ADAPTIVE KERNEL CANONICAL CORRELATION ANALYSIS ALGORITHMS FOR MAXIMUM AND MINIMUM VARIANCE

S. Van Vaerenbergh, J. Vía, J. Manco-Vásquez and I. Santamaría

Department of Communications Engineering, University of Cantabria, Spain

ABSTRACT

We describe two formulations of the kernel canonical correlation analysis (KCCA) problem for multiple data sets. The kernel-based algorithms, which allow one to measure nonlinear relationships between the data sets, are obtained as nonlinear extensions of the classical maximum variance (MAXVAR) and minimum variance (MINVAR) canonical correlation analysis (CCA) formulations. We then show how adaptive versions of these algorithms can be obtained by reformulating KCCA as a set of coupled kernel recursive least-squares algorithms. We illustrate the performance of the proposed algorithms on a nonlinear identification application and a cognitive radio detection problem.

Index Terms— Kernel methods, canonical correlation analysis, recursive least-squares, adaptive filtering

1. INTRODUCTION

Canonical Correlation Analysis (CCA) is a well-known technique in multivariate statistical analysis. The original CCA formulation was introduced by Hotelling in 1936 [1] as a way to measure the linear relationship between two multivariate variables. Given two data sets, CCA retrieves the linear projections of both that are maximally correlated. Since its original formulation, many extensions have been proposed to the standard CCA technique. Among others, there exist several generalized versions of CCA that deal with multiple data sets [2, 3], and also extensions to nonlinear versions of CCA [3, 4, 5], in particular kernel canonical correlation analysis (KCCA). An adaptive version of linear generalized CCA was proposed in [6], which allows one to perform CCA online and in time-varying environments. The range of fields in which CCA has been applied is wide and varied, including economy, meteorology, functional magnetic resonance imaging (fMRI) [7, 8, 9], blind source separation [5, 10], multivariate regression [6] and communication theory [11, 12].

In this paper we propose an online framework for generalized kernel canonical correlation analysis. This framework builds upon the linear adaptive CCA formulation from [6], which uses a set of coupled recursive least-squares (RLS) algorithms, one per data set, to retrieve the linear CCA solution in an online manner. We will resort to kernel methods [5], and in particular to kernel adaptive filtering [13, 14], to build a nonlinear version of this framework.

In order to retrieve the KCCA solution, the adaptive framework uses a set of coupled kernel recursive least-squares (KRLS) algorithms. While there exist several different KRLS implementations, the proposed KCCA framework requires an algorithm with tracking capability, which has become possible through recent advances in the field of kernel adaptive filtering [15]. A hybrid algorithm based on a coupling of a RLS and a sliding-window KRLS algorithm, which has only limited tracking capability, was previously employed in [12] to nonlinear channel identification. In this paper we formalize the kernel-based approach, we derive two practical adaptive KCCA algorithms and we perform simulations using the more sophisticated KRLS-T algorithm from [15].

The rest of this paper is structured as follows: In Section 2 we provide an overview of two basic generalized KCCA formulations, followed by a derivation of their adaptive versions in Section 3. The results of two numerical experiments are reported in Section 4, and, finally, the main conclusions of this work are listed in Section 5.

2. GENERALIZED KERNEL CANONICAL CORRELATION ANALYSIS

We are given M data sets $\{\mathbf{x}_i(1), \mathbf{x}_i(2), \ldots, \mathbf{x}_i(N)\}$, $i = 1, \ldots, M$, each containing N multivariate data. Kernel methods require the data to be transformed into a high-dimensional feature space, $\mathbf{x} \rightarrow \Phi(\mathbf{x})$, where the corresponding Gram matrices (or "kernel" matrices) $\mathbf{K}_i$ can be calculated as

$$\mathbf{K}_i(j,k) = \Phi_i(\mathbf{x}_i(j))^\top \Phi_i(\mathbf{x}_i(k)) = \kappa_i(\mathbf{x}_i(j), \mathbf{x}_i(k)), \quad (1)$$

in which $\kappa_i(\cdot,\cdot)$ represents a kernel function. The problem of KCCA consists in finding the projections of the transformed data sets, $\mathbf{z}_i = \mathbf{K}_i\boldsymbol{\alpha}_i$, that have maximal correlation [3, 5]. The generalized canonical correlation between the M transformed data sets is defined as

$$\rho = \frac{1}{M}\sum_{i=1}^{M} \rho_i, \quad (2)$$

This work was supported by MICINN (Spanish Ministry for Science and Innovation) under grants TEC2010-19545-C04-03 (COSIMA) and CONSOLIDER-INGENIO 2010 CSD2008-00010 (COMONSENS).

in which

$$\rho_i = \frac{1}{M-1} \sum_{\substack{j=1 \\ j \neq i}}^{M} \rho_{ij} \tag{3}$$

is a measure of the correlation associated to the i-th data set, $\rho_{ij} = \mathbf{z}_i^\top \mathbf{z}_j = \boldsymbol{\alpha}_i^\top \mathbf{K}_i \mathbf{K}_j \boldsymbol{\alpha}_j$. The trivial solution is avoided by applying the following constraint on the energy of the canonical variates

$$\frac{1}{M} \sum_{i=1}^{M} \|\mathbf{z}_i\|^2 = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{\alpha}_i^\top \mathbf{K}_i \mathbf{K}_i \boldsymbol{\alpha}_i = 1. \tag{4}$$

Overfitting problems can be avoided by adding a regularization factor c to the norm of the projectors in this restriction,

$$\frac{1}{M} \sum_{i=1}^{M} \boldsymbol{\alpha}_i^\top \mathbf{K}_i \mathbf{K}_i \boldsymbol{\alpha}_i + c \boldsymbol{\alpha}_i^\top \mathbf{K}_i \boldsymbol{\alpha}_i = 1, \tag{5}$$

see [3]. By defining the matrices

$$\mathbf{R} = \begin{bmatrix} \mathbf{K}_1\mathbf{K}_1 & \cdots & \mathbf{K}_1\mathbf{K}_M \\ \vdots & \ddots & \vdots \\ \mathbf{K}_M\mathbf{K}_1 & \cdots & \mathbf{K}_M\mathbf{K}_M \end{bmatrix}, \tag{6}$$

$$\mathbf{D} = \begin{bmatrix} \mathbf{K}_1(\mathbf{K}_1 + c\mathbf{I}) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{K}_M(\mathbf{K}_M + c\mathbf{I}) \end{bmatrix}, \tag{7}$$

the solutions of the KCCA problem can now be found by solving the following generalized eigenvalue problem (GEV)

$$\frac{1}{M} \mathbf{R} \boldsymbol{\alpha} = \beta \mathbf{D} \boldsymbol{\alpha}, \tag{8}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top, \ldots, \boldsymbol{\alpha}_M^\top]^\top$ and $\beta = \frac{1+(M-1)\rho}{M}$. The canonical weights $\boldsymbol{\alpha}_i$ that provide the projections that admit the best possible one-dimensional representation are retrieved as the eigenvector corresponding to the largest eigenvalue of the GEV problem (8). This general formulation, which in the case of a linear kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$ corresponds to the maximum variance (MAXVAR) formulation from [2], reduces to the standard KCCA algorithm from [3] in case only two data sets are used. With a linear kernel and only two data it reduces to the classical CCA formulation from [1].

Analogously to the MAXVAR generalization of CCA, the minimum variance (MINVAR) generalization of CCA is defined as the problem of finding the projections that admit the best possible $(M-1)$-dimensional representation [2]. In the context of KCCA, the KCCA-MINVAR generalization is defined as the problem of *minimizing* the generalized canonical correlation ρ subject to restriction (5), and thus it amounts to retrieving $\boldsymbol{\alpha}$ corresponding to the minimum eigenvalue of (8). This formulation is used for instance in [5] for kernel independent component analysis, where the goal is to filter nonlinear transformations that minimize the mutual information among the recovered signals. In the next section we will derive adaptive algorithms for both the MAXVAR and MINVAR criteria.

3. ADAPTIVE KCCA

3.1. MAXVAR

The GEV problem (8) can be interpreted as a set of M coupled kernel regression problems

$$\beta(\mathbf{K}_i + c\mathbf{I})\boldsymbol{\alpha}_i = \mathbf{z}, \qquad i = 1, \ldots, M, \tag{9}$$

where $\mathbf{z} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{z}_i$ and $\mathbf{z}_i = \mathbf{K}_i\boldsymbol{\alpha}_i$, and we have assumed that $\mathbf{K}_i$ is invertible. For a given $\mathbf{z}$, the weights $\boldsymbol{\alpha}_i$ can thus be retrieved by performing kernel least-squares regression,

$$\boldsymbol{\alpha}_i = \frac{1}{\beta}(\mathbf{K}_i + c\mathbf{I})^{-1}\mathbf{z}. \tag{10}$$

Let us assume now that we are operating in an online scenario, in which one datum $\mathbf{x}_i(n)$ of each data set is made available per time step n. In this scenario a *recursive* solution can be formulated for Eq. (10), which is known in the literature as kernel recursive least-squares (KRLS). Nevertheless, KRLS requires that the output $z(n)$ is known during training. Interestingly, however, the KCCA framework implies that the M KRLS algorithms are coupled and that they should produce the same outputs $z(n)$. Eq. (9) indicates that this can be achieved by estimating $z(n)$ in each iteration as

$$z(n) = \frac{1}{M}\sum_{i=1}^{M} z_i(n), \tag{11}$$

where $z_i(n)$ is the output of the i-th KRLS algorithm evaluated on $\mathbf{x}_i(n)$. Given the estimate $z(n)$ from Eq. (11), the n-th iteration of the algorithm concludes by training each KRLS individually on its corresponding input-output data pair $(\mathbf{x}_i(n), z(n))$.

3.1.1. Kernel recursive least-squares

A few remarks need to be made on this procedure. First, due to space restrictions, we will not discuss the inner mechanics of KRLS in more detail, but rather refer to [15]. For the scope of this paper it is sufficient to know that KRLS solves Eq. (10) recursively. The only two operations used by the adaptive KCCA framework are: 1) training the KRLS algorithm on input-output data pairs, and 2) evaluating it on input data.

Furthermore, since the design of online kernel methods presents certain difficulties, such as growing matrices and complexities, several different implementations of KRLS have been proposed in the last decade. One of the research goals in this area has also been to design a KRLS algorithm capable of tracking. This is a necessary property in the proposed KCCA framework, since the initial estimates of $z(n)$

Initialize the target output $z(1)$ randomly.
Initialize the i-th KRLS with $(\mathbf{x}_i(1), z(1)), \forall i$.
for $n = 2, 3, \ldots$ **do**
  Receive $\mathbf{x}_i(t)$, the input to the i-th KRLS, $\forall i$.
  Obtain the corresponding output $z_i(n), \forall i$.
  Calculate $z(n)$ through (11).
  Center and normalize $z(n)$.
  if *MAXVAR* **then**
    Train the i-th KRLS with $(\mathbf{x}_i(n), z(n)), \forall i$.
  else if *MINVAR* **then**
    Calculate $r_i(n)$ through (15), $\forall i$.
    Train the i-th KRLS with $(\mathbf{x}_i(n), r_i(n)), \forall i$.
  end
end

Algorithm 1: Adaptive KCCA using MAXVAR/MINVAR.

are likely to be erroneous and KRLS must have a mechanism to "forget" these data over time. Among the existing KRLS algorithms only a few are truly adaptive in this sense. In the simulations of Section 4 we will use the recently proposed KRLS-T algorithm from [15], which combines all the necessary properties.

3.2. MINVAR

In order to retrieve the eigenvector corresponding to the minimum eigenvalue of the GEV problem (8), we first rewrite it as follows

$$\left(\mathbf{D} - \frac{1}{M}\mathbf{R}\right)\boldsymbol{\alpha} = \gamma \mathbf{D}\boldsymbol{\alpha}, \tag{12}$$

where $\gamma = 1 - \beta$. Analogously to the MAXVAR case, the GEV problem (12) can be interpreted as a set of M coupled kernel regression problems,

$$\gamma(\mathbf{K}_i + c\mathbf{I})\boldsymbol{\alpha}_i = \mathbf{r}_i, \qquad i = 1, \ldots, M, \tag{13}$$

where $\mathbf{r}_i = \mathbf{z}_i - \mathbf{z}$. The weights $\boldsymbol{\alpha}_i$ can therefore be retrieved by solving the kernel least-squares regression problem

$$\boldsymbol{\alpha}_i = \frac{1}{\gamma}(\mathbf{K}_i + c\mathbf{I})^{-1}\mathbf{r}_i, \qquad i = 1, \ldots, M, \tag{14}$$

when $\mathbf{r}_i$ is given. Hence, an adaptive KCCA-MINVAR algorithm can be obtained by applying M coupled KRLS algorithms, which use the following estimate of the output

$$r_i(n) = z(n) - \frac{1}{M}\sum_{i=1}^{M} z_i(n), \tag{15}$$

as indicated by Eq. (13).

The proposed KCCA-MAXVAR and KCCA-MINVAR algorithms are summarized in Alg. 1.

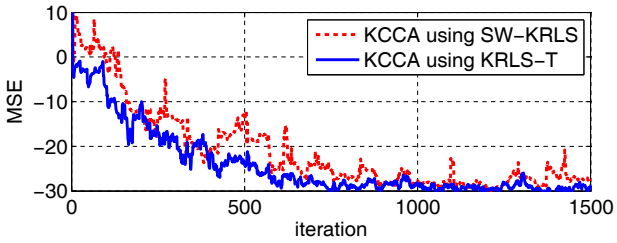

Fig. 1. Wiener system identification results of experiment 1.

4. NUMERICAL EXPERIMENTS

In this section we demonstrate the validity of the proposed algorithms through two different experiments. The KRLS algorithms used in these experiments are Matlab implementations from KMBOX[1]. The kernel we use is a radial basis function (RBF) kernel of the form $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2w^2)$, where w is the kernel width.

4.1. Wiener system identification

In the first experiment, we consider the online identification of a Wiener system, which is a block-based nonlinear system that consists of a static linear filter followed by a nonlinear channel. The experimental setup is taken from [12]. Specifically, the linear filter has impulse response $\mathbf{h} = [1, 0.3668, -0.5764, 0.2070]^\top$ and the nonlinearity is $y = \tanh(x)$.

In [12], a simplified version of the proposed algorithm was presented for two data sets, specifically designed for Wiener system identification. It couples a linear RLS algorithm , which identifies the linear channel, with a KRLS algorithm, which identifies the (inverse) nonlinearity. The KRLS algorithm used in [12] is sliding-window KRLS (SW-KRLS), which is capable of performing some tracking albeit with limited results. We repeat this experiment and compare the results to the proposed KCCA approach in which the newer KRLS-T algorithm from [15] is used, which is a more sophisticated tracker. Note that in the case of two data sets the formulations of KCCA-MAXVAR and KCCA-MINVAR coincide. For SW-KRLS and KRLS-T we use a memory of 20 bases, and the forgetting factor of KRLS-T is set to $\lambda = 0.99$. The remaining setup parameters can be found in [12].

The identification results are displayed in Fig. 1. The displayed MSE is measured between the true system's internal signal $z(n)$ and the estimate obtained by the RLS and KRLS algorithms. The results were averaged out over 10 simulations. As can be observed, using KRLS-T in adaptive KCCA has a positive effect, as it is capable of avoiding the error peaks produced by SW-KRLS.

[1] Available at `http://sourceforge.net/p/kmbox/`

4.2. Detection in cognitive radio

In the second experiment, we deal with a detection problem that uses a set of M sensors. Such detection problems appear for instance in multiantenna spectrum sensing for cognitive radio networks [16, 17, 18], where the secondary users perform spectral sensing in order to identify whether a wireless communication channel is in use by a licensed primary user or not. While in this scenario each sensor corresponds to a subchannel within the sensed spectrum band, one can easily imagine other applications such as sensor networks that fall under the description of this experiment.

Here, each sensor measures realizations of a zero-mean Gaussian distribution with variances $\boldsymbol{\sigma}$. We consider the following hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_0)), \qquad \boldsymbol{\sigma}_0 = [\sigma_{0,1}, \ldots, \sigma_{0,M}], \\ \mathcal{H}_1 &: \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_1)), \qquad \boldsymbol{\sigma}_1 = [\sigma_{1,1}, \ldots, \sigma_{1,M}]. \end{aligned} \tag{16}$$

Under hypothesis $\mathcal{H}_0$, the pdf measured by the i-th sensor is $\mathcal{N}(0, \sigma_{0,i}^2)$, for $i = 1, \ldots, M$. Under hypothesis $\mathcal{H}_1$, it is $\mathcal{N}(0, \sigma_{1,i}^2)$. Note that the measurements of the different sensors are conditionally independent given the hypothesis. For reasons of simplicity, we assume that $\sigma_{0,i} < \sigma_{1,i}$. If the variances $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}_1$ are known, one can apply the classical Neyman-Pearson test on each sensor individually, which in this case decides $\mathcal{H}_1$ if the test statistic

$$x_i[n](\sigma_{1,i} - \sigma_{0,i})x_i[n] \tag{17}$$

is greater than a threshold τ_i, where $x_i[n]$ is the sample received by sensor i at time n (see [19, Chapter 3] for further details). By exploiting the knowledge that all measurements at a given time follow the same hypothesis, the individual tests can be combined into the following optimal test

$$\sum_{i=1}^{M} x_i[n](\sigma_{1,i} - \sigma_{0,i})x_i[n] > \tau. \tag{18}$$

We consider the more challenging case in which the variances $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}_1$ are unknown. In this case, we can only exploit the knowledge that the test statistics should be correlated for all sensors. Since this implies that the individual hypothesis tests should be correlated, CCA can be applied. As discussed above, however, the optimal test statistics (17) are nonlinear (quadratic) functions of the data, and therefore the solution requires to use a nonlinear kernel.

The experimental setup is as follows: $M = 3$ sensors are considered, and $N = 300$ samples are used to blindly learn the hypothesis tests. The true variances of each distribution under the different hypotheses are $\boldsymbol{\sigma}_0 = [0.5, 0.5, 0.5]$ and $\boldsymbol{\sigma}_1 = [0.5, 2, 2]$. In particular, the variances for one of the sensors coincide under both hypotheses, while the variances for the other sensors are significantly different.

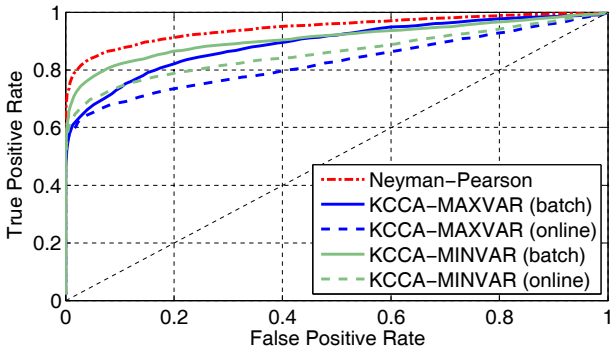

Fig. 2. Detection ROCs for experiment 2.

We compare the results of KCCA-MAXVAR and KCCA-MINVAR, both first in batch mode and then in online mode. The parameters of the batch and online algorithms are chosen as follows: the kernel width is fixed as $w = 0.5$, and the regularization is set to $c = 10^{-5}$. For adaptive KCCA the KRLS-T algorithm from [15] is used, with forgetting factor $\lambda = 0.999$ and a limited memory of 20 bases.

The receiving operator characteristic (ROC) curves of all algorithms, calculated on a test set of 10.000 points, are shown in Fig. 2. Several observations can be made. First of all, the ROCs of both batch KCCA algorithms fall very close to the optimal (Neyman-Pearson) detector, which has complete knowledge of the statistics while the KCCA algorithms operate completely blindly in this sense. Furthermore, the KCCA-MINVAR algorithms have a noticeable advantage over KCCA-MAXVAR. The reason is that they use the result of the $M-1$ most informative tests in order to explain the remaining test. KCCA-MAXVAR, on the other hand, uses the result of a single test in order to explain the $M-1$ remaining ones, which is clearly a disadvantage in this particular scenario. Finally, note that the adaptive algorithms perform slightly worse compared to their batch equivalents, since they are trained on the same data set but in an online manner. Nevertheless, their execution times are lower compared to the batch algorithms, since they are based on lower-complexity KRLS implementations.

5. CONCLUSIONS

We have proposed KCCA-MAXVAR and KCCA-MINVAR formulations that allow to perform nonlinear CCA with multiple data sets, both in batch and online (adaptive) form. The online algorithm is made possible thanks to recent advances in kernel adaptive filtering.

The first experimental results of these algorithms are very promising. In particular, we have described a new application of KCCA in the context of cognitive radio, in which it can be used to construct a hypothesis test for detection without using any knowledge of the signal statistics.

6. REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[2] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[3] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[4] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 5, pp. 365–377, 2000.

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

[6] J. Vía, I. Santamaría, and J. Pérez, "A learning algorithm for adaptive canonical correlation analysis of several data sets," *Neural Networks*, vol. 20, no. 1, pp. 139–152, Jan. 2007.

[7] O. Friman, J. Cedefamn, P. Lundberg, M. Borga, and H. Knutsson, "Detection of neural activity in functional MRI using canonical correlation analysis," *Magnetic Resonance in Medicine*, vol. 45, no. 2, pp. 323–330, 2001.

[8] D.R. Hardoon, J. Mourao-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fMRI data using kernel canonical correlation," *NeuroImage*, vol. 37, no. 4, pp. 1250–1259, 2007.

[9] N.M. Correa, T. Adali, Yi-Ou Li, and V.D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 39 –50, 2010.

[10] Y.O. Li, T. Adali, W. Wang, and V.D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[11] J. Vía, I. Santamaría, and J. Pérez, "Effective channel order estimation based on combined identification / equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3518–3526, Sept. 2006.

[12] S. Van Vaerenbergh, J. Vía, and I. Santamaría, "Adaptive kernel canonical correlation analysis algorithms for nonparametric identification of Wiener and Hammerstein systems," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 10–22, Apr. 2008.

[13] W. Liu, J.C. Príncipe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*, Wiley, 2010.

[14] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 97–123, 2011.

[15] S. Van Vaerenbergh, M. Lázaro-Gredilla, and I. Santamaría, "Kernel recursive least-squares tracker for time-varying regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1313–1326, Aug. 2012.

[16] S. Haykin et al., "Cognitive radio: brain-empowered wireless communications," *IEEE journal on selected areas in communications*, vol. 23, no. 2, pp. 201–220, 2005.

[17] D. Ramírez, J. Vía, I. Santamaría, and L.L. Scharf, "Detection of spatially correlated gaussian time series," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5006–5015, 2010.

[18] D. Ramírez, G. Vázquez-Vilar, R. López-Valcarce, J. Vía, and I. Santamaría, "Detection of rank-p signals in cognitive radio networks with uncalibrated multiple antennas," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3764–3774, 2011.

[19] S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Number v. 2 in Prentice Hall Signal Processing Series. Prentice Hall, 1993.