Distribution temps-fréquence à paramétrisation radialement Gaussienne optimisée pour la classification

Paul Honeine, Cédric Richard

Institut Charles Delaunay (FRE CNRS 2848) – Laboratoire de Modélisation et Sûreté des Systèmes
Université de Technologie de Troyes, 12 rue Marie Curie, BP 2060, 10010 Troyes, France
paul.honeine@utt.fr, cedric.richard@utt.fr

Résumé –
Cet article traite de l'optimisation des distributions temps-fréquence pour la résolution de problèmes de classification de signaux. On s'intéresse en particulier à la distribution à fonction de paramétrisation radialement Gaussienne, que l'on ajuste par optimisation de l'alignement noyau-cible. Initialement développé pour la sélection de noyau reproduisant en Machine Learning, ce critère présente l'intérêt de ne nécessiter aucun cycle d'apprentissage. On montre que l'on peut obtenir la fonction de paramétrisation radialement Gaussienne maximisant celui-ci en détournant une technique classique de réduction de termes interférentiels dans les représentations temps-fréquence. On illustre l'efficacité de cette approche à l'aide d'expérimentations.

Abstract –
In this paper, we design optimal time-frequency distributions with radially Gaussian kernel for classification. Our approach is based on the kernel-target alignment criterion, which has been investigated within the framework of Machine Learning, for selecting optimal reproducing kernels. One of its main interests is that it does not need any computationally intensive training stage and cross-validation process. We take advantage of this criterion to tune time-frequency distributions, and consider a classical optimization technique usually used for reducing their interference terms. We illustrate our approach with some experiments.

1 Introduction

Les distributions temps-fréquence et temps-échelle fournissent des outils puissants destinés à l'analyse des signaux non-stationnaires. Toutes déclinées de la distribution de Wigner, les distributions de la classe de Cohen offrent en particulier une variété d'espaces de représentation à la mesure des objectifs visés par les utilisateurs. Toutefois, ceci passe nécessairement par un choix soigneux de la fonction de paramétrisation de la distribution selon le type de signal analysé, les propriétés voulues et l'application visée. Les distributions temps-fréquence à fonction de paramétrisation radialement Gaussienne peuvent satisfaire à un vaste choix d'applications. On peut chercher par exemple à améliorer la lisibilité d'une représentation par une réduction de ses termes interférentiels [1, 2]. On peut encore en ajuster les paramètres afin de faciliter la résolution d'un problème de détection ou de classification de signaux non-stationnaires [3, 4].

A l'exception de quelques travaux récents [5, 6], ces différentes approches exploitent peu les récentes avancées des méthodes de reconnaissance des formes à noyau reproduisant. Initiées par les travaux de Vapnik sur les Support Vector Machines (SVM) pour la classification et la régression dans [7, 8], celles-ci regroupent à présent un vaste choix de techniques non-linéaires de traitement de données, de l'analyse en composantes principales [9] aux analyses discriminantes de Fisher [10] et généralisée [11], en passant par la séparation de sources [12]. Ces méthodes sont particulièrement attractives en raison de leur complexité algorithmique réduite, et parce qu'elles profitent des nouvelles avancées de la théorie statistique de l'apprentissage.

A caractère plus générique que [5, 6], cet article commence par rappeler que les méthodes à noyau les plus performantes et les plus diverses peuvent être mises en œuvre dans le plan temps-fréquence grâce à un choix approprié de noyau reproduisant, une stratégie que nous avons récemment développée dans [13, 14]. La sélection d'une représentation temps-fréquence adaptée à la résolution d'un problème de classification demeure toutefois une question récurrente. Nous avons montré que le domaine de la reconnaissance des formes peut lui apporter des éléments de réponse intéressants grâce aux techniques de sélection de noyau reproduisant en général, notamment avec le critère d'alignement noyau-cible [15]. Ce dernier présente l'intérêt de conduire à la sélection *a priori* d'un espace de représentation des observations, sans nécessiter la répétition de coûteuses phases d'apprentissage de règles de décision suivies de validation croisée. Nous avons étudié différentes applications potentielles dans ce même article, en particulier la sélection d'une distribution optimale parmi un ensemble de candidates.

Dans cet article, on utilise le critère d'alignement noyau-cible pour élaborer une représentation temps-fréquence optimum à paramétrisation radialement Gaussienne en vue de résoudre un problème de classification. Il s'avère que le problème d'optimisation résultant est similaire à celui étudié par Baraniuk et Jones dans [1], pour la réduction des termes interférentiels dans un contexte

d'analyse de signaux. La stratégie d'optimisation proposée par ces auteurs, reposant sur un algorithme alterné de montée de gradient, peut en conséquence être adaptée à notre problème. La suite du présent article est organisée ainsi. La section 2 introduit la famille des distributions temps-fréquence à paramétrisation radialement Gaussienne, et montre comment en améliorer la lisibilité par une atténuation des composantes interférentielles. En section 3, on présente un cadre général pour la mise en œuvre des méthodes à noyau dans le domaine temps-fréquence. La section 4 est consacrée à la présentation des méthodes de sélection de noyau reproduisant, où l'on s'intéresse plus particulièrement au critère d'alignement noyau-cible. Cette section est largement consacrée à l'usage de ce dernier pour l'élaboration de distributions temps-fréquence à paramétrisation radialement Gaussienne dans un contexte décisionnel. L'article s'achève par la section 5, qui vient illustrer la pertinence de l'approche proposée au travers d'expérimentations.

2 Distributions temps-fréquence radialement Gaussiennes

On s'intéresse aux distributions C_x de la classe de Cohen, caractérisées par une fonction de paramétrisation Π que l'on a choisie de considérer dans le plan Doppler-retard

$$C_x(t,f) = \iint \Pi(\nu,\tau)\, A_x(\nu,\tau)\, e^{-2j\pi(f\tau+\nu t)}\, d\nu\, d\tau, \quad (1)$$

où A_x désigne la fonction d'ambiguïté à bande étroite du signal x. Bien que la sélection de distributions vérifiant certaines propriétés théoriques puisse être systématisée par le choix d'une fonction de paramétrisation adéquate, rien ne garantit leur pertinence pour le problème posé. Afin de pallier cette difficulté, on peut envisager de recourir à une procédure d'optimisation destinée à en améliorer la lisibilité dans le cadre d'un problème d'analyse, ou le contraste inter-classe dans le cadre d'un problème de classification. Cette dernière question fait l'objet du présent article. La première a été étudiée par Baraniuk et Jones dans [1], où les auteurs recherchent un compromis entre la résolution temps-fréquence d'une part, et la suppression des termes interférentiels d'autre part, par optimisation d'une fonction de paramétrisation à profil radialement Gaussien. Il s'avère plus naturel de définir celle-ci en coordonnées polaires[1]. On considère dans ce but les variables radiale $r^2 = \nu^2 + \tau^2$ et angulaire $\theta = \arctan(\tau/\nu)$. La fonction de paramétrisation retenue s'exprime alors sous la forme

$$\Pi(r,\theta) = e^{-r^2/2\sigma^2(\theta)}.$$

La largeur de bande σ est fonction de l'angle θ, et détermine la forme de la fonction de paramétrisation.

Étant donné un signal x, Baraniuk et Jones proposent dans [1] de déterminer la largeur de bande $\sigma(\theta)$ qui maximise le volume de la fonction de caractérisation[2] sous une contrainte volumique sur Π, pénalisant ainsi les termes d'interférence distants de l'origine par nature. Le problème d'optimisation s'exprime en coordonnées polaires selon

$$\max_{\sigma} \int_0^{2\pi} \int_0^{\infty} r\, |A_x(r,\theta)|^2\ e^{-r^2/\sigma^2(\theta)}\, dr\, d\theta \quad (2)$$

sous la contrainte que $\int_0^{2\pi} \sigma^2(\theta)\, \mathrm{d}\theta$ soit égale à une constante donnée. Afin de traiter ce problème, une discrétisation des coordonnées polaires dans le plan Doppler-retard est nécessaire. Proposée dans [16], celle-ci permet de transformer le problème d'optimisation précédent ainsi :

$$\max_{\sigma} \sum_r \sum_\theta r\, |A_x(r,\theta)|^2\ e^{-(r\,\Delta_r)^2/\sigma^2(\theta)} \quad (3)$$

sous la contrainte

$$\sum_\theta \sigma^2(\theta) = \upsilon. \quad (4)$$

Ici, υ est le paramètre contrôlant le compromis lissage/interférences, et $\Delta_r = 2\sqrt{\pi/\ell}$ où ℓ est la taille du signal échantillonné x. Voir [16] pour plus de détails. Afin de résoudre ce problème, un algorithme de montée de gradient est proposé. Dans un premier temps, la solution $\sigma_{k+1}(\theta)$ à l'itération $k+1$ est déterminée par la mise-à-jour de $\sigma_k(\theta)$ selon

$$\sigma_{k+1}(\theta) = \sigma_k(\theta) + \mu_k\, \frac{\partial g}{\partial \sigma_k(\theta)}, \text{ pour } \theta = 0, 1, \ldots, \ell-1 \quad (5)$$

où μ_k est le pas, et g la fonction objectif à maximiser dans (3). Le gradient en $\sigma_k(\theta)$ est défini à partir de

$$\frac{\partial g}{\partial \sigma_k(\theta)} = \frac{2\Delta_r^2}{\sigma_k^3(\theta)} \sum_r |A_x(r,\theta)|^2\ r^3\ e^{-(r\ \Delta_r)^2/\sigma_k^2(\theta)}. \quad (6)$$

Dans un second temps, la contrainte (4) est prise en compte en normalisant $\sigma_{k+1}(\theta)$ par $||\sigma_{k+1}(\theta)||/\upsilon$ à chaque itération, ce qui correspond à une projection dans l'ensemble des fonctions admissibles.

3 Méthodes à noyau dans le domaine temps-fréquence

Le choix de la distribution temps-fréquence optimale pour une application donnée a suscité d'amples études [1, 4]. Par le lien que l'on établit ici entre distribution temps-fréquence et noyau reproduisant, on traduit ce problème par la recherche d'un noyau reproduisant optimal au sens d'un critère à définir. On rappelle ici le cadre général de la mise en œuvre des méthodes à noyau dans le domaine temps-fréquence, initialement présenté dans [13] et étudié plus récemment dans [14]. On s'intéresse ensuite au problème de sélection de noyau reproduisant en vue de le transposer à celui des représentations temps-fréquence.

3.1 Rappels du cadre théorique

Les méthodes à noyau sont pour la plupart issues d'algorithmes linéaires auxquels on a pu appliquer les deux résultats clés que sont le coup du noyau et le théorème de représentation [17]. Leur présentation fait ici suite à celle de la notion de noyau.

[1] Afin de simplifier la présentation, la même notation sera utilisée pour désigner la fonction en coordonnées rectangulaires et polaires, c'est-à-dire $\Pi(\nu,\tau)$ et $\Pi(r,\theta)$ respectivement.

[2] La fonction de caractérisation est la fonction d'ambiguïté pondérée par la fonction de paramétrisation, soit $A_x(r,\theta)\Pi(r,\theta)$.

3.1.1 Noyau défini positif et noyau reproduisant

On considère un sous-espace $\mathcal{X}$ de $\mathcal{L}_2(\mathbb{C})$, l'espace des signaux d'énergie finie à valeurs dans $\mathbb{C}$, que l'on munit du produit scalaire canonique $\langle x_i, x_j \rangle = \int_t x_i(t)\,\overline{x}_j(t)\,dt$ et de la norme associée. Un noyau désigne une fonction $\kappa(x_i, x_j)$ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$, à symétrie hermitienne. On rappelle les deux définitions fondamentales suivantes [18].

Définition 1. *Un noyau κ est dit défini positif sur $\mathcal{X}$ s'il vérifie*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i\,\overline{a}_j\,\kappa(x_i, x_j) \geq 0 \tag{7}$$

pour tout $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ et $a_1, \ldots, a_n \in \mathbb{C}$.

Définition 2. *Soit $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ un espace de Hilbert constitué de fonctions de $\mathcal{X}$ dans $\mathbb{C}$. La fonction $\kappa(x_i, x_j)$ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$ est le noyau reproduisant de $\mathcal{H}$, sous réserve qu'il en admette un, si et seulement si*

- *la fonction $\kappa_{x_i} : x_j \mapsto \kappa_{x_i}(x_j) = \kappa(x_i, x_j)$ appartient à $\mathcal{H}$, quel que soit $x_i \in \mathcal{X}$ fixé ;*
- *on a $\psi(x_i) = \langle \psi, \kappa_{x_i} \rangle_{\mathcal{H}}$ pour tout $x_i \in \mathcal{X}$ et $\psi \in \mathcal{H}$.*

On démontre que tout noyau défini positif κ est le noyau reproduisant d'un espace de Hilbert de fonctions de $\mathcal{X}$ dans $\mathbb{C}$ en exhibant directement celui-ci. On considère pour cela l'espace vectoriel $\mathcal{H}_0$ engendré par l'ensemble des fonctions $\{\kappa_x\}_{x \in \mathcal{X}}$, auquel on associe le produit scalaire

$$\langle \psi, \phi \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i\,\overline{b}_j\,\kappa(x_i, x_j), \tag{8}$$

avec $\psi = \sum_{i=1}^{n} a_i\,\kappa_{x_i}$ et $\phi = \sum_{j=1}^{m} b_j\,\kappa_{x_j}$ appartenant à l'espace $\mathcal{H}_0$. Il s'agit donc là d'un espace pré-hilbertien, que l'on complète conformément à [18] de sorte que toute suite de Cauchy y converge. On aboutit ainsi à l'espace de Hilbert $\mathcal{H}$ à noyau reproduisant κ recherché. Réciproquement, on peut démontrer que tout noyau reproduisant est défini positif [18]. A titre d'exemple, parmi les noyaux reproduisants les plus classiques, citons en particulier le noyau gaussien $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma_0^2)$, où σ_0^2 est un paramètre à fixer. D'autres noyaux, ainsi que des règles pour les combiner, sont présentés dans [8, 19].

3.1.2 Coup du noyau

Du second point de la définition 2 résulte une propriété fondamentale des espaces de Hilbert à noyau reproduisant. En remplaçant ψ par κ_{x_j}, on aboutit en effet à

$$\kappa(x_i, x_j) = \langle \kappa_{x_j}, \kappa_{x_i} \rangle_{\mathcal{H}} \tag{9}$$

quels que soient x_i et $x_j \in \mathcal{X}$. Ainsi le noyau κ fournit-il le produit scalaire des images dans $\mathcal{H}$ de toute paire d'éléments de $\mathcal{X}$, sans qu'il soit nécessaire d'expliciter ces images. Ce principe est appelé *coup du noyau*. Il permet d'élaborer des méthodes non-linéaires de traitement de données à partir d'approches linéaires, sous réserve que celles-ci puissent s'exprimer uniquement en fonction de produits scalaires des observations. Il suffit en effet de remplacer alors chacun de ces produits scalaires par un noyau non-linéaire. Ainsi la structure des algorithmes demeure-t-elle inchangée et le surcoût calculatoire dû à l'évaluation des noyaux négligeable.

Pour qu'il soit opérationnel, le coup du noyau nécessite souvent d'être associé au théorème de représentation [17]. Celui-ci établit que toute fonction ψ^* d'un espace de Hilbert à noyau reproduisant $\mathcal{H}$ qui minimise un coût

$$J((x_1, y_1, \psi(x_1)), \ldots, (x_n, y_n, \psi(x_n))) + h(\|\psi\|^2_{\mathcal{H}}), \tag{10}$$

impliquant n sorties $\psi(x_i)$ obtenues pour des entrées x_i et éventuellement n sorties désirées y_i, avec h une fonction monotone croissante sur $\mathbb{R}_+$, peut s'écrire sous la forme

$$\psi^* = \sum_{i=1}^{n} a_i^*\,\kappa_{x_i}. \tag{11}$$

On démontre ceci en notant que toute fonction ψ se décompose selon $\psi = \sum_{i=1}^{n} a_i\,\kappa_{x_i} + \psi^\perp$, où $\langle \psi^\perp, \kappa_{x_i} \rangle_{\mathcal{H}} = 0$ pour tout $i = 1, \ldots, n$. Puisque $\psi(x_j) = \langle \psi, \kappa_{x_j} \rangle_{\mathcal{H}}$, la valeur de $\psi(x_j)$ n'est donc pas affectée par $\psi^\perp$.

3.2 Mises en œuvre temps-fréquence

L'objectif de cette section est de montrer que les méthodes à noyau peuvent être mises en œuvre dans le domaine temps-fréquence grâce à un choix approprié de noyau reproduisant. On s'intéresse d'abord à la distribution de Wigner, par soucis de clarté, avant d'étendre l'étude aux autres distributions de la classe de Cohen. On suppose disposer dans la suite d'un ensemble d'apprentissage $\mathcal{A}_n$ comprenant n signaux $x_i \in \mathcal{X}$, éventuellement accompagnés d'une étiquette ou autre sortie désirée $y_i \in \mathcal{Y}$.

3.2.1 Distribution de Wigner

On désigne par W_x la distribution de Wigner de x, c'est-à-dire

$$W_x(t, f) = \int x(t + \tau/2)\,\overline{x}(t - \tau/2)\,e^{-2j\pi f \tau}\,d\tau. \tag{12}$$

Les classiques méthodes paramétriques de reconnaissance des formes, appliquées ici dans le plan temps-fréquence, reposent sur l'estimation de Ψ^* de sorte que la statistique

$$\psi^*(x) = \langle \Psi^*, W_x \rangle = \iint \Psi^*(t, f)\,W_x(t, f)\,dt\,df \tag{13}$$

optimise un critère J donné, par exemple celui de Fisher pour une analyse factorielle discriminante. En pratique,[3] la résolution d'un tel problème est rendue difficile par la taille des représentations W_{x_i} manipulées. Le coup du noyau et le théorème de représentation peuvent ici jouer un rôle déterminant. On considère pour cela le noyau suivant

$$\kappa_W(x_i, x_j) = \langle W_{x_i}, W_{x_j} \rangle. \tag{14}$$

Notons immédiatement qu'il ne nécessite aucun calcul de distributions de Wigner puisque, par la relation de Moyal, on a

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

[3] Les méthodes de discrétisation de la distribution de Wigner étant multiples, on a pris le parti de raisonner sur la définition (12) dans cet article. En pratique, on rappelle toutefois que les représentations temps-fréquence discrètes de signaux composés de ℓ échantillons sont des matrices de taille $(\ell \times \ell)$.

On vérifie qu'il s'agit bien d'un noyau défini positif. La condition (7), que l'on peut réécrire $\| \sum_i a_i W_{x_i} \|^2 \geq 0$, est en effet satisfaite. Il lui correspond donc un espace de Hilbert à noyau reproduisant $\mathcal{H}_W$ unique. On l'obtient en complétant l'espace fonctionnel $\mathcal{H}_0$ défini ci-dessous de sorte que toute suite de Cauchy y converge, ce que l'on note $\mathcal{H}_W = \overline{\mathcal{H}}_0$.

$$\mathcal{H}_0 = \{\psi : \mathcal{X} \to \mathbb{R} \mid \psi = \sum_i a_i \, |\langle \cdot \, , x_i \rangle|^2, a_i \in \mathbb{R}, x_i \in \mathcal{X}\}$$

3.2.2 Classe de Cohen

On se concentre à présent sur les distributions C_x de la classe de Cohen. Le noyau reproduisant associé, dans le plan Doppler-retard, s'exprime ainsi

$$\kappa(x_i, x_j) = \iint |\Pi(\nu, \tau)|^2 \; A_{x_i}(\nu, \tau) \; \overline{A}_{x_j}(\nu, \tau) \; d\nu \; d\tau, \quad (16)$$

où les fonctions d'ambiguïté sont données en coordonnées cartésiennes. En coordonnées polaires, le noyau reproduisant s'exprime suivant

$$\kappa(x_i, x_j) = \iint r \; |\Pi(r, \theta)|^2 \; A_{x_i}(r, \theta) \; \overline{A}_{x_j}(r, \theta) \; dr \; d\theta. \quad (17)$$

Dans le cas des distributions temps-fréquence à fonction de paramétrisation radialement Gaussienne, le noyau reproduisant associé est obtenu pour $\Pi_\sigma(r, \theta) = e^{-r^2/2\sigma^2(\theta)}$, ce qui permet d'écrire

$$\kappa_\sigma(x_i, x_j) = \iint r A_{x_i}(r, \theta) \overline{A}_{x_j}(r, \theta) e^{-r^2/\sigma^2(\theta)} \; dr \; d\theta. \quad (18)$$

La mise en œuvre des méthodes à noyau dans le domaine temps-fréquence nécessite une formulation discrète du noyau reproduisant associé. Dans le cas de distributions radialement Gaussiennes, on opte pour une discrétisation en coordonnées polaires comme proposé dans [1, 16]. Le noyau reproduisant correspondant est alors donné par

$$\kappa_\sigma(x_i, x_j) = \sum_{r,\theta} r \, A_{x_i}(r, \theta) \, \overline{A}_{x_j}(r, \theta) \; e^{-(r\Delta_r)^2/\sigma^2(\theta)}, \quad (19)$$

où $\Delta_r = 2\sqrt{\pi/\ell}$, ℓ étant la taille des signaux échantillonnés. Rien ne s'oppose désormais à l'association du noyau reproduisant κ_σ défini dans (19) à tout algorithme à noyau décrit dans la littérature, comme étudié dans [13, 14] pour d'autres distributions temps-fréquence. Dans ce qui suit, on rappelle succinctement les différentes techniques de sélection de noyau proposées dans le cadre des méthodes à noyau, avant d'étudier plus particulièrement le critère d'alignement noyau-cible.

4 Élaboration de la distribution radialement Gaussienne

Plusieurs résultats de la théorie de l'apprentissage montrent l'importance des informations *a priori*. Pour les méthodes à noyau, le théorème *no free kernel* [20] stipule qu'aucun noyau n'est optimal pour toutes les applications, et qu'une connaissance *a priori* pouvant mener à un choix approprié de celui-ci est nécessaire. Pour cela, on peut recourir à des algorithmes mettant en évidence une relation entre les données d'entrée et leurs cibles, appelées aussi étiquettes en classification. On s'intéresse dans la suite au critère d'alignement noyau-cible, après une brève présentation de différentes approches existantes pour la sélection de noyau.

4.1 Méthodes de sélection de noyau reproduisant : un aperçu

Outre les techniques de transformation conforme [21], on compte les approches de validation croisée et autres *leave-one-out*. Pour chaque noyau étudié, l'erreur de généralisation est estimée à partir de plusieurs classifieurs élaborés pour différents sous-ensembles de la base d'apprentissage et validés sur les données restantes [22]. Le noyau optimal, celui qui correspond à la plus faible erreur, est souvent obtenu par une recherche sur une grille de valeurs. Cette approche s'avère très coûteuse puisqu'on a recours à plusieurs phases d'apprentissage et de validation. Afin d'accélérer cette recherche, des méthodes reposant sur une borne de l'erreur de généralisation ont été récemment proposées, nécessitant un seul apprentissage de classifieur par noyau étudié. Parmi celles-ci, on a la borne fournie par la VC-dimension [23] et celle de rayon-marge [24]. D'autres possibilités sont également citées dans [25, 26]. Des méthodes permettant de combiner optimalement plusieurs noyaux sortent du cadre de ce travail [27, 28, 29].

Il est important de noter que les différentes approches présentées ci-dessus se transposent aisément au domaine temps-fréquence. Il en est de même pour leurs propriétés statistiques et théoriques. On se contente dans la suite de considérer un critère à complexité calculatoire réduite, le critère d'alignement noyau-cible, et de présenter sa mise en œuvre pour l'élaboration d'une distribution temps-fréquence radialement Gaussienne.

4.2 Alignement noyau-cible

Les performances des méthodes de reconnaissance des formes à noyau sont principalement déterminées par le noyau reproduisant considéré. Pour un algorithme donné, par exemple les SVM, deux noyaux produiront des résultats d'autant plus proches qu'ils sont similaires. L'alignement, introduit par Cristianini *et coll.* dans [30] est une mesure de similarité entre deux noyaux, ou entre un noyau et une fonction cible. Étant donné un ensemble d'apprentissage $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ de données étiquetées, l'alignement (empirique) des deux noyaux κ_1 et κ_2 est défini par

$$\mathcal{A}(K_1, K_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}}, \quad (20)$$

où $\langle \cdot, \cdot \rangle_F$ est le produit scalaire de Frobenius[4], et K_1 et K_2 les matrices de Gram de termes respectifs $\kappa_1(x_i, x_j)$ et $\kappa_2(x_i, x_j)$. Ainsi cette mesure correspond-t-elle plus concrètement au coefficient de corrélation entre les deux

[4] Le produit scalaire de Frobenius entre les matrices K_1 et K_2 est défini par $\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^n \kappa_1(x_i, x_j) \, \kappa_2(x_i, x_j)$.

matrices K_1 et K_2. On a $-1 \leq \mathcal{A}(K_1, K_2) \leq 1$ en général et $\mathcal{A}(K_1, K_2) \geq 0$ pour les matrices définies positives [31].

Le concept d'alignement peut aussi quantifier la similarité entre un noyau reproduisant et un noyau cible. Dans le cadre d'un problème de classification, rappelons que l'on cherche à identifier une règle décisionnelle ψ^* qui respecte la relation $\psi^*(x_i) = y_i$, où y_i est l'étiquette de l'observation x_i. Pour un problème à deux classes en particulier, on souhaite aboutir à $\psi^*(x_i) = \pm 1$ selon la classe d'appartenance de x_i. Dans ces conditions, le noyau cible correspondant est $\kappa^*(x_i, x_j) = y_i\, y_j$. La matrice de Gram idéale associée est alors $K^* = \boldsymbol{y}\boldsymbol{y}^\top$, où $\boldsymbol{y}$ est un vecteur colonne cible dont le $j^{\text{ème}}$ élément est y_j. Notons que

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

Dans ce qui suit, on se limite à l'étude d'un problème de classification bi-classes avec la matrice K^* définie ci-dessus pour cible. Toutefois, il est à noter qu'une généralisation au cas multi-classes est tout à fait envisageable.

Dans [20, 30], Cristianini *et coll.* suggèrent d'utiliser cette matrice cible $K^* = \boldsymbol{y}\boldsymbol{y}^\top$ et le critère d'alignement afin de rechercher le noyau reproduisant le mieux adapté. L'expression de l'alignement noyau-cible, entre un noyau reproduisant κ donné et le noyau optimal κ^*, s'exprime alors selon

$$\mathcal{A}(K, K^*) = \frac{\langle K, \boldsymbol{y}\boldsymbol{y}^\top \rangle_F}{\sqrt{\langle K, K \rangle_F \langle \boldsymbol{y}\boldsymbol{y}^\top, \boldsymbol{y}\boldsymbol{y}^\top \rangle_F}} = \frac{\boldsymbol{y}^\top K \boldsymbol{y}}{n \|K\|_F},$$

où K est la matrice de Gram de terme $\kappa(x_i, x_j)$. La pertinence du critère d'alignement est assurée par un lien avec l'erreur de généralisation, reposant sur le fait qu'un estimateur de Parzen de la forme $h(\cdot) = \frac{1}{n} \sum_{i=1}^n y_i \kappa(x_i, \cdot)$ admet des performances en généralisation d'autant meilleures que l'alignement est élevé [30]. Depuis les premiers travaux de Cristianini *et coll.*, le principe de maximisation de l'alignement noyau-cible a été étendu à d'autres problèmes, par exemple celui de la régression [32]. Il a également suscité plusieurs travaux sur l'apprentissage de métriques [33, 34], la décomposition en valeurs propres de la matrice de Gram [20, 32], ou encore la combinaison de noyaux en vue d'une amélioration des performances de l'ensemble [27, 28, 29].

4.3 Optimisation de l'alignement

La stratégie la plus simple pour ajuster les paramètres d'un noyau consiste à évaluer l'alignement sur une grille de valeurs possibles. Bien que cette approche puisse être envisagée pour un paramètre unique, elle s'avère coûteuse au-delà. Pour remédier à cela, on préconise une approche de montée de gradient. Cette dernière ne peut évidemment être envisagée que sous réserve que l'alignement soit differentiable par rapport aux paramètres considérés. Il en est ainsi pour les noyaux Gaussien et de Laplace par rapport à leur largeur de bande. Ceci n'est en revanche pas vrai pour le noyau polynomial $\kappa_q(x_i, x_j) = (p + \langle x_i, x_j \rangle)^q$ par rapport à son exposant q, bien que cela soit le cas par rapport à p.

Soit κ_σ un noyau reproduisant de paramètre σ. On considère le problème d'optimisation de l'alignement noyau-cible défini par

$$\sigma^* = \arg\max_\sigma \mathcal{A}(K_\sigma, K^*) = \arg\max_\sigma \frac{\langle K_\sigma, K^* \rangle_F}{n \|K_\sigma\|_F}, \quad (21)$$

où K_σ désigne la matrice de Gram du noyau κ_σ et K^* la matrice cible. Bien que le problème ne soit pas convexe à ce stade, on peut indiquer une méthode de gradient pour le résoudre. L'expression du gradient du numérateur dans (21) par rapport à σ s'écrit

$$\nabla_\sigma \langle K_\sigma, K^* \rangle_F = \sum_{i,j=1}^n y_i y_j \nabla_\sigma \kappa_\sigma(x_i, x_j).$$

Le gradient de $\|K_\sigma\|_F$ par rapport à σ s'exprime selon

$$\begin{aligned} \nabla_\sigma \|K_\sigma\|_F &= \nabla_\sigma \Big(\sum_{i,j=1}^n \kappa_\sigma^2(x_i, x_j) \Big)^{-\frac{1}{2}} \\ &= \Big(\sum_{i,j=1}^n \kappa_\sigma^2(x_i, x_j) \Big)^{-\frac{1}{2}} \sum_{i,j=1}^n \kappa_\sigma(x_i, x_j)\, \nabla_\sigma \kappa_\sigma(x_i, x_j) \\ &= \frac{1}{\|K_\sigma\|_F} \sum_{i,j=1}^n \kappa_\sigma^2(x_i, x_j)\; \nabla_\sigma \kappa_\sigma(x_i, x_j). \end{aligned}$$

En combinant les deux expressions, on peut alors écrire le gradient de l'alignement $A(K_\sigma, K^*)$ par rapport à σ ainsi

$$\begin{aligned} \nabla_\sigma \mathcal{A}(K_\sigma, K^*) \;&=\; \frac{1}{n\|K_\sigma\|_F} \sum_{i,j=1}^n y_i y_j \nabla_\sigma \kappa_\sigma(x_i, x_j) \\ &-\; \frac{\langle K_\sigma, K^* \rangle_F}{n \|K_\sigma\|_F^3} \sum_{i,j=1}^n \kappa_\sigma(x_i, x_j)\; \nabla_\sigma \kappa_\sigma(x_i, x_j). \end{aligned}$$

La méthode de montée de gradient pour la maximisation de l'alignement est principalement constituée de l'étape de mise-à-jour

$$\sigma_{k+1} = \sigma_k + \mu_k \nabla_\sigma \mathcal{A}(K_\sigma, K^*).$$

Des contraintes peuvent être prises en compte à chaque itération, dans un processus d'optimisation alternée. Une telle démarche est utilisée dans la suite pour l'optimisation de la distribution temps-fréquence à paramétrisation radialement Gaussienne.

4.4 Application aux distributions temps-fréquence

L'usage du critère d'alignement noyau-cible présente l'intérêt de ne nécessiter aucun apprentissage de la statistique de décision, la sélection du noyau étant pratiquée *a priori*. Dans [15], nous avons présenté plusieurs applications potentielles dans le cadre des distributions temps-fréquence. Parmi celles-ci, nous avons notamment étudié la sélection d'une distribution optimale parmi un ensemble de distributions candidates. On s'intéresse dans la suite à l'ajustement de la fonction de paramétrisation lorsqu'elle est de type radialement Gaussien, de sorte à maximiser le critère d'alignement noyau-cible.

En considérant les distributions temps-fréquence à fonction de paramétrisation radialement Gaussienne, on

s'intéresse à la fonction largeur de bande optimale $\sigma^*(\theta)$ maximisant l'alignement noyau-cible selon l'expression

$$\sigma^* = \arg\max_{\sigma} \frac{\langle K_\sigma, K^* \rangle_F}{n \|K_\sigma\|_F}. \quad (22)$$

Ce problème d'optimisation peut être traité comme préconisé dans la section précédente. Toutefois, afin d'établir le lien avec [1], nous allons considérer ci-dessous la maximisation du numérateur de l'alignement sous contrainte que son dénominateur soit constant :

$$\max_{\sigma} \sum_{i,j=1}^{n} y_i \, y_j \, \kappa_\sigma(x_i, x_j), \quad (23)$$

sous la contrainte

$$\sum_{i,j=1}^{n} \kappa_\sigma(x_i, x_j)^2 = \upsilon_0, \quad (24)$$

où υ_0 est un paramètre de normalisation. En développant la fonction objectif (3) à maximiser, on aboutit à

$$\sum_{i,j=1}^{n} y_i y_j \kappa_\sigma(x_i, x_j) = \iint r |A_{\rm eq}(r,\theta)|^2 \, e^{-r^2/\sigma^2(\theta)} \, dr \, d\theta \quad (25)$$

avec $|A_{\rm eq}(r,\theta)|^2 = \sum_{i,j} y_i y_j \, A_{x_i}(r,\theta) \, \overline{A}_{x_j}(r,\theta)$. On retrouve ainsi un problème équivalent à (2) dans lequel la partie dépendante du signal, soit $|A_x(r,\theta)|^2$, est remplacée par la représentation équivalente $|A_{\rm eq}(r,\theta)|^2$. Il est à noter que celle-ci ne dépend que des données de l'ensemble d'apprentissage, et qu'elle peut être évaluée préalablement à toute optimisation. On peut alors recourir à l'algorithme d'optimisation alternée proposé dans [1], avec la même complexité. Pour cela, on relâche la contrainte (24), coûteuse en temps de calcul, en la remplaçant par la contrainte volumique $\int \sigma^2(\theta) \, {\rm d}\theta = \upsilon'_0$ comme préconisé dans [1]. En reprenant la fonction objectif (3) avec la formulation discrète du noyau reproduisant, le problème d'optimisation s'écrit donc ainsi en coordonnées polaires

$$\max_{\sigma} \sum_{r,\theta} r \; |A_{\rm eq}(r,\theta)|^2 \; e^{-(r\Delta_r)^2 \sigma^2(\theta)}, \quad (26)$$

sous la contrainte

$$\sum_{\theta} \sigma^2(\theta) = \upsilon'_0. \quad (27)$$

Pour résoudre ce problème d'optimisation avec contrainte, on considère un algorithme de montée de gradient alternée similaire à celui présenté à la section 2. A l'itération $k+1$, on opère dans un premier temps une mise-à-jour de la solution selon

$$\sigma_{k+1}(\theta) = \sigma_k(\theta) + \mu_k \, \frac{\partial f}{\partial \sigma_k(\theta)}, \quad (28)$$

où μ_k est un paramètre contrôlant la vitesse de convergence et f la fonction objectif à maximiser dans (26). Le gradient de f en $\sigma_k(\theta)$ est défini à partir de

$$\frac{\partial f}{\partial \sigma_k(\theta)} = \frac{2\Delta_r^2}{\sigma_k^3(\theta)} \sum_r r^3 \, |A_{\rm eq}(r,\theta)|^2 \; e^{-(r\,\Delta_r)^2/\sigma^2(\theta)}. \quad (29)$$

Dans une seconde étape, on prend en compte la contrainte en projetant la solution sur l'ensemble des fonctions admissibles, ce qui revient à normaliser $\sigma_{k+1}(\theta)$ à chaque itération selon $\|\sigma_{k+1}(\theta)\|/\upsilon'_0$.

On insiste sur le fait que la représentation $|A_{\rm eq}(r,\theta)|^2$ peut être calculée dans une étape d'initialisation. De plus, celle-ci se prête à un calcul itératif ne nécessitant pas de conserver en mémoire l'ensemble des fonctions d'ambiguïté des signaux de l'ensemble d'apprentissage. Une fois la représentation $|A_{\rm eq}(r,\theta)|^2$ obtenue, la technique d'optimisation devient indépendante de la taille de la base d'apprentissage. Ceci n'est pas le cas d'approches précédemment proposées, en particulier dans [3], qui nécessitent l'évaluation des représentations temps-fréquence de chaque signal de l'ensemble d'apprentissage, et ce à chaque itération de l'algorithme d'optimisation. Pour cette raison certainement, les auteurs de cet article ont été contraints de se restreindre à un ensemble d'apprentissage de 15 signaux pour chaque classe.

5 Expérimentations

On considère successivement deux problèmes de classification de deux familles de 200 signaux de taille 64, à modulation fréquentielle linéaire, noyés dans un bruit blanc Gaussien de variance 4. Ceci correspond à un rapport signal-sur-bruit de l'ordre de -8 dB.

La première application concerne des signaux à modulation fréquentielle croissante, de 0.1 à 0.25 pour la première classe, et de 0.25 à 0.4 pour la seconde, en échelle fréquentielle normalisée. La Figure 1 (a) en haut présente la fonction de paramétrisation à profil radialement Gaussien ainsi obtenue, dont le profil dans le plan Doppler-retard s'avère parfaitement pertinent pour le problème de classification. La Figure 1 (a) en bas illustre son contour en rouge $\sigma(\theta)$, ainsi que le contour initial $\sigma_0(\theta)$ avant optimisation en bleu. Ce dernier est déterminé par la contrainte de volume, que l'on a fixé à $\upsilon'_0 = 2$. Dans une seconde application, on propose d'étudier le cas où les régions des signaux des deux classes sont distinctes dans le plan Doppler-retard. On considère pour cela des signaux comportant une modulation fréquentielle linéairement croissante de 0.1 à 0.4 pour la première classe, et décroissante de 0.4 à 0.1 pour la seconde classe. La Figure 1 (b) représente la fonction de paramétrisation obtenue à l'aide de notre algorithme. Elle correspond à un filtrage de l'information pertinente des deux régions d'intérêt pour le problème de classification traité. On représente sur la Figure 2 l'évolution moyenne de l'alignement au cours des itérations, sur 20 réalisations et pour chacun des problèmes.

Afin d'illustrer la pertinence de cette stratégie, on propose d'estimer l'erreur de classification obtenue à partir d'un classifieur de type SVM pour les distributions de Wigner et à fonction de paramétrisation radialement Gaussienne. Le Tableau 1 présente, en les moyennant sur 20 réalisations, le taux d'erreur obtenu sur un ensemble de test de 2000 signaux, ainsi que le nombre de vecteurs support correspondant. Non seulement la distribution optimale conduit à une erreur de classification plus faible, mais elle aboutit aussi à une division par deux environ du nombre de vecteurs support. Ceci est principalement dû, d'une part au caractère optimal de la distribution ainsi

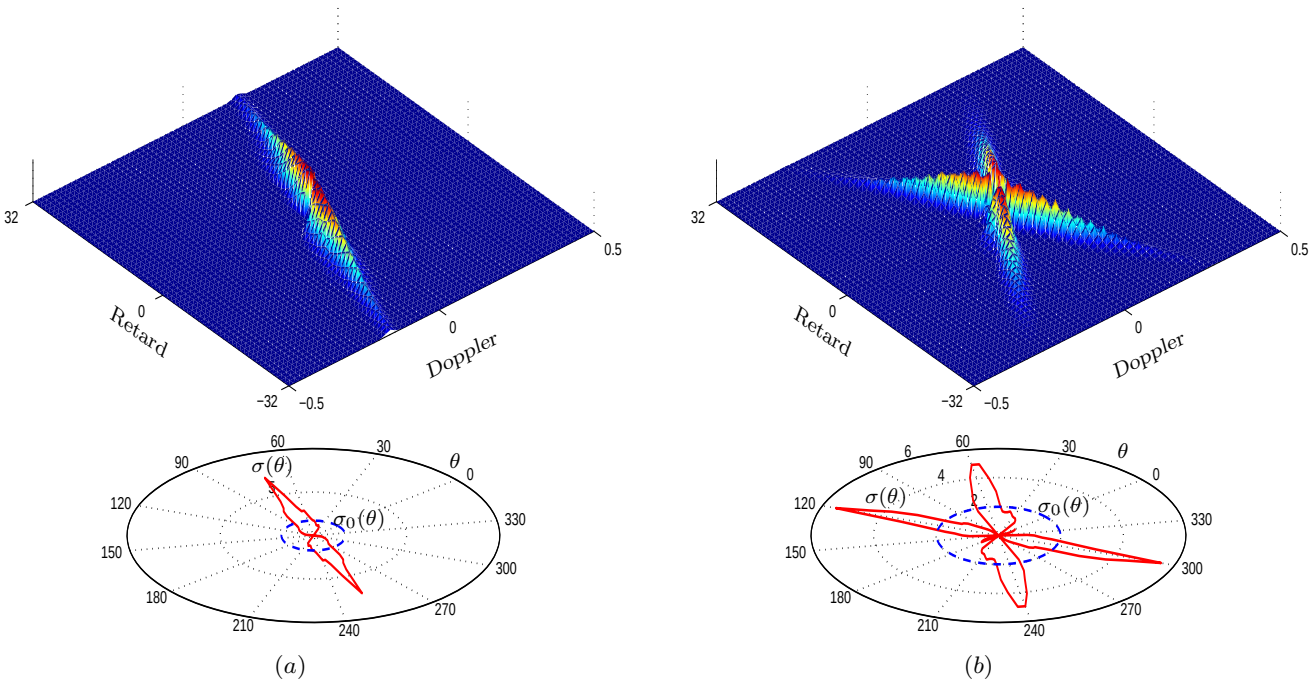

Figure 1: Résultats obtenus pour la 1ère application (a) et la 2e application (b). En haut: fonction de paramétrisation optimale résultante. En bas: son contour en rouge, et le contour initial en bleu, en coordonnées polaires.

	1ère application		2e application	
	Pourcentage d'erreur	Nombre de SV	Pourcentage d'erreur	Nombre de SV
Distribution de Wigner	19.10 ± 1.23	161.9 ± 4.97	19.41 ± 1.34	164.3 ± 5.62
Distribution optimale	16.89 ± 2.01	65.2 ± 4.46	17.81 ± 1.72	83.85 ± 5.65

Table 1: Comparaison du pourcentage d'erreur et du nombre de vecteurs support (SV) pour un classifieur SVM associé, d'une part à la distribution de Wigner, et d'autre part à la distribution optimale, pour chacune des deux applications.

obtenue, et d'autre part au caractère régularisant de ce traitement.

6 Conclusion

Emprunté aux méthodes à noyau pour la sélection de noyau reproduisant optimal, le critère d'alignement noyau-cible s'avère très pertinent pour élaborer des distributions temps-fréquence pour la classification de signaux non-stationnaires. Dans le cas particulier d'une distribution radialement Gaussienne, nous avons montré que la maximisation de ce critère se réduit à un problème d'optimisation pour lequel il existe une technique de résolution ayant fait ses preuves dans la communauté temps-fréquence. La pertinence de notre approche est soutenue par des expérimentations, qui montrent une amélioration sensible des performances en classification des SVM associés, ainsi qu'une diminution du nombre de vecteurs support.

References

[1] R. Baraniuk and D. Jones, "Signal-dependent time-frequency analysis using a radially gaussian kernel," *Signal Processing*, vol. 32, no. 3, pp. 263–284, 1993.

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

[3] M. Davy and C. Doncarli, "Optimal kernels of time-frequency representations for signal classification," in *Proc. of the IEEE International Symposium on Time-Frequency and Time-Scale analysis*, (Pittsburgh, USA), pp. 581–584, Oct. 1998.

[4] C. Doncarli and N. Martin, *Décision dans le plan temps-fréquence*. Paris: Hermès Sciences, Traité IC2, 2004.

[5] 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.

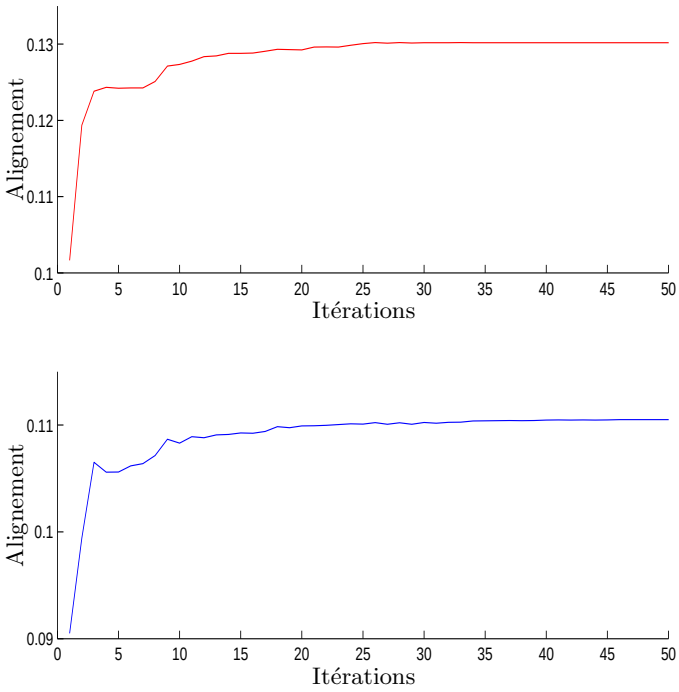

Figure 2: Evolution de l'alignement moyenné sur 20 réalisations pour le premier (en haut) et le second (en bas) problème.

[6] 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.

[7] B. Boser, I. Guyon, and V. Vapnik, "An training algorithm for optimal margin classifiers," in *Proc. 5th Annual Workshop on Computational Learning Theory*, pp. 144–152, 1992.

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

[9] 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.

[10] 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, USA), pp. 41–48, Morgan Kaufmann, 1999.

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

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

[13] P. Honeine, C. Richard, and P. Flandrin, "Reconnaissance des formes par méthodes à noyau dans le domaine temps-fréquence," in *Actes du XX^{ème} Colloque GRETSI sur le Traitement du Signal et des Images*, (Louvain-la-Neuve, Belgium), 2005.

[14] P. Honeine, C. Richard, and P. Flandrin, "Time-frequency learning machines," *IEEE Trans. Signal Processing*, vol. 55, pp. 3930–3936, July 2007.

[15] P. Honeine, 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 (ICASSP)*, (Toulouse, France), May 2006.

[16] R. Baraniuk, "Shear madness: Signal-dependent and metaleptic time-frequency representation," Tech. Rep. UILU-ENG-92-2226, Coordinated Science Laboratory, University of Illinois, Urbana, 1992.

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

[18] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[19] R. Herbrich, *Learning kernel classifiers. Theory and algorithms*. Cambridge, MA, USA: The MIT Press, 2002.

[20] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel target alignment," in *Innovations in Machine Learning: Theory and Application* (D. Holmes and L. Jain, eds.), pp. 205–255, Springer Verlag, 2006.

[21] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, pp. 783–789, July 1999.

[22] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, pp. 169–186, September 2003.

[23] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[24] K.-M. Chung, W.-C. Kao, C.-L. Sun, L.-L. Wang, and C.-J. Lin, "Radius margin bounds for support vector machines with the rbf kernel," *Neural Comput.*, vol. 15, no. 11, pp. 2643–2681, 2003.

[25] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, 2002.

[26] K. Duan, S. Keerthi, and A. Poo, "Evaluation of simple performance measures for tuning svm hyperparameters," *Neurocomputing*, vol. 51, pp. 41–59, 2003.

[27] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "Optimizing kernel alignment over combinations of kernels," Tech. Rep. 121, Department of Computer Science, University of London, 2002.

[28] J.-B. Pothin and C. Richard, "Kernel machines : une nouvelle méthode pour l'optimisation de l'alignement des noyaux et l'amélioration des performances," in *Actes du XX^{ème} Colloque GRETSI sur le Traitement du Signal et des Images*, (Louvain-la-Neuve, Belgium), 2005.

[29] J.-B. Pothin and C. Richard, "A greedy algorithm for optimizing the kernel alignment and the performance of kernel machines," in *Proc. EUSIPCO*, (Florence, Italy), 2006.

[30] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Proc. Neural Information Processing Systems (NIPS) 14* (T. G. Dietterich, S. Becker, and Z. Ghahramani, eds.), pp. 367–373, MIT Press, December 2001.

[31] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[32] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "On the extensions of kernel alignment," Tech. Rep. 120, Department of Computer Science, University of London, 2002.

[33] G. Wu, E. Y. Chang, and N. Panda, "Formulating distance functions via the kernel trick," in *Proc. 11th ACM International conference on knowledge discovery in Data mining*, pp. 703–709, 2005.

[34] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. Jordan, "Learning the kernel matrix with semi-definite programming," in *Proc. 19th International Conference on Machine Learning*, pp. 323–330, 2002.

EXTENDED ABSTRACT

Time-frequency and time-scale distributions offer a broad class of tools for nonstationary signal analysis. The price to pay is however the choice of a well-adapted, possibly signal-dependent, representation. Time-frequency distributions with radially Gaussian kernel (RGK) provide tunable representations for a large class of applications. This parameterization was initially introduced by Baraniuk and Jones in [1] for reducing interference terms in time-frequency distributions. The RGK has attracted the attention of many researchers since this pioneering work, for instance [3] where it is optimized for a given classification problem. However, all these works do not necessarily take advantage of the recent developments in pattern recognition with kernel machines. This paper provides a connection between the optimization of RGK for classifying non-stationary signals and some criteria coming from the machine learning literature. We consider more specifically the kernel-target alignment, a criterion that allows to select optimal reproducing kernels. We show that this criterion leads to an optimization problem similar to that initially proposed by Baraniuk and Jones for signal analysis. Experimental results demonstrate the relevance of our approach.

Over the last decade, a wide class of pattern recognition algorithms based on the theory of reproducing kernel Hilbert spaces have been proposed, with improved performance and low computational complexity. The latter is mainly due to the kernel trick, a key idea that allows to transform any conventional linear technique into a nonlinear one if it can be expressed only in terms of inner products. It suffices to replace them by a reproducing kernel, since it corresponds to an inner product between data mapped (implicitly) into a nonlinearly transformed space. Most popular kernel machines include the Support Vector Machines (SVM) for classification and regression [8], kernel principal component analysis [9] and kernel Fisher discriminant analysis [10].

Recently, we have presented in [13, 14] a new framework for nonstationary signal classification and analysis based on machine learning considerations. For this purpose, we considered the Cohen's class of time-frequency distributions as a natural choice for nonlinear transformations to be applied to nonstationary signals. It can be shown that the reproducing kernel associated to these spaces of representations can be written as (17). Here Π is the parameterizing function of the time-frequency distribution, expressed in polar coordinates, and A_x the ambiguity function. The cumbersome optimization of the two-dimensional function Π can be overcome by using the RGK, which is of the form $\Pi_\sigma(r,\theta) = e^{-r^2/2\sigma^2(\theta)}$. By injecting this expression into the general form of the reproducing kernel defined in (17), we get (18). Equivalently, by considering the discretization recommended in [16] that involves polar coordinates, we get (19) where $\Delta_r = 2\sqrt{\pi/\ell}$ with ℓ the signal length. By considering such a reproducing kernel, we can therefore use the most effective and innovative kernel machines to process time-frequency representations. We can also take advantage of the large spectrum of kernel selection criteria in the machine learning literature to optimize time-frequency representations. In what follows, we study one particular criterion, the kernel-target alignment.

Introduced by Cristianini *et al.* in [30], the alignment is a measure of similarity between two reproducing kernels, or between a reproducing kernel and a target function. Given a learning set, the alignment between two kernels κ_1 and κ_2 is defined by (20) where $\langle\cdot\,,\cdot\rangle_F$ is the Frobenius inner product, and K_1 and K_2 the Gram matrices with (i,j)-th entries $\kappa_1(x_i,x_j)$ and $\kappa_2(x_i,x_j)$. In [20, 30], Cristianini *et al.* suggest to determine the optimal reproducing kernel for a given classification problem as the one that maximizes the alignment with an ideal kernel. The ideal kernel for a bi-class problem is the one that transforms each x_i into its label y_i. The (i,j)-th entry of the resulting ideal Gram matrix then is $y_i y_j$. Therefore the kernel-target alignment criterion applied to a σ-tunable kernel is given by (22), namely, $\sigma^* = \arg\max_\sigma \langle K_\sigma, K^*\rangle_F / n\|K_\sigma\|_F$ where K_σ denotes the Gram matrix of the tunable kernel and K^* the ideal target matrix.

By considering the reproducing kernel associated with the RGK time-frequency distribution, we can further simplify the resulting optimization problem. For this purpose, we can write it as an optimization problem that consists of maximizing the numerator of the objective function in (22) subject to setting its denominator to a constant. This optimization problem is given by (23) subject to (24), where υ_0 is a normalization parameter. By injecting the RGK-based reproducing kernel, we can write the objective function as (25). We get an objective function similar to the one proposed in [1] within a signal analysis context. The signal-dependent entry in the latter, say $|A_x(r,\theta)|^2$, is substituted here by the so-called equivalent representation defined by $|A_{\text{eq}}(r,\theta)|^2 = \sum_{i,j} y_i y_j\, A_{x_i}(r,\theta)\, \overline{A}_{x_j}(r,\theta)$. We can therefore take advantage of the alternate-projection technique proposed in [1]. In particular, we relax the constraint (24) and replace it with a low-cost computation constraint on the volume of the parameter function as follows: $\int \sigma^2(\theta)\, d\theta = \upsilon_0'$. Compared with [1], it is worth noting that the computational complexity of the technique remains unchanged once we have evaluated the equivalent representation.

In order to solve the optimization problem, we consider an alternate-projecting technique inspired from [1] with two updating rules at each iteration. Let σ_k be the bandwidth parameter at iteration k. At iteration $k+1$, we perform a gradient ascent update using expression (28), where μ_k is the step-size parameter controlling the convergence of the algorithm, and f the objective function to be maximized in (26). The gradient of f is defined by the vector of entries given by expression (29) for $\theta = 0, 1, \ldots, \ell-1$. In order to take into account the constraint in the optimization problem, we project the solution onto the set of admissible functions. This can be implemented by normalizing $\sigma_{k+1}(\theta)$ at each iteration step with $\|\sigma_{k+1}(\theta)\|/\upsilon_0'$.

In order to illustrate the relevance of our approach, we propose to study two classification problems. Each one consists of two classes of two hundred 64-sample signals with a linear frequency modulation embedded in a white

Gaussian noise. The signal-to-noise ratio is approximately -8 dB. In the first problem, signals are characterized by an increasing frequency modulation, from 0.1 to 0.25 for the first class, and from 0.25 to 0.4 for the second one. In the second problem, signals are characterized by an increasing frequency modulation from 0.1 to 0.4, and by a decreasing one from 0.4 to 0.1. Figure 1 shows the resulting RGK parameterization function. We observe that the regions in the Doppler-lag plan are relevant for the classification problems we have considered. The convergence of the algorithm is illustrated in Figure 2 where we show, at each iteration, the value of the alignment, averaged over 20 trials. In order to illustrate the relevance of the proposed method, we estimate the generalization error of an SVM classifier associated with the reproducing kernel resulting from our approach, and we compare it with the one associated with the Wigner distribution. Table 1 shows that not only the classification error is reduced compared to the Wigner distribution, but also the number of Support Vectors which is divided by two.