], Cédric Richard[(1)], Patrick Flandrin[(2)]

[(1)]Institut Charles Delaunay (FRE CNRS 2848)
Laboratoire de Modélisation et Sûreté des Systèmes (LM2S)
Université de technologie de Troyes, 12 rue Marie Curie, 10010 Troyes, France
paul.honeine@utt.fr tel.: +33.3.25.71.56.25 fax.: +33.25.71.56.99
cedric.richard@utt.fr tel.: +33.3.25.71.58.47 fax.: +33.25.71.56.99

[(2)]Laboratoire de Physique (UMR CNRS 5672)
Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364, Lyon, France
patrick.flandrin@ens-lyon.fr tel.: +33.4.72.72.81.60 fax.: +33.4.72.72.80.80

NONSTATIONARY SIGNAL ANALYSIS WITH KERNEL MACHINES

ABSTRACT

This chapter introduces machine learning for nonstationary signal analysis and classification. It argues that machine learning based on the theory of reproducing kernels can be extended to nonstationary signal analysis and classification. The authors show that some specific reproducing kernels allow pattern recognition algorithm to operate in the time-frequency domain. Furthermore, the authors study the selection of the reproducing kernel for a nonstationary signal classification problem. For this purpose, the kernel-target alignment as a selection criterion is investigated, yielding the optimal time-frequency representation for a given classification problem. These links offer new perspectives in the field of nonstationary signal analysis, which can benefit from recent developments of statistical learning theory and pattern recognition.

INTRODUCTION

Time-frequency and time-scale distributions have become increasingly popular tools for analysis and processing of nonstationary signals. These tools map a one-dimensional signal into a two-dimensional distribution, a function of both time and frequency. Such joint description reveals the time-varying frequency content of nonstationary signals, unlike classical spectral analysis techniques *a la Fourier*. Over the years, a large variety of classes of time-frequency distributions have been proposed to explain the diversity of the treated problems. Linear and quadratic distributions have been extensively studied, and among them Cohen's class of time-frequency distributions as (quasi) energy distribution jointly in time and frequency, and most notably the Wigner-Ville distribution, see for instance (Cohen, 1989; Flandrin, 1999; Auger & Hlawatsch, 2008). From these, one can choose the optimal representation for the problem under investigation, such as increasing the representation immunity to noise and interference components (Auger & Flandrin, 1995, Baraniuk & Jones, 1993), or selecting the best class of representations for a given decision problem (Heitz, 1995; Till & Rudolph, 2000; Davy, Doncarly, & Boudreaux-Bartels, 2001).

Over the last decade, multiple analysis and classification algorithms based on the theory of reproducing kernel Hilbert space (RKHS) have gained wide popularity. These techniques take advantage of the so-called *kernel trick*, which allows construction of nonlinear techniques based on linear ones. Initiated by state-of-the-art support vector machines (SVM) for classification and regression (Vapnik, 1995), the most popular ones include the nonlinear generalization of principal component analysis or kernel-PCA (Schölkopf, Smola, & Müller, 1998), and nonlinear Fisher's discriminant analysis or kernel-FDA (Mika, Rätsch, Weston, Schölkopf, & Müller, 1999); see (Shawe-Taylor & Cristianini, 2004) for a survey of kernel machines. Kernel machines are computationally attractive, with outstanding performance, validated theoretically by the statistical learning theory (Vapnik, 1995; Cucker & Smale, 2002). Despite these advances, nonstationary signal analysis and classification still has not benefited from these developments, although such techniques have been brought to the attention of the signal processing community. Few work combine kernel machines and time-frequency analysis, of these Davy *et al.* (2002) apply the SVM algorithm for classification with a reproducing kernel expressed in the time-frequency domain. More recently, Honeine *et al.* (2007) applied a large panel of kernel machines for nonstationary signal analysis and classification, while in Honeine *et al.* (2006) and in Honeine and Richard (2007) the optimality of the representation space is treated.

This chapter shows how the most effective and innovative kernel machines can be configured, with a proper choice of reproducing kernel, to operate in the time-frequency domain. Further, this approach is extended to the selection of the optimal time-frequency domain for a given classification task. For this purpose, the strength of the kernel-target alignment criterion is investigated for time-frequency distributions. The performance of the proposed approach is illustrated with simulation results. But before, a brief review of the principal elements of the theory behind kernel machines is presented.

RKHS AND KERNEL MACHINES: A BRIEF REVIEW

The theory behind RKHS serves as a foundation of the kernel machines. The main building blocks of these statistical learning algorithms are the *kernel trick* and the Representer Theorem. In this section, these concepts are presented succinctly, after a short introduction on reproducing kernels.

REPRODUCING KERNELS AND RKHS

Let $\mathcal{X}$ be a subspace of $\mathcal{L}_2(\mathbb{C})$ the space of finite-energy complex signals, equipped with the usual inner product defined by $\langle x_i, x_j \rangle = \int_t x_i(t)\, x_j^*(t)\, dt$ and its corresponding norm, where $x_j^*(t)$ denotes the complex conjugate of the signal $x_j(t)$. A kernel is a function $\kappa(x_i, x_j)$ from $\mathcal{X}\times\mathcal{X}$ to C, with Hermitian symmetry. The basic concept of reproducing kernels is described by the following two definitions (Aronszajn, 1950).

Definition 1. A kernel $\kappa(x_i, x_j)$ is said to be positive definite on $\mathcal{X}$ if the following is true:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j^* \kappa(x_i, x_j) \geq 0 \tag{1}$$

for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{C}$.

Definition 2. Let $(\mathcal{H}, \langle \cdot\,, \cdot \rangle_{\mathcal{H}})$ be a Hilbert space of functions from $\mathcal{X}$ to C. The function $\kappa(x_i, x_j)$ from $\mathcal{X}\times\mathcal{X}$ to C is the reproducing kernel of $\mathcal{H}$ if, and only if,

1. the function $\kappa_{x_j} : x_i \mapsto \kappa_{x_j}(x_i) = \kappa(x_i, x_j)$ belongs to $\mathcal{H}$, for all $x_j \in \mathcal{X}$;
2. one has $\psi(x_j) = \langle \psi, \kappa_{x_j} \rangle_{\mathcal{H}}$ for all $x_j \in \mathcal{X}$ and $\psi \in \mathcal{H}$.

It can be shown that every positive definite kernel κ is the reproducing kernel of a Hilbert space of functions from $\mathcal{X}$ to C. It suffices to consider the space $\mathcal{H}_0$ induced by the functions $\{\kappa_x\}_{x\in\mathcal{X}}$, and equip it with the inner product

$$\langle \psi, \phi \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j^* \kappa(x_i, x_j) \tag{2}$$

where $\psi = \sum_{i=1}^{n} a_i \kappa_{x_i}$ and $\phi = \sum_{j=1}^{m} b_j \kappa_{x_j}$ are elements of $\mathcal{H}_0$. Then, this incomplete Hilbertian space is completed according to (Aronszajn, 1950), so that every Cauchy sequence converges in that space. Thus, one obtains the Hilbert space $\mathcal{H}$ induced by the reproducing kernel κ, usually called a *reproducing kernel Hilbert space*. It can also be shown that every reproducing kernel is positive

definite (Aronszajn, 1950). A classic example of a reproducing kernel is the Gaussian kernel $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma^2)$, where σ is a tunable parameter corresponding to the kernel bandwidth. Other examples of reproducing kernels, and rules for designing and combining them can be found in (Vapnik, 1995; Shawe-Taylor & Cristianini, 2004).

THE KERNEL TRICK, THE REPRESENTER THEOREM

Substituting ψ by κ_{x_i} in item 2 of Definition 2, one gets the following fundamental property of RKHS

$$\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}} \tag{3}$$

for all $x_i, x_j \in \mathcal{X}$. Therefore, $\kappa(x_i, x_j)$ gives the inner product in $\mathcal{H}$, the so-called *feature space*, of the images κ_{x_i} and κ_{x_j} of any pair of input data x_i and x_j, without having to evaluate them explicitly. This principle is called the *kernel trick*. It can be used to transform any linear data processing technique into a nonlinear 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 nonlinear 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 is a solid foundation of kernel machines for pattern recognition, such as SVM, kernel-PCA, and kernel-FDA (Schölkopf, Herbrich, & Smola, 2001).

Theorem (Representer Theorem). Any function ϕ of $\mathcal{H}$ minimizing a regularized cost function of the form

$$J\big((x_1, y_1, \phi(x_1)), \ldots, (x_n, y_n, \phi(x_n))\big) + \rho\big(\|\phi\|^2_{\mathcal{H}}\big) \tag{4}$$

with ρ a strictly monotonic increasing function on $\mathbb{R}_+$, can be written as a kernel expansion in terms of the available data, namely

$$\phi(\cdot) = \sum_{j=1}^{n} a_j \, \kappa(\cdot, x_j). \tag{5}$$

Sketch of proof. Any function φ of the space $\mathcal{H}$ can be decomposed as $\varphi(\cdot) = \sum_{j=1}^{n} a_j \kappa(\cdot, x_j) + \phi_\perp(\cdot)$, where $\langle \phi_\perp(\cdot), \kappa(\cdot, x_j) \rangle_{\mathcal{H}} = 0$ for all $j = 1, \ldots, n$. Using this with (3), it is obvious that $\phi_\perp$ does not affect the value of $\varphi(x_i)$, for all $i = 1, \ldots, n$. Moreover, one can verify that the n-order model defined in (5) minimizes ρ since $\rho(\|\varphi\|^2_{\mathcal{H}}) = \rho(\|\phi\|^2_{\mathcal{H}} + \|\phi_\perp\|^2_{\mathcal{H}}) \geq \rho(\|\phi\|^2_{\mathcal{H}})$ with equality only if $\varphi = \phi$. This is the essence of the Representer Theorem.

□

TIME-FREQUENCY KERNEL MACHINES: THE WIGNER-VILLE DISTRIBUTION

In this section, we investigate the use of kernel learning machines for pattern recognition in the time-frequency domain, by taking advantage of both the kernel trick and the Rerpesenter Theorem. To clarify the discussion, we will first focus on the Wigner-Ville distribution. This will be followed by an extension to other time-frequency distributions, linear and quadratic. Below, $\mathcal{A}_n$ denotes a training set of n instances $x_i \in \mathcal{X}$ and the desired outputs or labels $y_i \in \mathcal{Y}$, with $\mathcal{Y} = \{\pm 1\}$ for a binary (2 classes) classification problem.

THE WIGNER-VILLE DISTRIBUTION

The Wigner-Ville distribution is considered fundamental among the large class of time-frequency representations. This is mainly due to many desirable theoretical properties such as the correct marginal conditions for instance, as well as the unitary condition. The latter propells this distribution into a suitable candidate for detection based on the time-frequency domain. The Wigner-Ville distribution of a finite energy signal $x(t)$ is given by

$$W_x(t,f) = \int x(t+\tau/2)\, x^*(t-\tau/2)\, e^{-2j\pi f\tau}\, d\tau. \tag{6}$$

Consider applying conventional linear pattern recognition algorithms directly to time-frequency representations. This means that one should optimize a given criterion J of the general form (4), where each signal x_i, for $i=1,2,\ldots,n$, is substituted by its time-frequency distribution W_{x_i}. Therefore, one seeks to determine a function ϕ of the form

$$\phi(x) = \langle W_x, \Phi \rangle = \int\int W_x(t,f)\; \Phi(t,f)\; dt\, df, \tag{7}$$

or equivalently the time-frequency pattern $\Phi(t,f)$. The main difficulty encountered in solving such problems is that they are typically very high dimensional, the size of the Wigner-Ville distributions calculated from the training set being quadratic in the length of signals, leading to manipulating n representations of size l^2 for signals of length l. This makes pattern recognition based on time-frequency representations time-consuming, if not impossible, even for reasonably-sized signals. With both the kernel trick and the Representer Theorem, kernel machines eliminate this computational burden. For this purpose, one may consider the inner product between the Wigner-Ville distributions, which is given by the kernel

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

Note however that the time-frequency distributions, W_{x_i} and W_{x_j} here, do not need to be computed in order to evaluate κ_W. For this purpose, one may consider

the unitarity of the Wigner-Ville distribution illustrated by Moyal's formula $\langle W_{x_i}, W_{x_j} \rangle = |\langle x_i, x_j \rangle|^2$, yielding

$$\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2 . \tag{9}$$

It is evident that κ_W is a positive definite kernel, since the condition given by Definition 1 is clearly verified as $||\sum_j a_j W_{x_j}||^2 \geq 0$. Therefore, we are now in a position to construct the RKHS induced by this kernel, and denoted by $\mathcal{H}_W$. It is obtained by considering the space $\mathcal{H}_0$ defined below, and complete it with the limit of every Cauchy sequence

$$\mathcal{H}_0 = \{\phi : \mathcal{X} \to \mathbb{R} \mid \phi(\cdot) = \sum_j a_j |\langle \cdot, x_j \rangle|^2, a_j \in \mathbb{R}, x_j \in \mathcal{X}\}. \tag{10}$$

Thus, the kernel (9) can be considered with any kernel machine proposed in the literature to perform pattern recognition tasks in the time-frequency domain. By taking advantage of the Representer Theorem, the solution $\phi(\cdot) = \sum_j a_j |\langle \cdot, x_j \rangle|^2$ allows for a time-frequency distribution interpretation, since it can be written as $\phi(x) = \langle W_x, \Phi_W \rangle$, with the time frequency signature

$$\Phi_W = \sum_{j=1}^{n} a_j W_{x_j} . \tag{11}$$

This expression is directly obtained by combining (5) and (7). One should keep in mind that the n coefficients a_j that determine the solution are estimated without calculating any Wigner-Ville distribution, since only their inner products are required. Subsequently, the time-frequency pattern Φ_W can be determined with (11) in an iterative manner, without suffering the drawback of storing and manipulating a large collection of Wigner-Ville distributions. Moreover, most of the kernel machines speed-up the calculation of the time-frequency pattern Φ_W since a large number of the resulting coefficients a_j is null. This sparsity of the solution made a breakthrough in the machine learning community with the highly-performant SVM algorithms.

EXAMPLE OF TIME-FREQUENCY KERNEL MACHINE: THE WIGNER-VILLE-BASED PCA

In order to emphasize the main idea behind time-frequency kernel machines, the Wigner-Ville-based PCA approach is illustrated here, by considering the Wigner-Ville distribution and the kernel-PCA algorithm. The latter is a nonlinear form of PCA, and allows to extract principal components of variables that are nonlinearly related to the input variables. Given n centered observations $x_1, \ldots, x_n$, a dual formulation of the standard PCA algorithm consists of diagonalizing the Gram matrix K whose (i,j)-th entry is $\langle x_i, x_j \rangle$. The i-th coordinate of the k-th principal component is then given by $\sum_{j=1}^{n} a_{j,k} \langle x_i, x_j \rangle$, where the $a_{j,k}$'s are the components

of the k-th eigenvector of K. In kernel-PCA, a nonlinear transformation of the input data is carried out implicitly by replacing each inner product $\langle x_i, x_j \rangle$ with a kernel function $\kappa(x_i, x_j)$. A major interest of this method is that it performs PCA in feature spaces of arbitrarily large, possibly infinite, dimension. In particular, it can be adjusted to operate on the Wigner-Ville distributions of n signals $x_1, \ldots, x_n$ via an eigendecomposition of the Gram matrix K_W whose (i,j)-th entry is $\kappa_W(x_i, x_j)$. The k-th principal component can then be extracted from any signal x as follows

$$\phi_k(x) = \sum_{j=1}^{n} a_{j,k} \kappa_W(x, x_j), \tag{12}$$

where the $a_{j,k}$'s are the components of the k-th eigenvector of K_W. Now we can state an equivalent time-frequency formulation to the expression above: $\phi_k(x) = \langle W_x, \Phi_k \rangle$, with the signature

$$\Phi_k = \sum_{j=1}^{n} a_{j,k} W_{x_j}. \tag{13}$$

We call Φ_k the k-th principal distribution, although it may not be a valid time-frequency distribution.

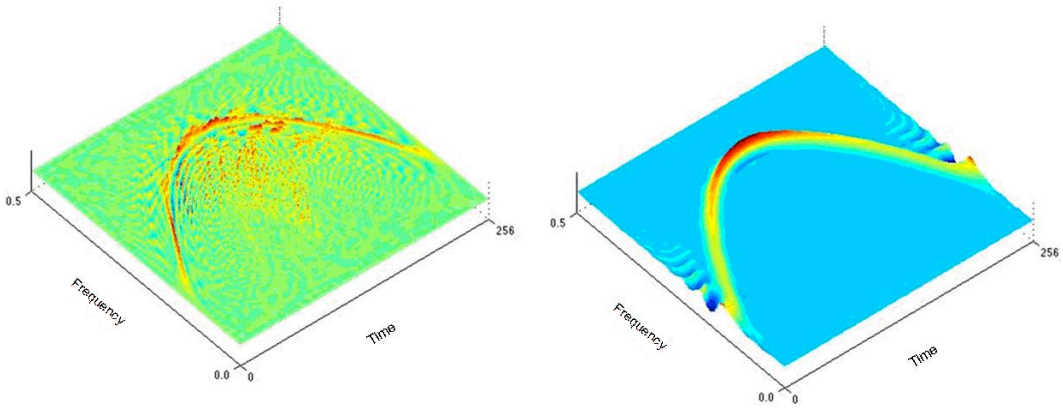

Figure 1: First principal distribution resulting from the Wigner-Ville-based kernel-PCA (left), and the Born-Jordan-based kernel-PCA (right).

To illustrate kernel-PCA as a potentially useful tool in nonstationary signal processing, a collection of 500 signals of 256 samples each is considered. Each signal consists of a quadratic frequency modulation between 0.1 and 0.4 in normalized frequency, corrupted with an additive white Gaussian noise at a signal-to-noise ratio of 0 dB. Figure 1 (left) shows the first principal distribution obtained by applying the kernel-PCA based on the Wigner-Ville time-frequency distribution. This is done by applying the classical kernel-PCA technique with the Wigner-Ville kernel (9), and then injecting the resulting weighting coefficients into (13). As illustrated in the figure, the quadratic frequency modulation and interference terms can be clearly identified in Φ_1. In order to reduce interference components in the resulting principal distribution, one may consider a different time-frequency distribution, as illustrated in Figure 1 (right) with Born-Jordan

distribution. In the following section, we extend the proposed approach to other time-frequency distributions of Cohen's class, whose elements include the Wigner-Ville distribution, the Born-Jordan distribution and the Choï-Williams distribution, only to name a few.

But before it is worth noting the gain in computational complexity by the proposed approach. Applying standard PCA directly to the set of Wigner-Ville distributions would lead to the same result. However, this approach suffers from the high computational cost of calculating, storing and manipulating a set of 500 matrices, each having size 256^2 by 256^2. In addition, it requires calculating and diagonalizing a 256^2 by 256^2 covariance matrix, which is computationally intensive if not impossible. This conclusion remains valid for standard pattern recognition machines such as FDA and GDA.

EXTENSION TO OTHER TIME-FREQUENCY DISTRIBUTIONS

Obviously, the concept of time-frequency machines for pattern recognition is not limited only to the Wigner-Ville distribution. This section illustrates it with other popular time-frequency distributions, linear and quadratic. More particularly, the choice of the optimal representation for a given classification task is studied in next section, with the kernel-target alignment criterion.

LINEAR REPRESENTATIONS

The short-time Fourier transform (STFT) remains the most widely used linear time-frequency distribution. For a given analysis window $w(t)$, well localized around the origin of the time-frequency domain, it is defined for a signal x by

$$F_x(t,f) = \int x(\tau)\, w^*(\tau - t)\, e^{-2\,j\pi f\tau}\, d\tau, \tag{14}$$

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

$$\kappa_F(x_i, x_j) = \|w\|^2 \langle x_i, x_j \rangle . \tag{15}$$

It is obvious that this is a positive definite kernel, since condition (1) takes the form $||\sum_i a_i x_i||^2 \geq 0$ which is trivially satisfied. Therefore, kernel κ_F induces a RKHS and can be used with any kernel machine to operate on the short-time Fourier domain. From the Representer Theorem, equation (5) offers a time-frequency distribution interpretation, as $\phi(x) = \langle F_x, \Phi_F \rangle$ with the time-frequency signature $\Phi_F = \sum_{j=1}^{n} a_j F_{x_j}$. Moreover, since the considered representation is linear with respect to the signal, this is equivalent to $\Phi_F = F_z$ with z being the signal $z = \sum_{j=1}^{n} a_j x_j$. Therefore, one needs to evaluate the short-time Fourier transform only once.

This illustrates the potential use of any kernel machine for pattern recognition with the short-time Fourier representation. Obviously, this is not limited only to this linear representation. For instance, one may consider the Fourier transform, defined by $\hat{x}(f) = \int x(\tau)\, e^{-2\,j\pi f\tau}\, d\tau$ for a signal x. In a more general case, one may also consider a time-scale representation, such as the wavelet transform as illustrated next.

The (continuous) wavelet decomposition is a very appreciated linear representation, relying on a time-translation of t, and a scaling factor a of a mother wavelet w, such as $w_{t,a}(\tau) = a^{-1/2}\, w((\tau - t)/a)$. The wavelet representation of a given signal x can be written as

$$\Omega_x(t,a) = \int x(\tau) \frac{1}{\sqrt{a}} w^*\big((\tau - t)/a\big)\, d\tau .$$

In order to be an inversible transformation, the admissible condition $c_w = \int |\hat{x}(f)|^2 df / |f| < +\infty$ must be satisfied. This not-too restrictive condition on the mother wavelet is verified for instance by the mexican hat, which is given by the second derivative of the Gaussian. This admissibility condition allows an identity similar to (15) with

$$\int\!\!\int \Omega_{x_i}(t,a)\; \Omega^*_{x_j}(t,a) \frac{dt\; da}{a^2} = c_w \int x_i(t)\; x^*_j(t)\; dt$$

where $dt\; da\; /\; a^2$ is a natural measure associated to the translation and scaling (Flandrin, 1999). Therefore, by normalizing the mother wavelet such that $c_w = 1$, we get a unitary transformation.

To summarize, both the short-time Fourier transform and the wavelet transform, as well as the Fourier transform, are linear with respect to the studied signal. Therefore, they share the same reproducing kernel defined by the linear kernel $\kappa(x_i, x_j) = \langle x_i, x_j \rangle$, upto a multiplicative normalization constant. Next, we extend this approach to the quadratic class of time-frequency distributions, as defined by Cohen's class.

QUADRATIC TIME-FREQUENCY DISTRIBUTIONS

The concept of quadratic forms for nonstationary signal analysis is mainly driven by the need to study its energy distribution over both time and frequency, simultaneously. This has produced a large body of theretical work over the years, as many different classes of solutions emerge naturally. From these, the Cohen class of time-frequency distributions gained considerable attention in recent years. These distributions are covariant with respect to time-frequency shifts applied to the signal under scutiny. For a finite energy signal $x(t)$ to be analyzed, a distribution belonging to Cohen class distributions is given by

$$C_x(t,f) = \iint \Pi(t'-t, f'-f)\; W_x(t',f')\; dt'df' \tag{16}$$

where W_x is the Wigner-Ville distribution of the signal x and Π is a two-dimensional weighting function. The latter determines the properties of the distribution. We can easily check that $\kappa_\Pi(x_i,x_j) = \langle C_{x_i}, C_{x_j} \rangle$ is a positive definite kernel. Then it can be used by any kernel machine, leading to the solution $\phi(x) = \langle C_x, \Phi_\Pi \rangle$, with $\Phi_\Pi = \sum_{j=1}^{n} a_j C_{x_j}$. The Π-tunable kernel κ_Π provides a large class of solutions adapted for a given task. For instance, one may require solutions that are relatively immune to interference and noise for analysis purpose. The kernel can also be exploited to improve classification accuracy, by maximizing a contrast criterion between classes.

The spectrogram is probably the most popular distribution of the Cohen class. Defined as the squared magnitude of the short-time Fourier transform (14), it is related to the kernel $\kappa_S(x_i,x_j) = \iint |\langle x_i, w_{t,f} \rangle \langle x_j, w_{t,f} \rangle|^2 dtdf$. Other examples of the Cohen class include distributions verifying the unitary condition, such as the Wigner-Ville distribution, the Rihaczek distribution and the Page distribution. While these distributions share the same kernel (9), a property resulting from Moyal's formula $\langle C_{x_i}, C_{x_j} \rangle = |\langle x_i, x_j \rangle|^2$, they differ by the time-frequency pattern Φ_C. The latter can be computed directly of using Φ_W with

$$\Phi_\Pi(t,f) = \iint \Pi(t'-t, f'-f)\; \Phi_W(t',f')\; dt'df' \tag{17}$$

Examples of the most used time-frequency distributions of Cohen's class are presented in the table below, with there definitions.

Wigner-Ville	$\int x(t+\tau/2)\; x^*(t-\tau/2)\; e^{-2j\pi f\tau} d\tau$
Born-Jordan	$\int \frac{1}{\lvert\tau\rvert} \int_{t-\lvert\tau\rvert/2}^{t+\lvert\tau\rvert/2} x(t'+\tau/2)\; x^*(t'-\tau/2)\; dt'\, e^{-2j\pi f\tau} d\tau$
Rihaczek	$x(t)\; \hat{x}^*(f)\; e^{-2j\pi f t}$
Margeneau-Hill	$\mathrm{Re}\{x(t)\; \hat{x}^*(f)\; e^{-2j\pi f t}\}$
Choï-Williams	$\iint \frac{\sigma}{\lvert\tau\rvert} e^{-2\sigma^2 (t'-t)^2/\tau^2} x(t'+\tau/2)\; x^*(t'-\tau/2)\; e^{-2j\pi f\tau} dt'\, d\tau$
Butterworth	$\iint \frac{\sqrt{\sigma}}{2\lvert\tau\rvert} e^{-\lvert t'\rvert \sigma 2/\lvert\tau\rvert} x(t+t'+\tau/2)\; x^*(t+t'-\tau/2)\; e^{-2j\pi f\tau} dt'\, d\tau$
Spectrogram	$\left\lvert \int x(\tau)\; w^*(\tau-t)\; e^{-2j\pi f\tau} d\tau \right\rvert^2$

Without loss of generality, we illustrate this extension to a particular distribution of Cohen's class, the Born-Jordan distribution. Figure 1 (right) shows the first principal distribution obtained by considering the kernel-PCA algorithm with the reproducing kernel associated to the Born-Jordan distribution, for the same data as

in Figure 1 (left) for the Wigner-Ville distribution. We recall that this time-frequency distribution is defined for a given signal $x(t)$ by

$$BJ_x(t,f) = \int \frac{1}{|\tau|} \int_{t-|\tau|/2}^{t+|\tau|/2} x(t'+\tau/2)\; x^*(t'-\tau/2)\; dt'\, e^{-2j\pi f\tau}\, d\tau,$$

leading to a reduced-interference distribution (Flandrin, 1999). This property is inherited into the principal distribution as illustrated in Figure 1 (right). Table 1 summarizes kernels associated to some time-frequency distributions, linear and quadratic, as illustrated all over this chapter.

While computing the kernel $\kappa_W(x_i, x_j)$ is efficient as illustrated for the Wigner-Ville-based PCA, this is not the case for any arbitrary non-unitary distribution, leading to a time consuming process for training a kernel machine. In applications where computation time is a major issue, a simple heuristic procedure can be considered to derive solutions of the form $\phi(x) = \langle C_x, \Phi_\Pi \rangle$. For this purpose, the kernel machine is trained by considering the easy-to-compute Wigner-Ville kernel $\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2$. By combining the distribution Φ_W resulting from (11) with equation (17) we get the time-frequency feature Φ_C. Clearly, this is a non-optimal strategy when the considered distribution does not satisfy the unitary condition.

	Distribution name	Associated reproducing kernel
linear	Fourier	$\kappa(x_i,x_j) = \langle x_i,x_j \rangle$
	Short-time Fourier (STFT)	$\kappa_F(x_i,x_j) = \|w\|^2 \langle x_i,x_j \rangle$
	Wavelet	$\kappa_{wav}(x_i,x_j) = c_w \langle x_i,x_j \rangle$
quadratic	Wigner-Ville	$\kappa_W(x_i,x_j) = \|\langle x_i,x_j \rangle\|^2$
	Spectrogram	$\kappa_S(x_i,x_j) = \iint \|\langle x_i,w_{t,f} \rangle \langle x_j,w_{t,f} \rangle\|^2 dtdf$
	Page, Rihaczek	$\kappa_W(x_i,x_j) = \|\langle x_i,x_j \rangle\|^2$
	Cohen	$\kappa_\Pi(x_i,x_j) = \langle C_{x_i},C_{x_j} \rangle$

Table 1. Some kernels associated to linear and quadratic time-frequency distributions.

OPTIMAL TIME-FREQUENCY REPRESENTATION: THE KERNEL-TARGET ALIGNMENT CRITERION

In previous sections, we have shown how kernel machines can be configured to operate in the time-frequency domain, with the proper choice of reproducing kernel. This section is devoted to the question of selecting the appropriate time-frequency distribution, and therefore the associated reproducing kernel, for a given classification task. Next, the kernel-target alignment criterion is studied for selecting a time-frequency representation, after this succinct review on different available strategies for kernel selection.

KERNEL SELECTION IN MACHINE LEARNING: A BRIEF REVIEW

Many results from the statistical learning theory emphasize on the crucial role of prior knowledge in machine learning problems. For the large class of kernel machines, the *no free kernel theorem* (Cristianini, Kandola, Elisseeff, & Shawe-Taylor, 2006) shows that no kernel is optimal for all applications, and therefore any prior knowledge must contribute to the choice of the appropriate kernel. For this purpose, many algorithms have been proposed in order to reveal the relationship between the data and their labels.

Cross-validation strategies, as well as *leave-one-out* techniques, are widely used in the literature to evaluate the performance for the model. For a given kernel, the generalization error is estimated by constructing several classifiers from various subsets of the available data, and then validated on the remaining data (Meyer, Leisch, & Hornik, 2003). This strategy is conducted on several candidate kernels, and the optimal kernel corresponds to the one with the smallest estimated error. For a tunable kernel, the optimal value of the tuning parameter is obtained from a grid search. The cross-validation approach turns out to be highly time consuming, since it requires a large number of train-and-validate stages. In order to speed up calculations, schemes to estimate a bound on the generalisation error have been recently proposed, requiring only one classifier per candidate kernel. From these, we recall the VC-dimension bound (Burges, 1998), the radius-margin criterion (Chung, Kao, Sun, Wang, & Lin, 2003), and the generalized approximate cross validation (Wahba, Lin, & Zhang, 2000). Other schemes are studied for instance in (Chapelle, Vapnik, Bousquet, & Mukherjee, 2002) (Duan, Keerthi, & Poo, 2003). While the framework defined by these methods is motivated from a theoretical point of view, their computational complexity is often cumbersome.

It is worth noting that these methods can be easily adapted to the time-frequency domain, as well as their statistical and theoretical properties. Next, a reduced computational complexity criterion is considered, the kernel-target alignment, and its adaptation for selecting time-frequency distributions is studied.

THE KERNEL-TARGET ALIGNMENT CRITERION

Performance of kernel-based pattern recognition methods is essentially influenced by the considered reproducing kernel. For a given algorithm such as the SVM, results obtained from two different kernels are as close as these kernels are similar. For this purpose, the authors of (Cristianini, Shawe-Taylor, Elisseeff, & Kandola, 2001) introduced the alignment as a measure of similarity between two reproducing kernels. Given a learning set $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of n data and their labels, the empirical alignment between two kernels κ_1 and κ_2 is defined by

$$A(\kappa_1,\kappa_2,\mathcal{A}_n) = \frac{\langle K_1,K_2\rangle_F}{\sqrt{\langle K_1,K_1\rangle_F \langle K_2,K_2\rangle_F}}, \tag{18}$$

where K_1 and K_2 the Gram matrices with entries of the form $\kappa_1(x_i, x_j)$ and $\kappa_2(x_i, x_j)$ respectively, and $\langle \cdot, \cdot \rangle_F$ denotes Frobenius inner product defined by $\langle K_1, K_2 \rangle_F = \sum_i \sum_j \kappa_1(x_i, x_j)\, \kappa_2(x_i, x_j)$. The alignment corresponds to a correlation coefficient between both matrices K_1 and K_2.

Recall that for a classification problem, one seeks a decision rule $\phi^*(\cdot)$ satisfying the relation $\phi^*(x_i) = y_i$ for i = 1, …, n. For the particular cas of a 2-class problem, this relation takes the form $\phi^*(x_i) = \pm 1$, depending on which class x_i belongs. The reproducing kernel corresponding to this ideal transformation can be defined by $\kappa^*(x_i, x_j) = y_i\, y_j$. The associated ideal Gram matrix $K^* = \boldsymbol{y}\,\boldsymbol{y}'$, where $\boldsymbol{y}$ denotes the target column vector whose i-th entry is y_i, which leads to

$$K^*(i,j) = \begin{cases} 1 & \text{si } y_i = y_j \\ -1 & \text{si } y_i \neq y_j \end{cases} \tag{19}$$

In what follows, the 2-class classification problem is studied, with K^* as defined above. But before, we emphasize on the simplicity of generalizing the proposed approach for a multi-class problem. For the case of c classes, the targets correspond to the c unit-norm and equiangular vectors, leading to $\kappa^*(x_i, x_j) = 1/(1 - c)$ if x_i and x_j belong to different classes, and $\kappa^*(x_i, x_j) = 1$ otherwise.

In (Cristianini, Shawe-Taylor, Elisseeff, & Kandola, 2001 ; Cristianini, Kandola, Elisseeff, & Shawe-Taylor, 2006), Cristianini *et al.* suggest to use the alignment to measure the similarity between a given reproducing kernel and the ideal target matrix $K^* = \boldsymbol{y}\,\boldsymbol{y}'$, in order to determine its relevance for the classification task in hand. Therefore, one considers the an optimization problem of the form

$$\kappa^* = \arg\max_{\kappa} \frac{\sum_{i,j=1}^{n} y_i y_j\, \kappa(x_i,x_j)}{n\left(\sum_{i,j=1}^{n} (\kappa(x_i,x_j))^2\right)^{1/2}}$$

where the definition of the alignment (18) is considered with $K^* = \boldsymbol{y}\,\boldsymbol{y}'$. The relevance of the alignment criterion is provided by a connection to the error of generalization, as demonstrated in (Cristianini, Shawe-Taylor, Elisseeff, & Kandola, 2001) on a Parzen estimator of the form $g(\cdot) = \frac{1}{n}\sum_{i=1}^{n} a_i\, \kappa(\cdot, x_i)$. Since the first study of Cristianini *et al.*, the concept of maximizing the kernel-target alignment has been extended to other problems, including regression problems (Kandola, Shawe-Taylor, & Cristianini, On the extensions of kernel alignment, 2002), metric learning (Wu, Chang, & Panda, 2005 ; Lanckriet, Cristianini, Bartlett, Ghaoui, & Jordan, 2002), as well as sudying the optimal combination of kernels in order to increase the performance of the classifier (Kandola, Shawe-Taylor, & Cristianini, 2002).

OPTIMAL TIME-FREQUENCY DISTRIBUTION

One should emphasize that the kernel-target alignment criterion does not require any learning of the decision rule. For this purpose, we investigate this approach to optimize a time-frequency representation for a classification task involving nonstationary signals. In a decisional framework, conventional strategies for determining the proper time-frequency representation mainly consist in selecting the one which yields the smallest estimated classification error (Heitz, 1995 ; Atlas, Droppo, & McLaughlin, 1997 ; Davy & Doncarli, 1998). However, this empirical approach requires multiple phases of learning the rule and cross-validating it. As we have seen so far, it is obvious that kernel machines for pattern recognition provide a unified context for solving a wide variety of statistical modelling and function estimation, for nonstationary signal analysis and classification.

The proposed approach involves kernels associated to time-frequency distributions, as defined in Table 1, and more particularly Cohen's class with the Π-tunable distributions. Therefore, the kernel-target alignment criterion can be written as

$$\frac{\sum_{i,j=1}^{n} y_i y_j \kappa_{\Pi}(x_i,x_j)}{n\,(\sum_{i,j=1}^{n}(\kappa_{\Pi}(x_i,x_j))^2)^{1/2}}, \tag{20}$$

and maximizing this score leads to the optimal time-frequency distribution, with optimality relative to the given classification task, defined by the available learning set of signals and their labels. The relevance of this approach is illustrated in next section, for nonstationary signals.

SIMULATION RESULTS

In this section, we study the relevance of the approach proposed in this chapter. Previously, we have illustrated the case of the kernel-PCA with the Wigner-Ville distribution. In Figure 1, we have shown the resulting principal time-frequency distributions for a conventional class of frequency modulated signals. In this section, we study learning a decisional problem for a classification/discrimination task, involving nonstationary signals.

NONSTATIONARY SIGNAL CLASSIFICATION: SVM AND KERNEL-FDA

This first set of experiments concerns a discrimination problem between two classes of 64-sample signals embedded in a white Gaussian noise of variance 1.25. The first class consists of 500 signals with a quadratic frequency modulation between 0.1 and 0.4 in normalized frequency, and the second one of 500 signals

with a Gaussian time-frequency atom in frequency at 0.1 and in time at the middle of the signals.

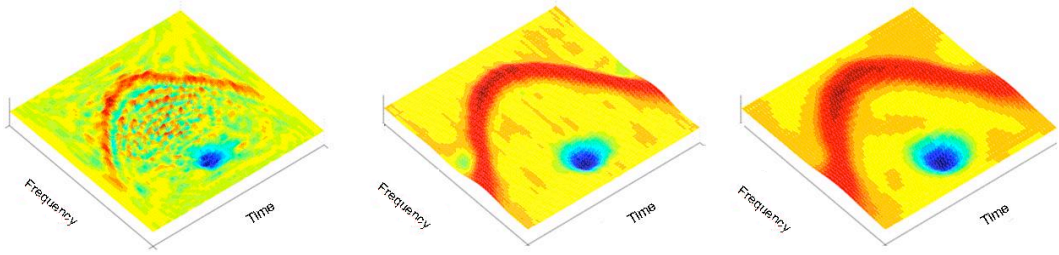

Figure 2: The resulting distributions obtained from the kernel-FDA algorithm with reproducing kernels associated to Wigner-Ville (left), the Choï-Williams (middle), and the spectrogram (right).

We apply the kernel-FDA algorithm with the quadratic kernel associated to the Wigner-Ville distribution. Figure 2 (left) presents the resulting distribution, where we get the two key components for the discrimination, on the one hand the quadratic frequency modulation with a positive orientation (red color) and on the other hand the Gaussian atom represented with a negative orientation (blue color). Moreover, some inference components are present in this signature, an inherent property of the Wigner-Ville distribution. In order to get rid of these interference components, one may consider other time-frequency distributions from Cohen's class, and use the associated reproducing kernel. For this purpose, we use both the Choï-Williams time-frequency distribution and the spectrogram, with their respective reproducing kernels. The resulting time-frequency signatures are illustrated in Figure 2 (middle) and Figure 2 (right), respectively. For each case, we found again the quadratic frequency modulation and the Gaussian atom, each in an opposed orientation. As opposed to the one obtained from the Wigner-Ville distribution, these signatures have reduced interferences. However, the price to pay is the relative high computational burden for evaluating the associated reproducing kernel. As illustrated in Figure 2, the proposed approach allows a time-frequency interpretation as opposed to the conventional linear techniques.

Moreover, we can study the relevance of several reproducing kernels for a given nonstationary signal classification problem, and using different learning algorithms. Two learning algorithms are investigated, the kernel-FDA and the classical SVM algorithms. The latter determines a separating hyperplan, with maximal-margin between the classes of the training set. This is mainly motivated by the statistical learning theory (Vapnik, 1995), and we wish to take advantage of this by applying it to the time-frequency domain, using an appropriate reproducing kernel as illustrated in this chapter. Therefore, we seek a time-frequency signature maximizing the distance with the time-frequency distributions of the training signals. For experiments, we consider the same discrimination problem as above, between quadratic frequency modulation signals and Gaussian atom signals. In order to compare the resulting estimated error, signals are corrupted further, with a signal-to-noise ratio of -10 dB. For now, we use the quadratic reproducing kernel associated to the Wigner-Ville distribution, and show in Figure 3 the resulting signatures for kernel-FDA and SVM algorithms.

The latter shows better performance that the former, as illustrated by comparing Figure 3 (right and left), a property resulting from the inherent regularized property of the SVM.

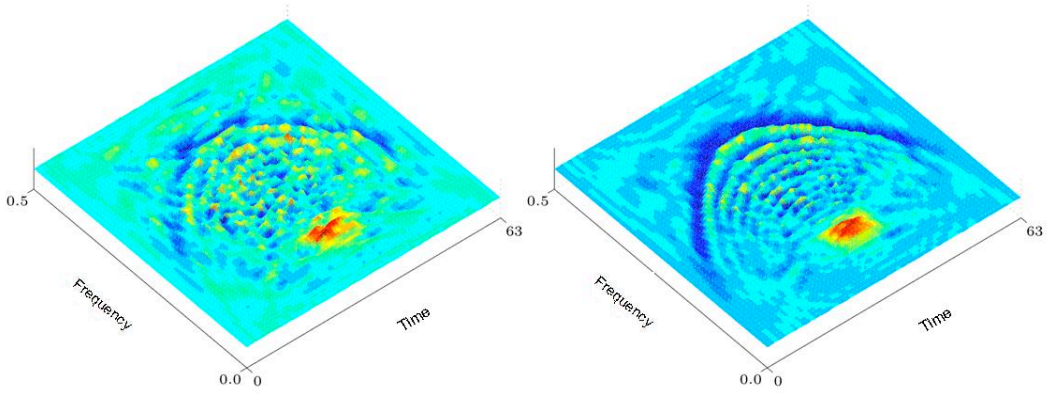

Figure 3: Time-frequency distributions obtained with the quadratic Wigner-Ville reproducing kernel, using the kernel-FDA algorithm (left) and the classical SVM algorithm (right).

In order to compare the performance of several time-frequency distributions, we estimate the generalization error from a set of 10 000 signals belonging to both classes. We show in Table 2 the resulting estimated errors, even though the regularization parameter is not optimaly tuned. The Wigner-Ville distribution is the best for the given classification task. The spectrogram and other *smooth* distributions are less adapted for this problem. This leads to the open question: how to choose the optimal distribution, without the computational burden of validating with a large set of test signals ? This is accessible with the kernel-target alignment criterion, illustrated next for different time-frequency distributions.

Time-frequency distribution	Kernel-FDA	SVM
Spectrogram	28.0	19.8
Smoothed Wigner-Ville	26.5	16.4
Choï-Williams	26.5	16.8
Wigner-Ville	19.3	13.5

Table 2: Estimated generalization error (in %) for several time-frequency distributions, using the kernel-FDA and the SVM algorithms.

OPTIMAL TIME-FREQUENCY DISTRIBUTION

Consider the classification of two families of signals of the form $e^{2j_{\pi\theta}(t)}$ corrupted by an additive white Gaussian noise, where $\theta(t)$ is a linear phase modulation for the first class, increasing from 0.1 to 0.4 normalized frequencies, and quadratic for the second class, between .1 and .4 normalized frequencies. It is worth noting that

signals from both classes share the same spectral support, which is not true in the time-frequency domain. For the learning task, consider 500 signals of length 256 for each family. On the one hand, the kernel-target alignment is computed from this training set, as given in expression (20) for different time-frequency distributions. On the other hand, these results are confronted with the performance of classifiers obtained from the same training set. For a given reproducing kernel, two classifiers are constructed, based independently on the SVM and on the kernel-FDA algorithms, and their generalization errors are estimated from a set of 20 000 signals. Figure 4: Error rate of a SVM classifier and a KFD classifier for different kernels associated to time-frequency distributions. The corresponding kernel-target alignment score is given between parentheses and determines the size of the dot. shows the performance of the kernels associated to the Wigner-Ville (wv), Margenau-Hill (mh), Choï-Williams (cw), Born-Jordan (bj), reduced interference with Hanning window (ridh), spectrogram (sp), and Butterworth (bu) distributions[1]. It shows the relationship between the error rate for both classifiers and the kernel-target alignment score. Moreover, the ease with which the latter can be estimated using only training data, prior to any computationally intensive training, makes it an interesting tool for time-frequency distribution selection.

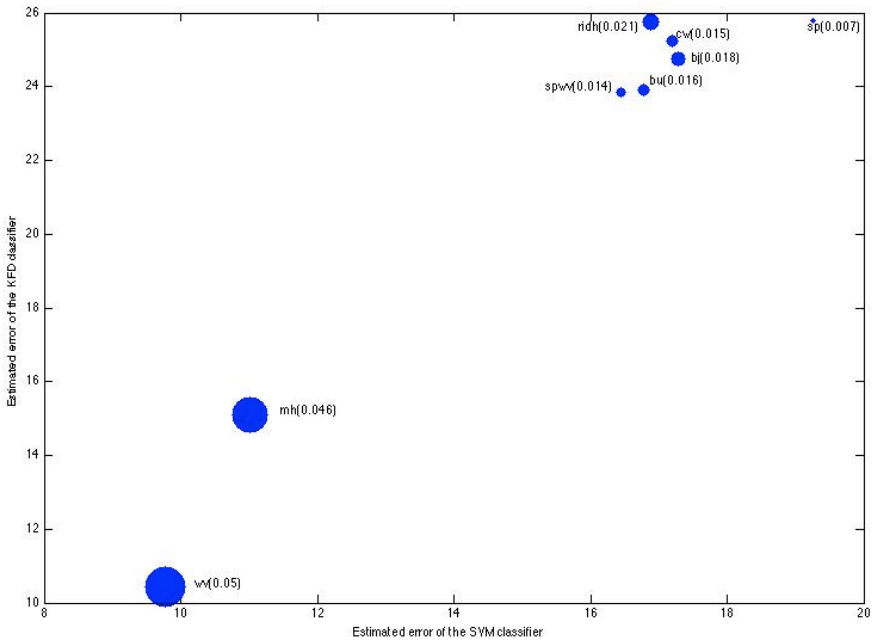

Figure 4: Error rate of a SVM classifier and a KFD classifier for different kernels associated to time-frequency distributions. The corresponding kernel-target alignment score is given between parentheses and determines the size of the dot.

[1] For tuning parameters, we consider those given by default in the Time-Frequency Toolbox.

CONCLUSION

In this chapter, the authors presented the time-frequency kernel machines, a new framework for nonstationary signal analysis and processing. It is shown that pattern recognition algorithms based on the reproducing kernels can operate in the time-frequency domain, with some specific kernels. The choice of the proper kernel for a given classification task is studied from the kernel-target criterion, yielding new and efficient techniques for optimal time-frequency distribution selection. All these links offer new perspectives in the field of nonstationary signal analysis since they provide an access to the most recent developments of pattern recognition and statistical learning theory.

REFERENCES

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society , 68*, 337-404.

Atlas, L., Droppo, J., & McLaughlin, J. (1997). Optimizing time-frequency distributions for automatic classification. *SPIE-The International Society for Optical Engineering, , 3162*, 161-171.

Auger, F., & Flandrin, P. (1995). Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing , 43* (5), 1068-1089.

Auger, F., & Hlawatsch, F. (2008). *Time-Frequency Analysis.* Wiley-ISTE.

Baraniuk, R. G., & Jones, D. L. (1993). A signal-dependent time-frequency representation: Optimal kernel design. *IEEE Transactions on Signal Processing , 41*, 1589-1602.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *data Min. Knowl. Discov. , 2* (2), 121-167.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning , 46* (1-3), 131-159.

Chung, K.-M., Kao, W.-C., Sun, C.-L., Wang, L.-L., & Lin, C.-J. (2003). Radius margin bounds for support vector machines with the RBF kernel. *Neural Comput., 15* (11), 2643–2681.

Cohen, L. (1989). Time-frequency distributions-a review. *Proceedings of the IEEE , 77* (7), 941-981.

Cristianini, N., Kandola, J., Elisseeff, A., & Shawe-Taylor, J. (2006). On kernel target alignment. In *Innovations in Machine Learning* (Vol. 194, pp. 205-256). Springer Berlin / Heidelberg.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. (2001). On kernel-target alignment. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Ed.), *Neural Information Processing Systems (NIPS) 14* (pp. 367–373). MIT Press.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society , 39* (1), 1-49.

Davy, M., & Doncarli, C. (1998). Optimal kernels of time-frequency representations for signal classification. *IEEE International Symposium on TFTS* , 581-584.

Davy, M., Doncarly, C., & Boudreaux-Bartels, G. (2001). Improved optimization of time-frequency based signal classifiers. *IEEE Signal Processing Letters , 8* (2), 52-57.

Davy, M., Gretton, A., Doucet, A., & Rayner, P. W. (2002). Optimised support vector machines for nonstationary signal classification. *IEEE Signal Processing Letters , 9* (12), 442-445.

Duan, K., Keerthi, S., & Poo, A. (2003). Evaluation of simple performance measures for tuning svm hyperparameters. *Neurocomputing , 51*, 41-59.

Flandrin, P. (1999). *Time-Frequency/Time-Scale Analysis.* San Diego: Academic Press.

Heitz, C. (1995). Optimum time-frequency representations for the classification and detection. *Applied Signal Processings , 3*, 124-143.

Heitz, C. (1995). Optimum time-frequency representations for the classification and detection of signals. *Applied Signal Proceedings , 3*, 124-143.

Honeine, P., & Richard, C. (2007). Signal-dependent time-frequency representations for classification using a radially gaussian kernel and the alignment criterion, *Proc. of IEEE Statistical Signal Processing Workshop*. Madison (WI), USA.

Honeine, P., Richard, C., & Flandrin, P. (2007). Time-frequency kernel machines. *IEEE Trans. Signal Processing*, *55*, 3930-3936.

Honeine, P., Richard, C., Flandrin, P., & Pothin, J.-B. (2006). Optimal selection of time-frequency representations for signal classification: a kernel-target alignment approach, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. Toulouse, France.

Kandola, J., Shawe-Taylor, J., & Cristianini, N. (2002). *On the extensions of kernel alignment.* University of London, Department of Computer Science, London.

Kandola, J., Shawe-Taylor, J., & Cristianini, N. (2002). *Optimizing kernel alignment over combinations of kernels.* University of London, Department of Computer Science, London.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L. E., & Jordan, M. (2002). Learning the kernel matrix with semi-definite programming. *Proc. 19th International Conference on Machine Learning*, (pp. 323-330).

Meyer, D., Leisch, F., & Hornik, K. (2003). The Support Vector Machine under Test. *Neurocomputing , 55*, 169-186.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K. R. (1999). Fisher discriminant analysis with kernels. In Y. H. Hu, J. Larsen, E. Wilson, & S.

Douglas (Ed.), *Advances in neural networks for signal processing* (pp. 41-48). San Mateo, CA: Morgan Kaufmann.

Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. *14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory (COLT '01/EuroCOLT '01)* (pp. 416-426). London, UK: Springer-Verlag.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation , 10* (5), 1299-1319.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge Uniersity Press.

Till, M., & Rudolph, S. (2000). Optimized time-frequency distributions for signal classification with feed-forward neural networks. *Applications and science of computational intelligence III. 4055*, pp. 299-310. Orlando: SPIE proceedings series.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York, USA: Springer-Verlag.

Wahba, G., Lin, Y., & Zhang, H. (2000). Generalized approximate cross validation for support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, & C. Schuurmans (Ed.), *Advances in Large Margin Classifiers* (pp. 297-309). Cambridge: MIT Press.

Wu, G., Chang, E. Y., & Panda, N. (2005). Formulating distance functions via the kernel trick. *11th ACM International conference on knowledge discovery in Data Mining*, (pp. 703-709).

KEY TERMS AND THEIR DEFINITIONS

Positive definite kernel
A two-variable function defined on $\mathcal{X}$ that satisfies $\sum_{i,j} a_i a_j^* \kappa(x_i, x_j) \geq 0$ for all $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathrm{C}$.

Reproducing kernel Hilbert space
A Hilbert space of functions from $\mathcal{X}$ to C that possesses a reproducing kernel, i.e. a (positive definite) kernel $\kappa(x_i, x_j)$ with the properties: (1) $\kappa_x = \kappa(x, \cdot)$ belongs to that space, and (2) $\langle \psi, \kappa_x \rangle = \psi(x)$, for all $x \in \mathcal{X}$ and $\psi \in \mathcal{H}$

Kernel-PCA
A nonlinear extension of the classical Principal Component Analysis algorithm based on the kernel paradigm, yielding a powerful feature extraction technique.

Kernel-target alignment criterion
A criterion to select and tune a kernel for a given learning task, prior to any leaning, by comparing it to an ideal kernel obtained from the available training data.

Wigner-Ville distribution
A High resolution joint time-frequency distribution for nonstationary signals analysis, defined by $W_x(t,f) = \int x(t+\tau/2)\, x^*(t-\tau/2)\, e^{-2j\pi f\tau}\, d\tau$ for a given signal x.

Cohen's class of time-frequency distributions
The class of distributions covariant with respect to time and frequency shifts applied to the studied signal. For a given signal x, a distribution belonging to Cohen's class is given by $C_x(t,f) = \iint \Pi(t'-t, f'-f)\, W_x(t', f')\, dt' df'$ where W_x is the Wigner-Ville distribution of x and Π is a tunable function.

Short-time Fourier transform
A linear time-frequency representation of a signal, defined by $F_x(t,f) = \int x(\tau) w^*(\tau - t) e^{-2j\pi f\tau}\, d\tau$ for a given analysis window w.

Wavelet representation
A linear time-frequency representation relying on a time-translation and a scaling of a mother wavelet w, and defined by $\Omega_x(t,a) = \int x(\tau)\, a^{-1/2} w^*\left((\tau - t)/a\right) d\tau$

BIOGRAPHY

Paul Honeine received the Dipl.-Ing. degree in mechanical engineering in 2002 and the M.Sc. degree in industrial control in 2003, both from the Faculty of Engineering, the Lebanese University, Lebanon. In 2007, he received the Ph.D. degree in System Optimisation and Security from the University of Technology of Troyes, France, and was a Postdoctoral Research associate with the Systems Modeling and Dependability Laboratory, from 2007 to 2008. Since September 2008, he has been an assistant Professor at the University of Technology of Troyes, France. His research interests include nonstationary signal analysis, nonlinear adaptive filtering, machine learning, and wireless sensor networks.

Cédric Richard received the M.S. degree in 1994 and the Ph.D. degree in 1998 from Compiègne University of Technology, France, in electrical and computer engineering. Since 2003, he is a Professor at the Systems Modelling and Dependability Laboratory, Troyes University of Technology. His current research interests include statistical signal processing and machine learning. Dr. Richard is the author of over 100 papers. He is a Senior Member of the IEEE, and serves as an associate editor of the IEEE Transactions on SP since 2007. He is a member of the SPTM technical committee of the IEEE SP society since 2009.

Patrick Flandrin is currently with the Physics Department at Ecole Normale Supérieure de Lyon, where he works as a CNRS senior researcher. His research interests include mainly nonstationary signal processing at large (with emphasis on time-frequency and time-scale methods), scaling processes and complex systems. He was awarded the Philip Morris Scientific Prize in Mathematics in 1991, the SPIE Wavelet Pioneer Award in 2001and the Michel Monpetit Prize from the French Academy of Sciences in 2001. He is a Fellow of the IEEE since 2002.