KERNEL PCA BASED ESTIMATION OF THE MIXING MATRIX IN LINEAR INSTANTANEOUS MIXTURES OF SPARSE SOURCES

Frédéric Desobry and Cédric Févotte

Signal Processing Group
Engineering Department, University of Cambridge
Trumpington St, Cambridge CB2 1PZ, UK.
http://www-sigproc.eng.cam.ac.uk/
E-mail : {fd238,cf269}@eng.cam.ac.uk

ABSTRACT

Source sparsity based methods have become a common approach to blind source separation (BSS) problems, especially in the underdetermined case (more sources than sensors). If the sources are not sparse in the time-domain, in most cases, they can be mapped to a transformed domain (*e.g.*, wavelets, time-frequency, Fourier) in which this assumption is verified. In this paper, we are solely interested in the estimation of the mixing matrix. As observed by Zibulevski and coauthors, the data represented in the scatter plot of the observations tend to cluster along the mixing matrix columns. Each column can be seen as one of the principal components of the data in a higher (possibly infinite) dimension space, and these components can be estimated with a Kernel Principal Component Analysis (KPCA) based approach. The theoretical framework is derived, and excellent performance is observed both on synthetic and audio signals.

1. INTRODUCTION

Blind Source Separation (BSS) consists in estimating n signals (the sources) from the sole observation of m mixtures of them (the observations). In this paper we consider linear instantaneous mixtures, such that

$$\mathbf{x}_t = \mathbf{A}\,\mathbf{s}_t \tag{1}$$

where $\mathbf{x}_t = [x_{1,t}, \dots, x_{m,t}]^T$ is the vector containing the observations at time $t = 1, \dots, N$, $\mathbf{s}_t = [s_{1,t}, \dots, s_{n,t}]^T$ is the vector containing the sources, $\mathbf{A}$ is the $m \times n$ mixing matrix.

A now common approach to source separation consists of using *source sparsity* assumptions, in particular for underdetermined mixtures ($m < n$). This approach has been thoroughly developed in several papers from Zibulevsky and coauthors (see *e.g.* [1, 2]), with preliminary ideas found in [3]. The assumption of sparsity means that only a few coefficients of the sources are significantly non-zero. If the sources are not sparse in their original domain (*e.g.*, the time domain for audio signals), they might be sparse in a transformed domain (*e.g.*, the Fourier domain, wavelet transform).

Source separation methods exploiting source sparsity fall into two categories. One category consists of methods estimating jointly the mixing matrix and the sources under appropriate models, and usually within a Bayesian framework (see [4] and references therein). The second category consists of two steps methods: first estimate the mixing matrix, then infer the sources from the observations and the estimated mixing matrix. If the mixture is determined, the second step simply consists of applying the inverse of the mixing matrix estimate to the observations. If the mixture is underdetermined, the sources can be either estimated from heuristic approaches (based on projection of the observations onto the closest mixing matrix column [3,5–7]) or from Bayesian approaches using a Laplacian source prior (which boils down to shortest path decomposition using linear programming [1, 2]), a mixture prior (a weighted sum of a Dirac at 0 and another distribution) [5, 8], or a Student t prior [4].

Frédéric Desobry acknowledges support from the Boeing company under the University of Cambridge grant RG40562. Cédric Févotte acknowledges support from the European Commission funded Research Training Network HASSIP (HPRN-CT-2002-00285).

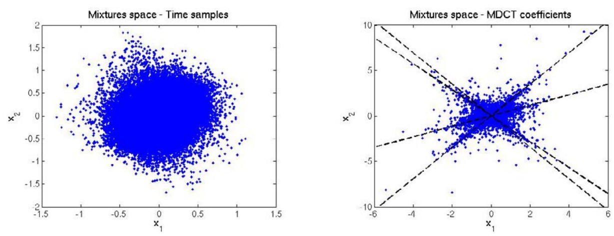

Fig. 1. (a) Scatter plot of x_1 w.r.t x_2 in the time-domain. (b) Scatter plot of x_1 w.r.t x_2 in the MDCT domain - the dashed lines represent the directions of the mixing matrix.

In this paper we are solely interested in the estimation of the mixing matrix. The interest of this task is twosome: 1) it corresponds to the first step of the methods in the second category, 2) it can provide a good initialization point of the mixing matrix and thus accelerate the convergence of the methods of the first category.
Because of the sparsity of the sources, the observations tend to cluster along the mixing matrix columns. This is illustrated on Fig. 1. The left plot shows a scatter plot of the two observations of a 2×4 mixture of four audio sources in the time domain. The right plot shows the scatter plot of the observations after a Modified Discrete Cosine Transform (a projection on a family of local cosines) was applied to them. The latter plot clearly shows three *clusters*, colinear to the mixing matrix columns.

Thus, estimating the mixing matrix in a mixture of sparse sources amounts to identify linear clusters passing through the origin (if the sources are zero-mean) and then merely defined by their angle to the origin. When the sources are sparse the distribution of $\theta_t = \tan^{-1}(x_{2,t}/x_{1,t})$ is thus n-modal, and many papers have addressed our task as a problem of estimating the modes of this distribution from a finite set of values. This is the approach adopted by Vielva, Erdoğmuş, Príncipe and co-authors in [7–11]. More precisely in [8, 9] the authors use a Parzen kernel based estimation of the density and gradient-based search of its modes, which is similar to the approach of Bofill and Zibulevsky described in [2] where the authors search the modes of a *potential function* of θ_t. A technique inspired from spectral analysis is described in [10]. The latter papers consider the $m = 2$ case, generalizations to $m > 2$ are studied in [7, 11].

Other works addressing the estimation of $\mathbf{A}$ include the simple nearest neighbors classification method mentioned in [12], a method inspired from image processing techniques for edge detection in [3], and finally a method based on self organizing maps in [6].

Our approach is based on nonlinear Principal Component Analysis: the lines in the scatter plot are principal components of the data, and the estimates of the mixing columns are the first n components. We derive an original extension of Kernel Principal Component Analysis in reproducing kernel Krein spaces. The remainder of this article is organized as follows: in Section 2.1, we recall basics about reproducing kernel Krein spaces, then in Section 2.2 we present the chosen nonlinear Principal Component Analysis method, namely Kernel Principal Component Analysis (KPCA). In Section 2.3, we describe our choice of a (reproducing) kernel. Simulation results are presented for both synthetic and audio sources in Section 3.

2. KERNEL PCA FOR MIXING MATRIX ESTIMATION

In the following we assume that the sources $s_1, \ldots, s_n$ in Eq. (1) are, to some extent, sparse. If the sources in their original domain (time, image) are not sparse *per se*, we assume that a linear transform yielding a sparse representation of sources (short-time Fourier, local cosine transforms, wavelets) was applied *a priori* to the observations. For the sake of clarity, we present our approach for $m = 2$, though it remains valid for any m. The algorithm for the use of KPCA for possibly underdetermined matrix identification is presented in Algorithm 1.

Algorithm 1: KPCA for mixing matrix estimation.

Step 0: Initialization

- Set η, θ_0 (m, N and n are known);
- Define reproducing kernel $k(\cdot,\cdot)$ as in Eq. (6);
- Define $\mathbf{x}_t = [x_{1,t}\ x_{2,t}]^T$; perform shrinkage with threshold η on the $\|\mathbf{x}_t\|_{\mathcal{X}}$'s; map (by symmetry) all remaining $\mathbf{x}_t$'s to the positive half-plane.

Step 1: Krein kernel principal component analysis

- Define matrix $K = [k_{tt'}]_{t,t'=1,\ldots,N}$ as in eq (4);
- Find the eigenvalues and eigenvectors for K.

Step 2: Estimation of the mixing matrix A

- Keep the eigenvectors corresponding to the n largest eigenvalues;
- Express the corresponding n directions in $\mathcal{X}$ using Eq. (5).

2.1. Reproducing kernel Krein spaces

Reproducing kernel Hilbert spaces (RKHS) now are a common tool in Signal Processing and Machine Learning, and their framework has given birth to a whole family of algorithms, known as *kernel methods* [13]. However, in many applications (such as ours, see Section 2.3), the required positive definiteness of the reproducing kernel (rk) is too restrictive, and this constraint cannot be met. It is however possible to define more general reproducing kernel spaces, such as reproducing kernel Krein spaces (RKKS) (see, *e.g.*, [14, 15]).

A Krein space[1] $(\mathcal{K}, \langle\cdot,\cdot\rangle_{\mathcal{K}})$ is a vector space which can be decomposed as the sum of a Hilbert space $(\mathcal{H}_+, \langle\cdot,\cdot\rangle_{\mathcal{H}_+})$ and a Hilbert antispace $(\mathcal{H}_-, -\langle\cdot,\cdot\rangle_{\mathcal{H}_-})$, with the inner product in $\mathcal{K}$ defined as $\langle\cdot,\cdot\rangle_{\mathcal{K}} = \langle\cdot,\cdot\rangle_{\mathcal{H}_+} - \langle\cdot,\cdot\rangle_{\mathcal{H}_-}$. As in the more popular Hilbert framework, an RKKS is a Krein space in which the evaluation functional

$$\begin{array}{cccc} \delta_{\mathbf{x}}: & \mathcal{H} & \to & \mathbb{R} \\ & f & \mapsto & f(\mathbf{x}) \end{array} \quad \text{for } \mathbf{x} \in \mathcal{X} \qquad (2)$$

is continuous; its rk being expressed, *e.g.*, as: $k(\cdot,\cdot) = k_+(\cdot,\cdot) - k_-(\cdot,\cdot)$, with k_+ (resp., k_-) the reproducing kernel in $\mathcal{H}_+$ (resp., $\mathcal{H}_-$). The following proposition in [14] gives a condition for a kernel to be a Krein rk:

Proposition 1: Condition for Krein reproducing kernel, from [14]. *If $k(\cdot,\cdot)$ is a symmetric function with values in $\mathbb{R}$, the following assertions are equivalent: (i) $k(\cdot,\cdot)$ is the reproducing kernel of a Krein space $\mathcal{K}$ of functions on $\mathcal{X}$; (ii) There exists a positive definite $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that l dominates k (*i.e.*, $l - k$ is positive definite).*

Calculations in RKKSs are very similar to what is commonly done in a RKHS, as the reproducing property holds:

$$\forall f \in \mathcal{H}, \forall x \in \mathcal{X}, \langle f(\cdot), k(\mathbf{x},\cdot)\rangle_{\mathcal{H}} = f(\mathbf{x}) \qquad (3)$$

This implies that functions in $\mathcal{K}$ can be expressed as linear combinations of kernels. Basically, the main difference between Krein and Hilbert RKSs lies in the existence of zero-norm non-zero functions, which makes the definition of projections more complex. This also alters the use of norm of differences of functions and the search of quadratic forms optima. We refer to [17] for deeper insight on these issues.

2.2. Kernel principal component analysis

Kernel PCA is a functional generalization of Principal Component Analysis (PCA), as are Isomap, Locally Linear Embedding or spectral clustering methods. It allows for as many principal components as there are data samples in the training set, and these directions can be nonlinear.

The rationale of our approach is to consider that the straight lines we need to recover from the n-dimension space (see Fig. 1) are principal components of the data. Though they are linear, PCA cannot be used as:

1. The directions we are looking for are not orthogonal;
2. There are more directions to find than the dimension of the space (underdetermined context).

Kernel PCA was initially defined in RKHSs (see, *e.g.*, [13]), but the same framework applies to any reproducing kernel space, provided it is endowed with a representer theorem.

We refer to Algorithm 1 for a precise description of our algorithm involving KPCA in an underdetermined context. Basically,

[1]For further details on (resp. reproducing) Krein spaces, we refer to [16] (resp., [14, pp. 234-254] and [15]).

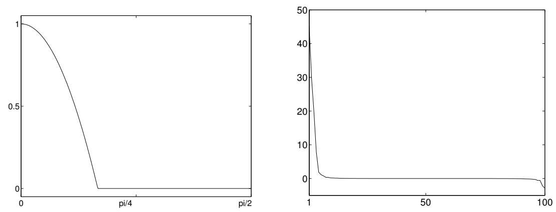

Fig. 2. Reproducing kernel $k(\cdot,\cdot)$ (see Eq. (6)) plotted as a function of $\widehat{\mathbf{x}_t, \mathbf{x}_{t'}}$ (left); eigenvalues for $k(\cdot,\cdot)$ (right).

for a given (reproducing) kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and data set $\{\mathbf{x}_t\}_{t=1,\ldots,N}$, one performs the eigendecomposition of the kernel matrix defined as:

$$K = [k_{tt'}]_{t,t'=1,\ldots,N} \qquad \text{with:} \qquad k_{tt'} = k(\mathbf{x}_t, \mathbf{x}_{t'}) \tag{4}$$

(Recall that in PCA, this task is performed over the correlation matrix $E[\mathbf{x}\mathbf{x}^T]$). The directions corresponding to the n largest eigenvalues are estimates of the lines in the scatter data.

We still need to assure that the eigenvectors of K can be recovered, *i.e.* that they can be expressed as a finite linear combination of kernels evaluated on the training data. This is obtained through the representer theorem in [18] which extends Wahba and Kimeldorf's to Krein reproducing spaces, therefore allows one to express these eigenvectors as:

$$\mathbf{v}_i(\mathbf{x}) = \sum_{t=1}^{N} \alpha_t^i k(\mathbf{x}, \mathbf{x}_t) \tag{5}$$

with $\{\alpha_1^i, \ldots, \alpha_N^i\}$ in $\mathbb{R}^N$, for any $\mathbf{x}$ in $\mathbb{R}^2$, and $\mathbf{v}_i$ the eigenvector corresponding to the i^{th} largest eigenvalue. Eq. (5) also proves that KPCA yields *linear* directions iff the kernel $k(\cdot,\cdot)$ itself is linear.

Next paragraph is dedicated to presenting the kernel used in Algorithm 1.

2.3. Kernel definition

The kernel used in KPCA needs to be:

- linear, as we are looking for lines in $\mathbb{R}^m$;
- parameterized so that KPCA can be tuned according to the problem difficulty (sparsity of the sources?, noisy data?);
- designed in the Krein framework (the obvious linear kernel $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ which is a Hilbert rk cannot be parameterized in terms of angles).

Denoting $(\widehat{\mathbf{x}_t, \mathbf{x}_{t'}})$ the angle between $\mathbf{x}_t$ and $\mathbf{x}_{t'}$, the kernel

$$k_{\theta_0}(\mathbf{x}_t, \mathbf{x}_{t'}) = \begin{cases} \frac{\cos(\widehat{\mathbf{x}_t, \mathbf{x}_{t'}}) - \cos\theta_0}{1-\cos\theta_0} & \text{if } (\widehat{\mathbf{x}_t, \mathbf{x}_{t'}}) \leq \theta_0 \\ 0 & \text{else.} \end{cases} \tag{6}$$

satisfies all the required conditions (we recall that if $k(\cdot,\cdot)$ is a RKKS kernel, then so is $(\mathbf{x}, \mathbf{x}') \mapsto \frac{k(\mathbf{x},\mathbf{x}')}{\sqrt{k(\mathbf{x},\mathbf{x})k(\mathbf{x}',\mathbf{x}')}}$, and that $\cos(\widehat{\mathbf{x}_t, \mathbf{x}_{t'}}) = \langle\mathbf{x}_t, \mathbf{x}_{t'}\rangle_{\mathcal{X}}$ as the $\mathbf{x}_t$'s have unit norm). This kernel is plotted in Fig. 2, along with its eigenvalues.

3. RESULTS AND DISCUSSION

3.1. Synthetic sources

We first consider synthetic sources generated as:

$$s_{i,t} \overset{\text{i.i.d.}}{\sim} p\delta_0 + (1-p)\mathcal{N}(0,1) \tag{7}$$

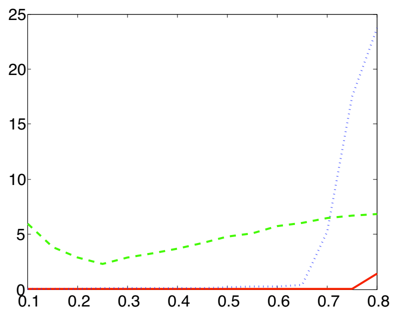

Fig. 3. Mean square error between the angles estimated by the three ICA approaches and θ, as a function of the sparsity index p, for KPCA- (line), density- (dotted) and clustering-based (dash) approaches. The approach introduced in this paper exhibits superior performance, and the error rate keeps low even when p reaches high values.

(for $i = 1, \ldots, n$, $t = 1, \ldots, N$). We consider a mixture of $n = 4$ sources with $m = 2$ observations, with mixing matrix $\mathbf{A}$ defined as

$$\mathbf{A} = \begin{bmatrix} \cos(\theta_1) & \cos(\theta_2) & \cos(\theta_3) & \cos(\theta_4) \\ \sin(\theta_1) & \sin(\theta_2) & \sin(\theta_3) & \sin(\theta_4) \end{bmatrix} \tag{8}$$

with $\boldsymbol{\theta} = [\theta_1\ \theta_2\ \theta_3\ \theta_4]^T = [-60^o\ \ -30^o\ 30^o\ 60^o]^T$. We use $N = 1000$ samples and study the performance of various methods upon several values of p, varying from 0.1 to 0.8 (by increments of 0.05). In the following, we give Monte Carlo estimates of the mean square error (MSE) $\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \widehat{\theta}_i)^2$ of the angles $\boldsymbol{\theta}$ over 100 realizations of the sources for each value of p.

We compare the behavior of our KPCA-based approach to two state-of-the-art methods: the clustering approach (C) in [2] and the density estimation approach (DE) found in [8, 9]. The various parameters are chosen as follows. In the (KPCA) method, the parameters are the kernel shrinking parameter $\theta_0 = .01^\circ$ and the threshold $\eta = 0$ (no shrinkage). Clustering (C) is performed via the *kmeans* algorithm. As in [8, 9], the density estimator is a Parzen estimator used with a Gaussian window with optimized width. The modes of the density are found by gradient-based optimization initialized with the n modes of the angles histogram.

Results are presented in Fig. 3 where the MSEs are plotted *vs* the sparsity index p. For $p \leq 0.6$, both our approach and (DE) exhibit excellent behavior, as virtually no error is made in the estimation of the angles vector $\boldsymbol{\theta}$. For values of p greater than 0.6, this remains correct only for our approach, as the MSE for (DE) starts increasing rapidly. The clustering approach presents an overall stable behavior with approximately a MSE of 5.

3.2. Audio signals

We now apply the above approaches to a mixture of $n = 4$ audio sources with $m = 2$ observations, with mixing matrix $\mathbf{A}$ defined as in Eq. (8) with: $\boldsymbol{\theta} = [-60^o\ \ -55^o\ 30^o\ 60^o]$. White Gaussian noise is added to the mixtures, yielding approximately 13dB SNR on each observation. As in [4], we first preprocess the audio signals by applying a Modified Discrete Cosine Transform (MDCT) to the observations. The MDCT is a projection on a family of local cosines and is known to provide sparse decomposition of audio signals (see, *e.g.*, [19]). The scatter plot of the noise-free observations is represented in Fig. 1.

Approaches	KPCA	C	DE
Noise-free signals	15	/	60
Signals with noise	23	/	66

Table 1. Mean square error between the angles estimated by kernel principal component analysis, clustering, and density estimation techniques.

We compare again the results obtained by (KPCA), (C) and (DE) on both the noise-free and noisy observations. For our approach, parameters are tuned as follows: (noise-free signals) $\eta = 0.4$ and $\theta_0 = 15$; (noisy signals) $\eta = 0.5$ and $\theta_0 = 1.5$.

The results are presented in Table 1. Our approach is more robust to the addition of noise that (DE), and yields better performance both on noisy and noise-free signals. The clustering approach does not converge in both situations.

3.3. Extension to $m > 2$

The extension to the more general case $m > 2$ is straightforward as the framework defined in Section 2 remains the same: what changes is not the dimension of the matrix K, which is $N \times N$, but the domain over which the rk $k(\cdot, \cdot)$ is defined. As the relation between the inner product and the angle is valid whatever the dimension of the space, no modification is even required in the definition of the kernel. Preliminary experiments involving m varying from 2 to 5 are not presented here for the sake of brevity but yielded excellent performance.

4. CONCLUSION

In this paper, we introduced an original approach for dealing with the estimation of possibly underdetermined mixing matrices in the source sparsity based framework proposed by Zibulevsky and coauthors. In the space spanned by $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, we identify the coefficients of the mixing matrix $\mathbf{A}$ as the principal components in the scatter plot of the data using Kernel Principal Component Analysis (KPCA). The need for a theoretical framework and the restrictive constraints inherent to the Hilbert context led us to introduce KPCA in a reproducing kernel Krein space. Experiments on synthetic sources as well as audio illustrate the excellent behavior of our approach. Further work will involve the estimation of the number of sources n.

5. REFERENCES

[1] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev, "Blind source separation by sparse decomposition," in *Independent Component Analysis: Principles and Practice*, S. J. Roberts and R. M. Everson, Eds. Cambridge University Press, 2001.

[2] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[3] J. K. Lin, D. G. Grier, and J. D. Cowan, "Feature extraction approach to blind source separation," in *Proceedings of IEEE Workshop on Neural Networks for Signal Processing (NNSP'97)*, Amelia Island Plantation, Florida, 1997, pp. 398–405.

[4] C. Févotte and S. Godsill, "A Bayesian approach to blind separation of sparse sources," *IEEE Transactions on Speech and Audio Processing*, 2005, To appear.

[5] L. Vielva, D. Erdoğmuş, and J. C. Príncipe, "Underdetermined blind source separation using a probabilistic source sparsity model," in *Proceedings of 3rd International Conference on independent component analysis and blind signal separation (ICA'01)*, San Diego, California, Dec. 2001, pp. 675–679.

[6] M. Van Hulle, "Clustering approach to square and non-square blind source separation," in *Proceedings of IEEE Workshop on Neural Networks for Signal Processing (NNSP'99)*, Madison, Wisconsin, Aug. 1999, pp. 315–323.

[7] L. Vielva, Y. Pereiro, D. Erdoğmuş, J. Pereda, and J. C. Príncipe, "Inversion techniques for underdetermined bss in an arbitrary number of dimensions," in *Proceedings of ICA'03*, Nara, Japan, Apr. 2003, pp. 131–136.

[8] L. Vielva, D. Erdoğmuş, C. Pantaleón, I. Santamaría, J. Pereda, and J. C. Príncipe, "Underdetermined blind source separation in a time-varying environment," in *Proceedings of ICASSP'02*, Orlando, Florida, May 2002, pp. 3049–3052.

[9] D. Erdoğmuş, L. Vielva, and J. C. Príncipe, "Nonparametric estimation and tracking of the mixing matrix for underdetermined blind source separation," in *Proceedings of 3rd International Conference on independent component analysis and blind signal separation (ICA'01)*, San Diego, California, Dec. 2001, pp. 675–679.

[10] L. Vielva, I. Santamaría, C. Pantaleón, J. Ibánez, D. Erdoğmuş, J. Pereda, and J. C. Príncipe, "Estimation of the mixing matrix for underdetermined blind source separation using spectral estimation techniques," in *Proceedings of EUSIPCO'02*, Toulouse, France, Sep. 2002, pp. 557–560.

[11] L. Vielva, I. Santamaría, D. Erdoğmuş, and J. C. Príncipe, "On the estimation of the mixing matrix for underdetermined blind source separation in an arbitrary number of dimensions," in *Proc. 5th International Conference on Independent Component Analysis and Blind Source Separation (ICA'04)*, Granada, Spain, Sep. 2004, pp. 185–192.

[12] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

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

[14] L. Schwartz, "Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés.," *Journal d'Analyse Mathématique*, pp. 115–256, 1964.

[15] X. Mary, *Hilbertian Subspaces, Subdualities and Applications*, Ph.D. thesis, INSA Rouen, 2003.

[16] J. Bognàr, *Indefinite inner product spaces*, Springer-Verlag, 1974.

[17] B. Hassibi, A. Sayed, and T. Kailath, "Linear estimation in krein spaces - part i: Theory," *IEEE Transactions on Automatic Control*, pp. 18–33, 1996.

[18] C.S. Ong, X. Mary, S. Canu, and A.J. Smola, "Learning with non positive kernels," in *ICML 2004*, 2004.

[19] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1998.