SIGNAL-DEPENDENT TIME-FREQUENCY REPRESENTATIONS FOR CLASSIFICATION USING A RADIALLY GAUSSIAN KERNEL AND THE ALIGNMENT CRITERION

Paul Honeine, Cédric Richard

Institut Charles Delaunay (FRE CNRS 2848) - LM2S - Université de Technologie de Troyes,
12 rue Marie Curie, BP 2060, 10010 Troyes cedex, France - fax. +33.3.25.71.56.99
paul.honeine@utt.fr (tel. +33.3.25.71.56.25) – cedric.richard@utt.fr (tel. +33.3.25.71.58.47)
www.cedric-richard.fr

ABSTRACT

In this paper, we propose a method for tuning time-frequency distributions with radially Gaussian kernel within a classification framework. It is based on a criterion that has recently emerged from the machine learning literature: the kernel-target alignement. Our optimization scheme is very similar to that proposed by Baraniuk and Jones for signal-dependent time-frequency analysis. The relevance of this approach of improving time-frequency classification accuracy is illustrated through examples.

1. INTRODUCTION

Bilinear time-frequency distributions provide a powerful tool for non-stationary signal analysis since they reveal their time-varying frequency content. Amongst the myriad of existing distributions, selecting the optimal one for a given application has always been a critical issue. The most notorious is the adaptive distribution by Baraniuk and Jones, whose radially Gaussian kernel (RGK) is designed to smooth interference terms that usually limit the interpretability of time-frequency representations while preserving signal components [1, 2]. This distribution was also used in [3] to improve performance of time-frequency based classifiers.

Over the last decade, the theory of reproducing kernels has made a major breakthrough in the field of pattern recognition. This has led to new algorithms, such as Support Vector Machines (SVM), with improved performance and lower computational cost [4]. In a recent work [5], we extended this framework to time-frequency analysis and proposed a new class of powerful tools for non-stationary signal analysis and classification. In that case, reproducing kernels simply correspond to an inner product between time-frequency distributions. Choosing a suitable time-frequency distribution for a given classification problem can then be seen as optimal reproducing kernel selection. A solution to objectively pick time-frequency distributions that best facilitate the classification task has recently been developed in [6] through the concept of kernel-target alignment. This criterion was originally proposed in [7] to select appropriate reproducing kernels for classification without training any kernel machine.

In this paper, we use the alignment criterion to estimate the parameters of a time-frequency distribution with RGK kernel. The problem of maximizing this criterion reveals to be similar to that proposed in [1] for signal analysis. The gradient ascent algorithm introduced in this paper is therefore adapted to a classification framework. This paper is organized as follows. We start by introducing time-frequency distributions, in particular the RGK distribution, and their use in machine learning. Next, we propose a classification framework within which the parameters of RGK kernel are estimated through maximization of the alignment criterion. Finally, we illustrate the proposed approach.

2. TIME-FREQUENCY KERNEL MACHINES

Any time-frequency distribution of Cohen's class with a parameter function Φ is defined in the Doppler-Lag domain by

$$C_x(t,f) = \iint \Phi(\nu,\tau)\, A_x(\nu,\tau)\, \mathrm{e}^{-j2\pi(f\tau-t\nu)} \mathrm{d}\nu\, \mathrm{d}\tau,$$

where $A_x(\nu,\tau)$ is the ambiguity function of signal x given by $A_x(\nu,\tau) = \int x(t+\tau/2)\, \overline{x(t-\tau/2)}\, e^{2j\pi\nu t} \mathrm{d}t$ in rectangular coordinates. While the two-dimensional function Φ determines the properties of the distribution, one often seeks to parameterize it with a one-dimensional function, say σ, and denote it Φ_σ. This is the essence of the RGK time-frequency distribution [1], whose parameter function is defined by $\Phi_\sigma(\nu,\tau) = \mathrm{e}^{-(\nu^2+\tau^2)/2\sigma^2(\theta)}$, where $\theta = \arctan(\tau/\nu)$ is the angle between the radial line through (ν,τ) and the Doppler axis. It is more convenient to express the RGK kernel in polar coordinates[1], with $\Phi_\sigma(r,\theta) = \mathrm{e}^{-r^2/2\sigma^2(\theta)}$ where $r = \sqrt{\nu^2+\tau^2}$ is the radial variable, as opposed to the angular variable θ. The function $\sigma(\cdot)$ is called the spread function. It determines the shape of Φ_σ in the ambiguity plane, and the properties of the time-frequency distribution.

[1] For convenience of notation, we denote by $\Phi_\sigma(\nu,\tau)$ and $\Phi_\sigma(r,\theta)$ the function Φ_σ represented respectively in rectangular and polar coordinates of the ambiguity plane, as well as $A_x(\nu,\tau)$ and $A_x(r,\theta)$.

Kernel machines are non-linear pattern recognition techniques obtained from classical linear ones by using the kernel trick and a reproducing kernel. The latter corresponds to an inner product in a transformed space. Taking advantage of new theoretical advances, kernel machines are attractive by their reduced algorithmic complexity, mainly due to the *kernel trick*. This key idea exploits the fact that a great number of pattern recognition techniques does not depend explicitly of the data itself, but rather of their inner products. A generalization of these are the reproducing kernels, corresponding to an inner product of implicitly transformed data, while every reproducing kernel determines the transformation, up to a unitary transformation.

For non-stationary signal analysis and classification, a natural class of signal transformations are Cohen's class of time-frequency distributions. Given any pair of signals (x_i, x_j), the reproducing kernel associated to such spaces of transformations can be expressed by

$$\kappa(x_i, x_j) = \iint |\Phi(\nu,\tau)|^2 \; A_{x_i}(\nu,\tau) \; \overline{A_{x_j}(\nu,\tau)} \; \mathrm{d}\nu \; \mathrm{d}\tau,$$

where the ambiguity functions are in rectangular coordinates, or equivalently in polar coordinates

$$\kappa(x_i, x_j) = \iint r \; |\Phi(r,\theta)|^2 \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)} \; \mathrm{d}r \; \mathrm{d}\theta.$$

In the particular case of the RGK distribution, its reproducing kernel is obtained by injecting $\Phi_\sigma(r,\theta) = \mathrm{e}^{-r^2/2\sigma^2(\theta)}$ in the expression above, which leads to the expression

$$\kappa_\sigma(x_i, x_j) = \iint r \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)} \; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}} \; \mathrm{d}r \; \mathrm{d}\theta, \quad (1)$$

where the ambiguity functions are in polar coordinates. The use of this reproducing kernel allows a wide class of pattern recognition methods to operate on the RGK distribution, as studied in [5] for other time-frequency distributions. In what follows, we consider a criterion initially proposed within the framework of kernel machines, in order to optimize the parameters of the reproducing kernel (1), and therefore the corresponding RGK distribution.

3. CLASSIFICATION-DEPENDENT TIME-FREQUENCY DISTRIBUTION

We consider a 2-class classification problem of signals, from a training set $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ of n signals x_k with their labels $y_k = \pm 1$. Let K_σ be the Gram matrix of the training set, whose (i,j)-th entry is $\kappa_\sigma(x_i, x_j)$ defined in equation (1), and K_t the target matrix whose (i,j)-th entry is $y_i y_j$ (product of the outputs of the ideal classifier, given the input x_i and x_j). To measure the similarity between the reproducing kernel and the class labels, we consider the kernel-target alignment, defined by

$$\mathcal{A}(K_\sigma, K_t) = \frac{\langle K_\sigma, K_t \rangle_F}{\|K_\sigma\|_F \|K_t\|_F}, \quad (2)$$

where $\langle \cdot, \cdot \rangle_F$ is Frobenius scalar product, defined by summing up the products of the corresponding components of both input matrices, and $\|\cdot\|_F$ its norm, that is $\|\cdot\|_F^2 = \langle \cdot, \cdot \rangle_F$. In [8], Cristianini *et al.* proposed to select appropriate reproducing kernels by maximizing this score. Theoretical and experimental results show that good generalization performance may be expected by using kernels with large alignment score [7]. Note that this criterion does not require any computational intensive stage for designing and testing classifiers.

The optimal spread function $\sigma^*(\cdot)$ is determined by maximizing the alignment score, with

$$\sigma^* = \arg\max_\sigma \mathcal{A}(K_\sigma, K_t).$$

This can be formulated as a constrained optimization problem where the numerator of (2) is maximized subject to a constant denominator, namely,

$$\max_\sigma \sum_{i,j=1}^{n} y_i \, y_j \, \kappa_\sigma(x_i, x_j), \quad (3)$$

subject to

$$\sum_{i,j=1}^{n} \kappa_\sigma(x_i, x_j)^2 = V_0, \quad (4)$$

where V_0 is a preset normalization parameter. By expanding the objective functional in (3), we can write

$$\begin{aligned} &\sum_{i,j=1}^{n} y_i \, y_j \, \kappa_\sigma(x_i, x_j) \\ &\quad = \sum_{i,j=1}^{n} y_i y_j \iint r \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)} \; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}} \; \mathrm{d}r \; \mathrm{d}\theta \\ &\quad = \iint r \Big[\sum_{i,j=1}^{n} y_i y_j \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)} \Big] \; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}} \; \mathrm{d}r \; \mathrm{d}\theta \end{aligned} \quad (5)$$

We obtain the same form of objective functional to be maximized as in [1], which was $\iint r |A_x(r,\theta)|^2 \; \mathrm{e}^{-r^2/\sigma^2(\theta)} \; \mathrm{d}r \; \mathrm{d}\theta$, where the signal dependent term $|A_x(r,\theta)|^2$ is substituted by the equivalent representation $\sum_{i,j} y_i y_j \; A_{x_i}(r,\theta) \; \overline{A_{x_j}(r,\theta)}$. Since the latter depends only on the training signals and their labels, we can evaluate it prior to any optimization scheme. Exactly the same algorithm as in [1] can then be used to solve this problem. In particular, we relax the computationally expensive constraint (4) by substituting it with a constraint on the volume of the parameter function, i.e., $\int \sigma^2(\theta) \mathrm{d}\theta = V_0'$ as recommended in [1]. We shall now describe the algorithm in more details.

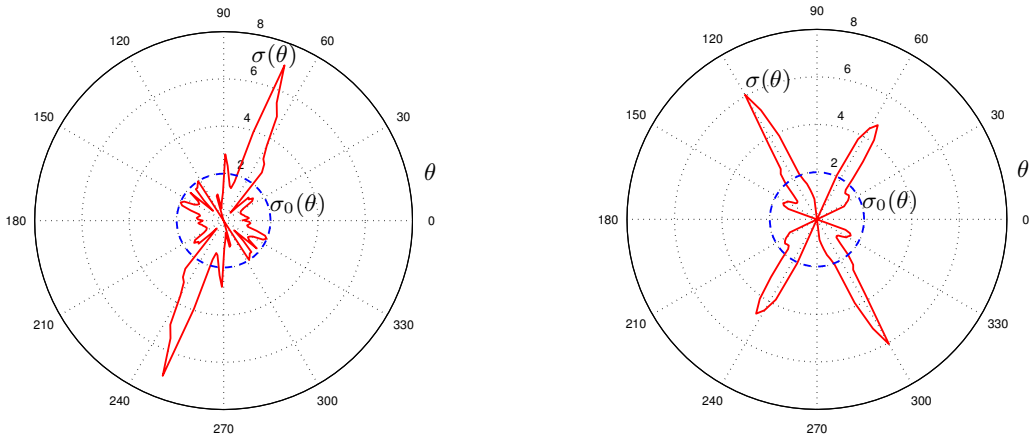

Fig. 1. Results obtained in the first (left) and second (right) studied cases. The optimal spread function is presented in polar coordinate, from the initial shape $\sigma_0(\theta)$ to the final one $\sigma(\theta)$, where the initial shape is obtained from the constraint $V_0' = 2$.

4. THE ALGORITHM

By considering discrete signals of l samples each, the resulting Doppler-Lag plan is often sampled on a rectangular grid, of size $l^2 \times l^2$. For the discrete RGK kernel, it is natural to consider a polar grid, obtained by an interpolation from the rectangular one, as described in [1]. With some abuse of notation, we define the resulting discrete RGK kernel by

$$\Phi_{\boldsymbol{\sigma}}(r,\theta) = \mathrm{e}^{-(r\Delta_r)^2/\boldsymbol{\sigma}^2(\theta)},$$

where henceforth r and θ designate the radial and angular[2] discrete variables, respectively, and $\Delta_r = 2\sqrt{\pi/l}$ is the radial step size. In this expression, $\boldsymbol{\sigma}(\theta)$ designates the θ-th entry of the spread vector, obtained by sampling the spread function. The reproducing kernel associated to the discrete RGK distribution can be written as

$$\kappa_{\boldsymbol{\sigma}}(x_i,x_j) = \sum_{r,\theta} r\, A_{x_i}(r,\theta)\, \overline{A_{x_j}(r,\theta)}\, \mathrm{e}^{-(r\Delta_r)^2/\boldsymbol{\sigma}^2(\theta)}.$$

From its continuous form (5), the discrete objective functional can be expressed as

$$\sum_{r,\theta} r\Big[\sum_{i,j=1}^{n} y_i y_j\, A_{x_i}(r,\theta)\, \overline{A_{x_j}(r,\theta)}\Big]\, \mathrm{e}^{-(r\Delta_r)^2/\boldsymbol{\sigma}^2(\theta)}, \quad (6)$$

while the constraint can be written as

$$\sum_{\theta} \boldsymbol{\sigma}^2(\theta) = V_0'. \quad (7)$$

To solve the (discrete) constrained optimization problem above, we adopt a classical iterative scheme with two steps, a gradient ascent step to maximize (6) and a projection step to take into account the constraint (7). Prior to the optimization step, we evaluate the equivalent representation of the training set,

$$\Psi(r,\theta) = \sum_{i,j=1}^{n} y_i y_j\, A_{x_i}(r,\theta)\, \overline{A_{x_j}(r,\theta)},$$

which leads to the objective functional $f_{\boldsymbol{\sigma}} = \sum_{r,\theta} r\, \Psi(r,\theta)\ \mathrm{e}^{-(r\Delta_r)^2/\boldsymbol{\sigma}^2(\theta)}$. The gradient of this functional evaluated at the spread vector $\boldsymbol{\sigma}(\theta)$ can be written as

$$\nabla f_{\boldsymbol{\sigma}} = \left[\frac{\partial f_{\boldsymbol{\sigma}}}{\partial \boldsymbol{\sigma}(0)}, \cdots, \frac{\partial f_{\boldsymbol{\sigma}}}{\partial \boldsymbol{\sigma}(l-1)}\right], \quad (8)$$

where, for any value of $\theta = 0, \cdots l-1$, we have

$$\frac{\partial f_{\boldsymbol{\sigma}}}{\partial \boldsymbol{\sigma}(\theta)} = \frac{2\Delta_r^2}{\boldsymbol{\sigma}^3(\theta)} \sum_{r} r^3\, \Psi(r,\theta)\, e^{-(r\,\Delta_r)^2/\boldsymbol{\sigma}^2(\theta)}.$$

We are now in a position to apply a gradient ascent scheme to maximize the objective functional $f_{\boldsymbol{\sigma}}$, with the updating recursion defined at iteration $k+1$ by

$$\boldsymbol{\sigma}_{k+1} = \boldsymbol{\sigma}_k + \mu_k\, \nabla f_{\boldsymbol{\sigma}_k},$$

where μ_k is a step-size control parameter.

While this allows an increase in the objective function, such a recursion yields an increase in the volume of the RGK kernel, and thus a violation of the constraint (7). To take into consideration this constraint, we project the solution onto the feasible set of spread vectors verifying (7), which can be written as $\|\boldsymbol{\sigma}_{k+1}\| = V_0'$ for the $k+1$ iteration, where $\|\cdot\|$ is the vector Euclidean norm. This can be done by rescaling the spread vector at each iteration with $\|\boldsymbol{\sigma}_{k+1}\|/V_0'$. This step does not affect the shape of the RGK kernel corresponding to this spread vector. Now we are ready to do some experimentations.

[2] In its discrete version, $\theta = 0, \cdots, l-1$ spans $[0,\pi]$, which can be extended into the whole Doppler-Lag domain by considering the symmetry in the latter.

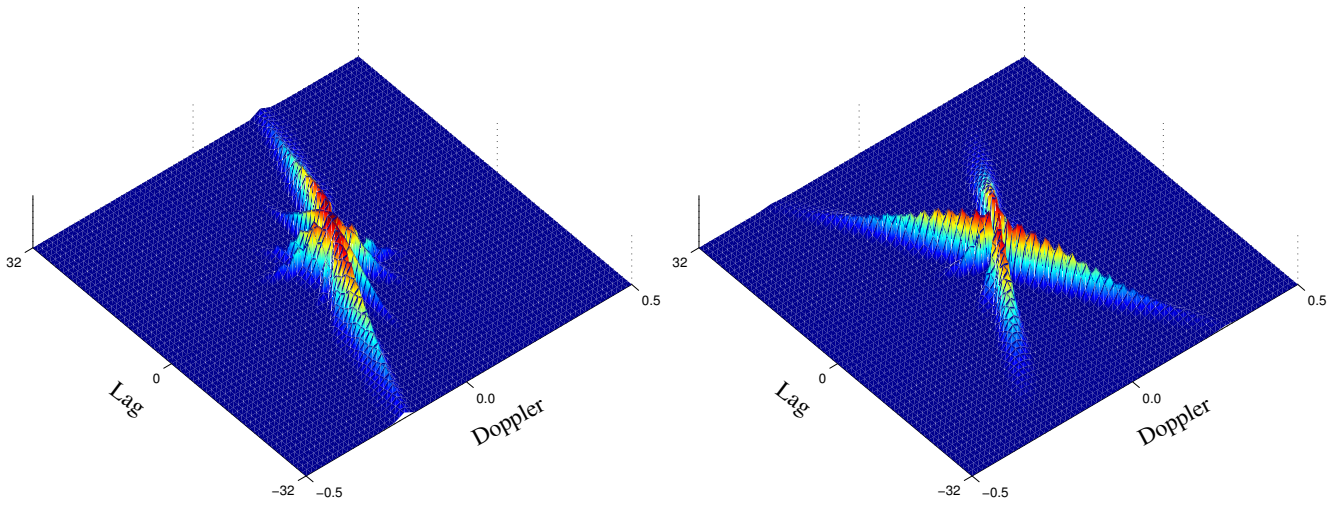

Fig. 2. The optimal RGK kernel obtained in the first (left) and second (right) studied cases.

5. SIMULATIONS

In what follows, we illustrate the proposed approach with two classification problems, each consisting of two sets of 200 signals of 64 samples, with a monocomponent embedded in a white Gaussian noise background of variance 4, leading to a signal-to-noise ratio of -9 dB. In the first case, signals have a linear frequency modulation (chirp) for the first class, from 0.2 to 0.4 (in normalized frequency), and signals from the second class have a Gaussian atom, at time and frequency indexes 32 and 0.2, respectively. In the second case, the signals have a chirp of increasing modulation for the first class, from 0.1 to 0.4, and a decreasing modulation, from 0.4 to 0.1 for the second class. In both cases, the relevant regions for the classification are overlapping in the time domain, as well as in the frequency domain, while they are distinguished in the time-frequency plane.

By applying the proposed optimization algorithm to each case, we got optimal RGK functions that correspond to relevant regions, in terms of classification, of the ambiguity domain. We illustrate the shape of the optimal spread function in Fig. 1(a) and in Fig. 1(b), obtained from the first and second studied cases, respectively. These shapes determine the relevant regions for classification, in the Doppler-Lag plan. The resulting optimal RGK obtained from these spread functions, are illustrated in Fig. 2(a) and in Fig. 2(b), for both studied cases. We underline the fact that these results are obtained from the algorithm with the constraint (4) on the volume being substituted by $\|\boldsymbol{\sigma}\|/V_0'$, where in simulations we set $V_0' = 2$. To illustrate the convergence of the algorithm, we represent in Fig. 3 the evolution of the alignment score at each iteration, averaged over 20 realizations.

The relevance of using the kernel-target alignment criterion to improve classification is illustrated with an SVM classifier

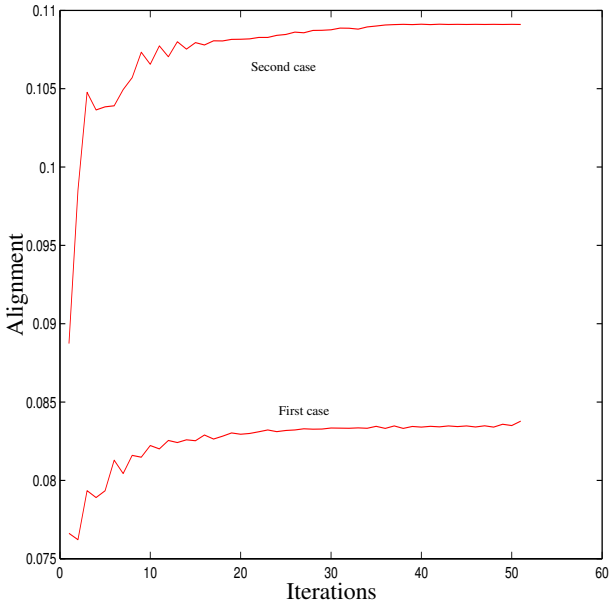

Fig. 3. Evolution of the alignment score versus iterations, averaged on 20 realizations, for the first (lower) and second (upper) studied cases.

associated with the Wigner distribution or the optimal RGK distribution, for each studied case. Table 1 represents the error rate, estimated on a test set of 2000 signals, and the number of support vectors, both averaged over 20 realizations. Note that the optimal RGK distribution minimizes the classification error, and also results in almost half the number of support vectors as compared to the Wigner distribution. This is mainly due to the optimality of the resulting distribution on the one hand, and on the other hand to its regularity, i.e. robustness caused by reduced interference terms.

First case	Error rate (%)	Number of *SV*
Wigner distribution	23.9 ± 1.3	172.6 ± 5.5
Optimal distribution	21.4 ± 2.0	87.5 ± 4.1

Second case	Error rate (%)	Number of *SV*
Wigner distribution	19.4 ± 1.5	163.7 ± 4.6
Optimal distribution	17.8 ± 1.5	84.5 ± 6.1

Table 1. Comparison of the error rates (%) and the number of support vectors (SV) obtained from a SVM algorithm applied to the Wigner distribution on the one hand, and to the optimal RGK distribution on the other hand, for both studied cases.

6. CONCLUSION

In this paper, we presented a classification-dependent time-frequency distribution, based on the kernel-target alignment. We showed that this strategy can benefit of a previously proposed optimization scheme for constructing signal-dependent distributions, by using the radial Gaussian kernel.

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

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

[6] 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.

[7] 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, 2001.

[8] 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.