Complexity-based discrepancy measures applied to detection of apnea-hypopnea events

R.E. Rolón[a], I.E. Gareis[a,c], L.E. Di Persia[a], R.D. Spies[b], H.L. Rufiner[a,c]

May 2, 2018

[a] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH−UNL/CONICET, Santa Fe, Argentina
[b] Instituto de Matemática Aplicada del Litoral, IMAL, FIQ−UNL/CONICET, Santa Fe, Argentina
[c] Laboratorio de Cibernética, Fac. de Ing., Univ. Nacional de Entre Ríos, Argentina

Abstract

In recent years an increasing interest in the development of discriminative methods based on sparse representations with discrete dictionaries for signal classification has been observed. It is still unclear, however, what is the most appropriate way for introducing discriminative information into the sparse representation problem. It is also unknown which is the best discrepancy measure for classification purposes. In the context of feature selection problems, several complexity-based measures have been proposed. The main objective of this work is to explore a method that uses such measures for constructing discriminative sub-dictionaries for detecting apnea-hypopnea events using pulse oximetry signals. Besides traditional discrepancy measures, we study a simple one called difference of conditional activation frequency (DCAF). We additionally explore the combined effect of over-completeness and redundancy of the dictionary as well as the sparsity level of the representation. Results show that complexity-based measures are capable of adequately pointing out discriminative atoms. Particularly DCAF yields competitive averaged detection accuracy rates of 72.57% at low computational cost. Additionally, ROC curve analyses show averaged diagnostic sensitivity and specificity of 81.88% and 87.32%, respectively. This shows that discriminative sub-dictionary construction methods for sparse representations of pulse oximetry signals constitute a valuable tool for apnea-hypopnea screening.

Keywords: Discriminative information, discrepancy measures, sparse representation, apnea-hypopnea events, pulse oximetry signal.

1 Introduction

Although it is widely used and accepted, the notion of complexity has very often avoided a rigorous formalization. It is therefore not surprising that no universally accepted measure exists yet for quantifying such a concept. In particular, within information theory, the complexity of any element of a code, or of any feature of a signal representation in the context of signal processing, is known to be strongly related to the information it carries or, more precisely, to the value of its entropy. It is important to point out however that, in the context of signal classification, the more informative features (in terms of classification) are not necessarily the ones with larger entropy. Hence more "ad-hoc" measures are needed. In fact, any appropriate complexity measure corresponding to a given feature should be instead, strongly related to the amount of information about class membership provided by such a feature. One could then think of using as measure of complexity the conditional entropy of the class given the feature. However, features providing the most discriminative information regarding a class are almost always those with lower conditional entropy values, and hence the best features for classification purposes will be the least complex ones.

Information theory was originally based on the engineering of noisy communication channels, and it is closely associated to a large number of disciplines such as signal processing, artificial intelligence, complex systems and pattern recognition, to name only a few. We are particularly interested in the latter. Pattern recognition is a discipline which is mainly oriented to the generation of algorithms or methods that can decide an action based upon certain recognized similarities (patterns) in the input data. Within signal classification, which is perhaps one of the most important subfields of pattern recognition, several discrepancy measures have been used in problems coming from a wide variety of areas such as machine learning [1], image and speech processing [2], neural networks [3] and biomedical signal processing [4, 5], among others. Among them the most commonly used is probably the Kullback-Leibler (KL) divergence [6, 7]. This divergence, also known as relative entropy, was used as a discriminative measure for selecting, from a large collection of orthonormal bases, the one attaining maximum information [1]. A more recent approach was introduced by Gupta *et.al.* [8] who used this divergence as a discrepancy measure in the traditional k-nearest neighbor (k-NN) algorithm, yielding competitive classification performances in the context of raw electroencephalographic signal classification. Although it provides certain computational and theoretical advantages, the lack of symmetry of the KL divergence has motivated the development of several symmetric versions such as the so called J-divergence [9] and the well known and widely used Jensen-Shannon divergence [10].

Sparse representation of signals constitute a useful technique which has drawn wide interest in recent years due to its success in many applications such as signal and image processing [11]. This technique allows the analysis of the signals by means of only a few well-defined basic waveforms. Due to its advantages, such as robustness to noise and dimension reduction, among others, sparse representation has acquired a large popularity in the area of biomedical signal processing. For example, this technique has been successfully applied to several problems including the estimation of the human respiratory rate [12] and electrocardiographic signal processing, both for signal enhancement and QRS complex detection, for improving heart disease analysis and diagnosis [13]. It is timely to point out however that, up to our knowledge, no applications of discrepancy measures to sparse representation for signals classification are known yet.

All reconstructive methods, such as principal components analysis (PCA), independent components analysis (ICA) and the previously mentioned sparse representations [14], produce particular types of signal representations minimizing a given cost functional which usually involves both fidelity and regularization terms. These methods have been successfully applied in a wide variety of problems such as signal denoising, missing data and outliers, among many others. On the other hand, discriminative methods such as linear discriminant analysis (LDA) are oriented to find optimal decision boundaries to be used for classification tasks. It is well known that for signal classification, which is our main interest in this work, discriminative methods generally outperform reconstructive methods. It is mainly for this reason that several authors have recently developed supervised approaches based on sparse representation which are simultaneously reconstructive and discriminative [15, 16].

The obstructive sleep apnea-hypopnea (OSAH) syndrome [17] is one of the most common sleep disorders and more often that not it remains undiagnosed and therefore not treated. This syndrome is caused by repeated events of partial or total blockage of the upper airway during sleeping, which correspond to events of hypopnea and apnea, respectively. To evaluate the severity degree of the OSAH syndrome, medical physicians have created the so called apnea-hypopnea index (AHI), which is defined as the average number of apnea-hypopnea events per hour of sleep. In terms of this index OSAH is classified as normal, mild, moderate or severe depending on whether such an index falls in the interval $[0, 5)$, $[5, 15)$, $[15, 30)$, or $[30, \infty)$, respectively. The gold standard test for OSAH diagnosis is a study called polysomnography (PSG). However, PSG is both costly and lengthy and the accessibility to this type of study is limited. Additionally, PSG studies require of information coming from a variety of physiological signals such as electroencephalography (EEG), airflow, pulse oximetry (SaO_2), etcetera. It is known however that cessation of breathing associated with apnea-hypopnea events are always accompanied by a drop in the oxygen saturation level in the SaO_2 signal record, although quite often such a drop is very small and almost impossible to detect by a human observer.

The main objective of this work is precisely to develop a technique based on sparse representations and the use of appropriate discriminative information that be able to accurately and efficiently detect apnea-hypopnea events by using only the SaO_2 signal. Several ways exists for combining discriminative information and sparse representations within the context of signal classification. We shall follow one consisting of using the discriminative information for detecting those atoms having the most frequent activations in order to provide them as input for a classifier. This approach was initially introduced in [4] where two methods using the absolute value of the activation differences of the atoms as a measure of the discriminative information for the detection of OSAH were presented. In this work a rigorous formalization of such a measure is introduced and compared with several other discrepancy measures for classifying apnea-hypopnea events. Also, the combined effect of using different sizes of non-redundant dictionaries and different sparsity degree is explored in detail. Results show clearly that the proposed measure is capable of adequately pointing out discriminative atoms in a full dictionary, yielding competitive accuracy rates in the detection of individual apnea-hypopnea events. Additionally, this new approach is computationally very cheap. In fact, it has proved to be at least twice faster than those associated to all other discrepancy measures.

The rest of this article is organized as follows: in Section 2 the obstructive sleep apnea-hypopnea syndrome is explained. Sparse representation of signals is introduced in Section 3. In Section 4 several discriminative information measures are presented. Section 6 contains a detailed description about the performed experiments. Results and discussions are introduced in Section 7 while conclusions are presented in Section 8.

2 Sleep apnea-hypopnea

Apnea-hypopnea events occur as a consequence of a functional-anatomic disturbance of the upper airway producing its partial or total blockage. At the end of an apnea-hypopnea event, a pronounced desaturation of the blood hemoglobin commonly occurs. These desaturations generate characteristic patterns in the pulse oximetry record known as intermittent hypoxemias. The hypoxemia-reoxygenation cycles promote oxidative stress, angiogenesis and tumor growth, favor the sympathetic activation with increment of blood pressure and systemic and vascular inflammation with endothelial dysfunction which contributes to multi-organic chronic morbility, metabolic abnormalities and cognitive impairment [18]. Additionally, strong correlations between neoplastic diseases and the OSAH syndrome have been described in [19]. Also, a recent study among male mice suggests that OSAH's intermittent hypoxia can be associated to fertility reduction [20]. Currently this pathology affects more than 4% of the human population around the world [21]. Additionally, it was found that aging, male gender, snoring and obesity are all risk factors for OSAH syndrome [22].

Although very limited in many countries, overnight polysomnography (PSG) is currently the gold standard tool for diagnosing OSAH syndrome. As previously mentioned, a full PSG consists of the simultaneous measurement of several physiological signals such as EEG, electrocardiography (ECG), respiratory effort, airflow, SaO_2 and electrical activity produced by skeletal muscles (EMG), etc. Mainly due to its ease of acquisition, we are particularly interested in the SaO_2 signal. Figure 1 shows a typical temporal plot of just a few physiological signals coming from a full PSG. This figure also depicts a portion of an original raw airflow signal as well as the corresponding portion of the SaO_2 signal. The corresponding labels of apnea-hypopnea events (dashed lines) are also shown. Finally, at the bottom of this figure, the electrical activity of the heart as well as the sleep stages are shown. In a typical PSG study, after a normal period of sleep the recorded signals are provided to medical experts who analyze the whole record and mark the apnea-hypopnea events and sleep stages, needed for the posterior evaluation the AHI index. Due to its complexity and cost, a few alternatives to PSG have been adopted. One of the most popular ones is the so called home respiratory polygraphy (HRP) [23] which requires no neurophysiological signals. Although studies have shown that there exists a high correlation between AHI values generated by HRP and PSG studies [24], HRP still needs of several physiological signals, whose acquisition strongly affects the normal sleeping of the person. It is therefore highly desirable to develop a reliable OSAH screening system which makes use of as few as possible physiological signals. In this regard, pulse oximetry, being a cheap and non-invasive technique, has become a suitable alternative for screening purposes [25].

In this work we shall develop a method for the detection of apnea-hypopnea events that uses only the SaO_2 signals. Our approach leads to a binary classification problem whose main purpose is the detection of the presence (or not) of events of apnea and hypopnea. It is timely to point out that although our method does take into consideration an appropriate fidelity term, we are by no means interested in achieving accurate signal representation.

3 Sparse representations

As previously mentioned, one of the most popular reconstructive methods is based on sparse representations of the signals involved. Sparsity can be enforced by including upper bounds for the number of non-zero coefficients in the representation of the given signals in terms of atoms in a dictionary.

Formally, the problem of sparse representations of signals can be separated into two sub-problems, the so-called sparse coding problem and the dictionary learning problem. We shall now proceed to describe in detail each one of these sub-problems. To be more precise, let $\mathbf{x} \in \mathbb{R}^N$ be a discrete signal and let $\Phi \in \mathbb{R}^{N \times M}$ (generally with $M \geq N$) be a dictionary whose columns $\boldsymbol{\phi}_j \in \mathbb{R}^N$ are atoms that we want to use for obtaining a representation of $\mathbf{x}$ of the form $\mathbf{x} = \Phi\mathbf{a}$. Here, and in the sequel, we shall refer to the vector $\mathbf{a} = [a_1\ a_2\ \cdots\ a_M]^T \in \mathbb{R}^M$ as a "representation" of $\mathbf{x}$. Sparsity consists essentially of obtaining a representation with as few non-zero elements as possible. A way of obtaining such a representation

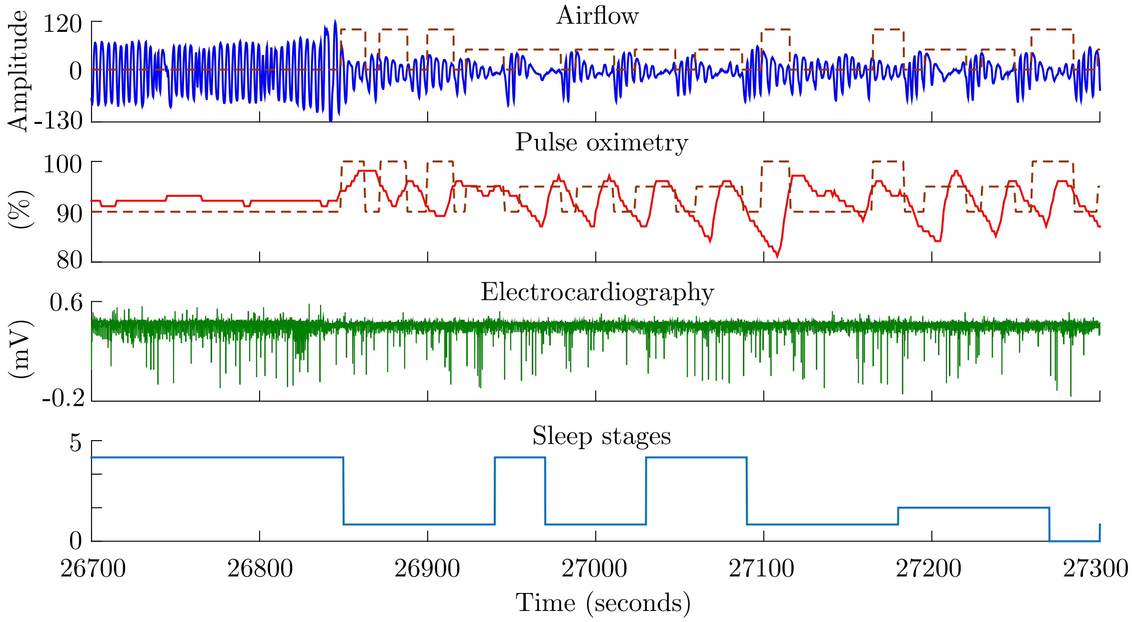

Figure 1: A portion of a few number of physiological signals coming from a full PSG. Dashed lines (brown) are apnea-hypopnea labels introduced by the medical expert.

consists of solving the following problem:

$$(P_0): \quad \min_{\mathbf{a}} ||\mathbf{a}||_0 \text{ subject to } \mathbf{x} = \Phi\mathbf{a},$$

where $||\mathbf{a}||_0$ denotes the l_0 pseudo-norm, defined as the number of non-zero elements of $\mathbf{a}$.

Several questions regarding problem (P_0) immediately arise. Among them: *i*) does there exist an exact representation $\mathbf{x} = \Phi\mathbf{a}$?, *ii*) if an exact representation exists, is it unique?, *iii*) in the case of non-uniqueness, how do we find the "sparsest" representation?, *iv*) how difficult is it, from the computational point of view, to solve problem (P_0)?. Although it is not an objective of this article to get into details about the answers to these questions, it turns out that imposing exact representation is most often a too restrictive and therefore inappropriate constrain and, on the other hand, solving (P_0) is generally an NP hard problem yielding this approach highly unsuitable for most applications. For more details we refer the reader to [26, §1.8].

In order to overcome some of the difficulties which entail solving problem (P_0), several relaxed versions of it have been considered. One of them consists of allowing a small representation error while imposing an upper bound on the l_0 pseudo-norm of the representation:

$$(P_0^q): \quad \min_{\mathbf{a}} ||\mathbf{x} - \Phi\mathbf{a}||_2 \text{ subject to } ||\mathbf{a}||_0 \leq q,$$

where q is a prescribed integer parameter. This formulation takes into account the existence of possible additive noise terms; in other words it assumes that $\mathbf{x} = \Phi\mathbf{a} + \mathbf{e}$ where $\mathbf{e} \in \mathbb{R}^N$ is a small energy noise term. Thus, this approach is particularly suitable in most real applications (such as biomedical signal processing) where measured signals are always contaminated by noise. Several greedy strategies have been proposed for solving problem (P_0^q) [27, 28]. Among them, orthogonal matching pursuit (OMP) [28] is perhaps the most commonly used strategy. This greedy algorithm guarantees convergence to the projection of $\mathbf{x}$ into the span of the dictionary atoms, in no more than q iterations. Figure 2 shows an example of the values of a particular coefficient a_{j^*} associated to the atom ϕ_{j^*} obtained by applying the OMP algorithm for a large number (almost half a million) of segments of SaO_2 signals and its corresponding activation histogram.

Although pre-constructed dictionaries, such as the well known wavelet packets [29], typically lead to fast sparse coding, they are almost always restricted to certain classes of signals. Is is mainly for this reason that new approaches introducing data-driven dictionary learning techniques emerged. A dictionary learning (DL) problem consists of simultaneously finding a dictionary Φ and representations of n signals $\mathbf{x}_i$, $1 \leq i \leq n$, (in terms of atoms of such a dictionary) complying with a sparsity constraint for each one of the n signals, while minimizing the total representation error. The (DL) problem associated to the data: q, M, $N \in \mathbb{N}$, $M \geq N$ and n signals in $\mathbb{R}^N$, $\mathbf{x}_1, \cdots, \mathbf{x}_n$, can be formally written as:

$$(DL): \quad \min_{\substack{\Phi \in \mathbb{R}^{N \times M} \\ \mathbf{a}_i \in \mathbb{R}^M,\ ||\mathbf{a}_i||_0 \leq q,\ 1 \leq i \leq n.}} \quad \sum_{i=1}^{n} ||\mathbf{x}_i - \Phi \mathbf{a}_i||_2$$

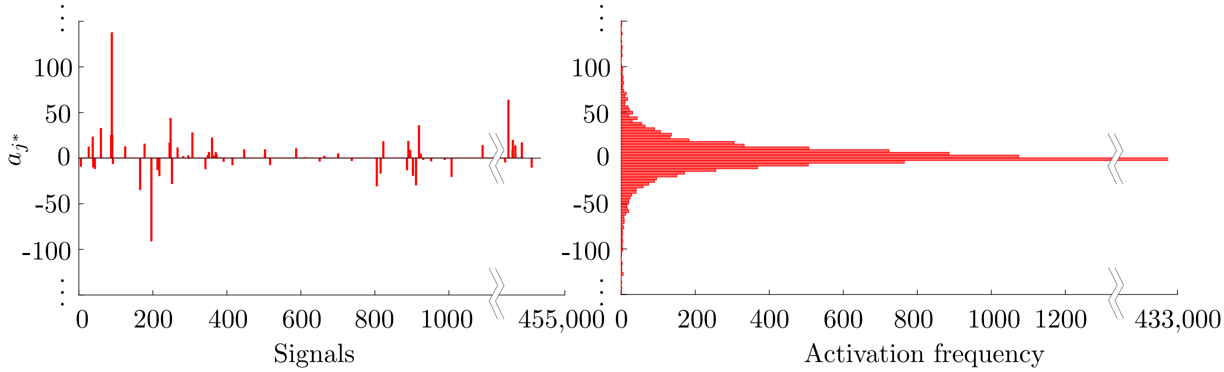

Figure 2: The values of the activations of a particular atom for each signal (left) and the corresponding histogram of activations (right).

The first data-based dictionary learning algorithms were originally developed almost three decades ago [30, 31, 32]. Some of them have their roots in probabilistic frameworks by considering the observed data as realizations of certain random variables [30, 31]. In [31] for example, the authors developed an algorithm for finding a redundant dictionary that maximizes the likelihood function of the probability distribution of the data. In that work, an analytic expression for the likelihood function was derived by approximating the posterior distribution by Gaussian functions. An iterative approach for dictionary learning, known as the "method for optimal directions" (MOD), was presented in [32]. The sparse coding stage of this method makes use of the OMP algorithm followed by a simple dictionary updating rule. A new iterative algorithm was recently proposed by Aharon *et.al.* in [14]. This new approach, called "K singular value decompositions" (KSVD), consists mainly of two stages: a sparse coding stage and a dictionary learning stage. The OMP algorithm is used for the sparse coding stage, which is followed by a dictionary updating step where the atoms are updated one at a time and the representation coefficients are allowed to change in order to minimize the total representation error.

4 Discriminative sub-dictionary construction

Although data-driven dictionary learning algorithms produce sparse representations of signals which are robust against noise and missing data, such representations turn out to be unsuitable if the final objective is signal classification. This is mainly so because those algorithms do not take into account any a-priori or available information concerning class membership. In order to overcome this difficulty, some strategies which incorporate appropriate class information have been proposed [4, 33, 16]. In [33], for instance, the authors developed a discriminative dictionary learning method by efficiently integrating a single predictive linear classifier into the cost function of the KSVD algorithm. A method incorporating a discriminative term into the cost function of the standard KSVD algorithm was presented in [16]. This method finds an optimal dictionary which is simultaneously representative and discriminative for face recognition tasks. In this work, we make use of a simple approach for detecting discriminative atoms from a previously learned dictionary and using them to build a new sub-dictionary. This approach, which was originally presented in [4], consists of solving two problems, namely: *i*) the above mentioned full (DL) problem and *ii*) a discriminative sub-dictionary (DSD) construction problem. We shall now proceed to describe problem *ii*). One way to obtain discriminative sub-dictionaries consists of maximizing an appropriate discriminative value functional $G(\cdot)$. Given a data matrix $\mathbf{X} \in \mathbb{R}^{N \times n}$, a class label vector $\mathbf{c} \in \mathcal{C}^n$ (where $\mathcal{C}$ is the set of all classes; in the binary case $\mathcal{C} = \{c_1, c_2\}$), a dictionary $\Phi \in \mathbb{R}^{N \times M}$ and $p \in \mathbb{N}$ (with $p < M$), the most discriminative sub-dictionary $\hat{\Phi}^{\mathbf{d}} \in \mathbb{R}^{N \times p}$, according to an appropriate prescribed discriminative value functional $G_{\mathbf{X},\mathbf{c},\Phi} : \mathbb{R}^{N \times p} \to \mathbb{R}_0^+$ is defined as:

$$(DSD): \quad \hat{\Phi}^{\mathbf{d}} = \operatorname*{argmax}_{\substack{\mathbf{d} \doteq [i_1\ i_2\ \cdots\ i_p] \\ i_j \in \{1,2,\cdots,M\} \\ i_j \neq i_k \forall j \neq k,}} \quad G_{\mathbf{X},\mathbf{c},\Phi}\left(\Phi^{\mathbf{d}}\right)$$

where for $\mathbf{d} \doteq [i_1\ i_2\ \cdots\ i_p]$, $\Phi^{\mathbf{d}}$ denotes the $N \times p$ matrix whose j^{th}-column is the i_j^{th}-column of Φ.

The function G, which must be provided, quantifies the discriminative power of each sub-dictionary $\Phi^{\mathbf{d}}$. Thus, large values of G correspond to highly discriminative sub-dictionaries while small values of G are associated to sub-dictionaries with low discriminability.

Several questions concerning problem (DSD) clearly emerge. Among them: i) how do we find an appropriate discriminative value function G?, ii) given the functional G, does problem (DSD) have a solution?, iii) if it does, is it unique?, iv) in the case of non-uniqueness, how do we decide which sub-dictionary, among the optimizers, is the best for our classification purposes?, v) how difficult is it, in terms of computational cost, to solve problem (DSD)?. Although this problem has not been extensively studied, is it known that solving (DSD) is computationally very challenging for $p > 1$, mainly due to the combinatorial explosion problem. A way to overcome the computational complexities entailed by problem (DSD) consists of defining an appropriate discriminative value functional G for $p = 1$. In that way G is independently evaluated at each one of the atoms (columns) of Φ and the discriminative sub-dictionary $\Phi^{\mathbf{d}} \in \mathbb{R}^{N \times p^*}$ is constructed by stacking side-by-side the first p^* ranked columns of Φ with largest G values. This simplification is based on the assumption that each atom in the dictionary is used to model specific characteristics that are not completely modeled by the other atoms. Thus, the discriminative information provided by a particular atom will be different from the information contributed by other atoms.

5 Discriminative value functions for atom selection

Several ways for appropriately constructing discriminative value functions G exists. In this section we present two different approaches to define such a function. Namely $i)$ using traditional discrepancy measures and $ii)$ using a new discriminative measure to which we shall refer as the "difference of conditional activation frequency" (DCAF). We shall previously need to introduce an appropriate setting and terminology regarding probability density functions (PDFs) in the context of sparse representations for signal classification.

Here, and in the sequel, we shall consider the vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ as realizations of a particular random vector $\mathcal{X}$. Any sparse representation of those vectors will result in the PDFs of each coefficient a_j (associated to the atom ϕ_j) showing a very concentrated peak at zero with heavy tails (as depicted in Figure 2). In the context of binary signal classification it is reasonable to think that if a given atom ϕ_{j^*} is highly discriminative, then the conditional PDFs $\pi(a_{j^*}|c_1)$ and $\pi(a_{j^*}|c_2)$ will be significantly different. Thus, if a dictionary Φ is poorly discriminative, then one should expect $\pi(a_j|c_1) \approx \pi(a_j|c_2)$ for all j.

Although the elements a_j of the representation vector $\mathbf{a}$ are in general real numbers, for practical reasons it is appropriate to discretize them. That can be done in the usual way by partitioning the real line $\mathbb{R}$ into intervals $I_k \doteq \left(\left(k - \frac{1}{2}\right)\Delta, \left(k + \frac{1}{2}\right)\Delta\right]$, $k \in \mathbb{Z}$, of length Δ and the associated discretized random variable $\mathcal{K}_j \doteq \sum_{k \in \mathbb{Z}} k \, \chi_{I_k}(a_j)$. The corresponding probability mass function (PMF) is $p_{\mathcal{K}_j}(k) = P(a_j \in I_k) = \int_{I_k} \pi(a_j) \, da_j$, $k \in \mathbb{Z}$. Figure 3 shows the estimated PMF and the corresponding conditional PMFs (given each one of the two classes), both for a non-discriminative and a discriminative atom using SaO_2 signals.

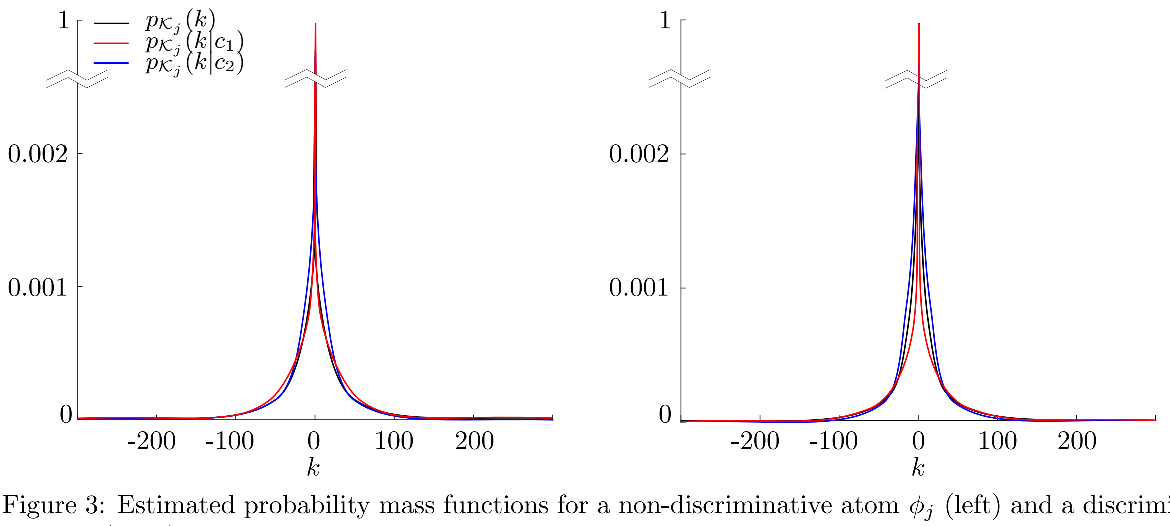

Figure 3: Estimated probability mass functions for a non-discriminative atom ϕ_j (left) and a discriminative one (right).

We shall now proceed to define how we compute the discriminative value function G. Given the data matrix $\mathbf{X} \in \mathbb{R}^{N\times n}$, the corresponding class label vector $\mathbf{c} \in \mathcal{C}^n$ and a full dictionary $\Phi \in \mathbb{R}^{N\times M}$, the first step consists of obtaining the sparse matrix $\mathbf{A} \doteq [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n] \in \mathbb{R}^{M\times n}$ by applying the OMP algorithm. The j^{th}-row of this sparse matrix is then used for estimating the conditional PMFs $p_{\mathcal{K}_j}(\cdot|c_1)$ and $p_{\mathcal{K}_j}(\cdot|c_2)$. Finally, the value of G at the atom ϕ_j is computed as the discrepancy (as quantified by an appropriate discrepancy measure) between these two PMFs. In what follows, we introduce the discrepancy measures that we shall use in this work.

5.1 Traditional discrepancy measures

A great diversity of measures whose purpose is performing comparisons between probability distributions exists [34]. In this work the best known and more commonly used ones are compared in terms of their performance for selecting the most discriminative atoms in a dictionary. The KL, J and JS divergence measures were utilized, along with the Fisher score (F).

The KL divergence [7] is probably the most widely used information "distance" measure from a theoretical framework and it was successfully applied in numerous problems for signal classification [1, 35, 36]. To compare the two conditional PMFs associated with the activation of the j^{th}-atom the KL distance was used as follows:

$$\text{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq \sum_{k\in\mathbb{Z}} p_{\mathcal{K}_j}(k|c_1) \log\left(\frac{p_{\mathcal{K}_j}(k|c_1)}{p_{\mathcal{K}_j}(k|c_2)}\right), \tag{1}$$

assuming that $0 \log(0) \doteq 0$.

Despite the computational and theoretical properties provided by KL distance, what usually becomes a trouble in many problems of signal classification is its lack of symmetry. It can be easily seen that altering the order of the arguments in (1) can change the output value. To solve this issue a symmetric version of the KL distance can be used such as the J-divergence [9], which, even though was not initially created as a symmetric version of the KL distance, is the sum of the two possible KL distances between probability distributions. In this article the J-divergence is defined as follows:

$$\text{J}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq \text{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) + \text{KL}\left(p_{\mathcal{K}_j}(\cdot|c_2), p_{\mathcal{K}_j}(\cdot|c_1)\right). \tag{2}$$

Another symmetric smoothed version of the KL distance is the JS divergence [10]. For the problem of comparing the two conditional probabilities associated to each class it is defined as:

$$\text{JS}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq w_1\text{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), q_{\mathcal{K}_j}(\cdot)\right) + w_2\text{KL}\left(p_{\mathcal{K}_j}(\cdot|c_2), q_{\mathcal{K}_j}(\cdot)\right), \tag{3}$$

where $q_{\mathcal{K}_j}(\cdot) = w_1 p_{\mathcal{K}_j}(\cdot|c_1) + w_2 p_{\mathcal{K}_j}(\cdot|c_2)$ and w_1 and w_2 are the weights associated to each of the conditional PMFs, with $w_1, w_2 \geq 0$ and $w_1 + w_2 = 1$. An interesting feature of the JS-distance is the fact that different values of weights (w_1 and w_2) can be assigned to the probability distributions according to their importance. In this work $w_1 = P(c_1)$ and $w_2 = P(c_2)$ i.e. the weights are associated with the a-priori probabilities of the classes. Note that computing the JS-distance as defined here is the same as computing the mutual information between the class and the activations, i.e. $\text{JS}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) = \text{MI}(\mathcal{K}_j, \mathcal{C})$.

Within signal classification problems, F is a measure which has been extensively used. Unlike the other measures presented here, that require estimations of the conditional PMFs, F uses just two parameters of the distributions (the means and standard deviations). This makes this measure much less expensive computationally speaking, but implicitly assumes certain characteristics of the distribution under study (i.e. second order characteristics). In the case of univariate binary problem at hand the F can be defined as:

$$\text{F}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)) \doteq \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{4}$$

where μ_ℓ and σ_ℓ^2 are the mean and standard deviation of $p_{\mathcal{K}_j}(\cdot|c_\ell)$ [37].

Although the above mentioned discrepancy measures provide, in a certain sense, "measures" of distance between two probability distribution functions, most of them (such as the KL divergence and those symmetric variants) are not strictly a metric. For instance, the KL divergence is a non-symmetric discrepancy measure where the triangular inequality is not satisfied. Nevertheless, $\text{KL}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2))$ is a non-negative measure, i.e. $\text{KL}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)) \geq 0$ and $\text{KL}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)) = 0$ if and only if $p_{\mathcal{K}_j}(\cdot|c_1) = p_{\mathcal{K}_j}(\cdot|c_2)$.

5.2 Difference of conditional activation frequency

In a previous work a method called most discriminative column selection (MDCS) for the construction of a discriminative sub-dictionary was originally presented [4]. The sparse representations of the signals in terms of sub-dictionaries constructed using MDCS provided good performance in the detection of apnea-hypopnea events. In the mentioned work, the most discriminative atoms were identified by comparing the difference of conditional activation frequency (DCAF).

The candidates to be considered as "most discriminative" according to [4] are those atoms with higher absolute difference between conditional activation probabilities given the class. That is, an atom is considered as highly discriminative if it is active, in proportion, more times for one of the classes. The use of this approach as a measure of discriminative power follows from the idea that one of the most expressive parameters regarding the importance of a given atom is its activation probability. Moreover, if certain atoms are active mostly for a given class, then it is assumed they represent features of importance in the description of that particular class.

Following this idea, DCAF is defined as:

$$\text{DCAF}(\eta_1^j, \eta_2^j) \doteq |\eta_1^j - \eta_2^j|, \tag{5}$$

where:

$$\eta_\ell^j \doteq \frac{\text{number of activations of the } j^{\text{th}}\text{-atom for } c_\ell}{\text{number of } c_\ell \text{ samples}}. \tag{6}$$

The measure defined in (5) is symmetric, its value is always ≥ 0, and is inexpensive in terms of computing[1].

It can easily be seen that the definition of η_ℓ^j in (6) is equal to the maximum likelihood estimation of the conditional probability of activation, i.e.:

$$p_{\mathcal{K}_j}(k \neq 0 | c_\ell) \approx \eta_\ell^j. \tag{7}$$

Replacing this expression in (5) we can write,

$$\begin{aligned} \text{DCAF}(\eta_1^j, \eta_2^j) &\approx |p_{\mathcal{K}_j}(k \neq 0|c_1) - p_{\mathcal{K}_j}(k \neq 0|c_2)|, \\ &\approx \big|\big(1 - p_{\mathcal{K}_j}(k = 0|c_1)\big) - \big(1 - p_{\mathcal{K}_j}(k = 0|c_2)\big)\big|, \\ &\approx |p_{\mathcal{K}_j}(k = 0|c_2) - p_{\mathcal{K}_j}(k = 0|c_1)|, \end{aligned} \tag{8}$$

finally expressing the DCAF in terms of the complementary conditional probabilities that the atoms will not be activated. With the exception of the F, all the measures presented in Section 5.1 can be expressed as summations, where only one of the terms is computed using the probabilities that $k = 0$. However, due to the high sparsity of the representations the terms associated with $k = 0$ are particularly important. This fact allows us to expect some correlation between the results obtained with the different discrepancy measures and the DCAF.

Figure 4 shows a representation of the conditional PMFs associated to the activations of two different atoms (left side) as well as an illustration of such functions where the peaks centered at zero ($k = 0$) were discarded (middle). It is important to note that, when excluding the zero-centered peak from the graphic, a significant reduction in the magnitude of the y-axis scale is produced which highlights the importance of the activation probability of sparse representations. However, the discrepancy between the distributions is not only due to the atoms activation probability, since slight differences between the probability values for all $k \neq 0$ exist (zoom-in region). Additionally, the absolute values of these differences are represented by the gray regions. It is also important to point out that, these area values shown in gray ($\sum_{k\neq 0} |p_{\mathcal{K}_\ell}(k|c_1) - p_{\mathcal{K}_\ell}(k|c_2)|$) are not necessarily equal to those corresponding to the DCAF values. Nevertheless, for symmetric PMFs with high kurtosis and heavy tails (such is the case of the PMFs used in this work), the conditional and a-priori distributions are similar and therefore both area values are close to each other.

6 Experimental setup

This section presents the proposed system and its configuration settings, aimed at detecting patients suspected of suffering from moderate-severe OSAH syndrome. It also describes the database used for training and testing the method along with the measures selected for assessing its performance.

The main objective of our research is to explore the effect of using discrepancy measures to rank the atoms according to their discriminative power. Also, the experiments are designed to determine the

[1] If the classes are balanced the DCAF can be replaced just by simply counting, without the necessity of dividing with the number of samples.

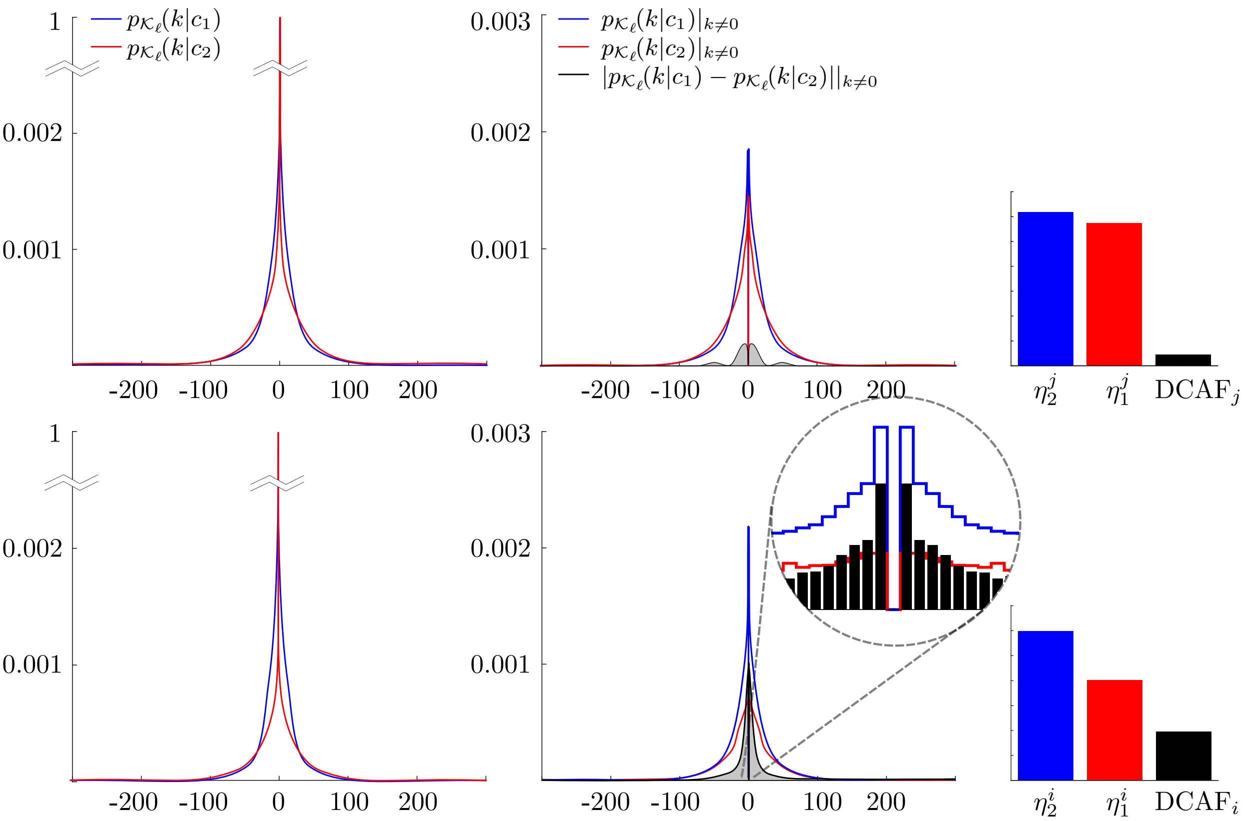

Figure 4: A representation of the conditional PMFs corresponding to the activations of two different atoms (left side), the same functions excluding the peaks centered at zero ($k = 0$) and the absolute value of their differences (middle) and a graphical interpretation of the DCAF (right side). The top row corresponds to a non-discriminative atom (ϕ_j) while the bottom row corresponds to a discriminative one (ϕ_i).

effect of using dictionaries with different degree of over-completeness (redundant dictionaries) for the detection of apnea-hypopnea events. Additionally, the performance of the system for different sizes of sub-dictionaries and sparsity degrees is analyzed.

Figure 5 shows a simplified block diagram of the presented system. It can be observed that our system comprises a training phase (above) and a testing phase (below). To clarify the system's description, we divided it into three different stages, namely Stage I, Stage II and Stage III. It can be seen that stages I and II are included into training and testing phases while Stage III is only used during testing. Stage I is composed by a pre-processing block whose inputs are the raw SaO_2 signals and its outputs are filtered segments of such signals, as described in Section 6.1. At the training phase, Stage II receives segmented signals and finds an optimal discriminative sub-dictionary. During the testing phase, Stage II obtains a sparse matrix in terms of the previously found sub-dictionary. These processes are thoroughly described in Section 6.2. Finally, the obtained sparse codes are used as input of Stage III. This stage detects apnea-hypopnea events and estimates the AHI value, as described in Section 6.3.

6.1 Database and signal's pre-processing

The sleep heart health study (SHHS) dataset [38, 39] was originally designed to study correlations between sleep-disordered breathing and cardiovascular diseases. This dataset includes a large number of PSG studies, each of them containing several physiological signals such as EEG, ECG, nasal airflow, SaO_2, among others. Medical expert annotations of sleep stages, arousals and apnea-hypopnea events are also provided. In this work, only the SaO_2 signal (sampled at 1Hz) and its corresponding apnea-hypopnea labels are considered for performing the experiments. In this article, the first online version of such a database (SHHS-2) is used. This version of the database contains a total of 995 freely available PSG

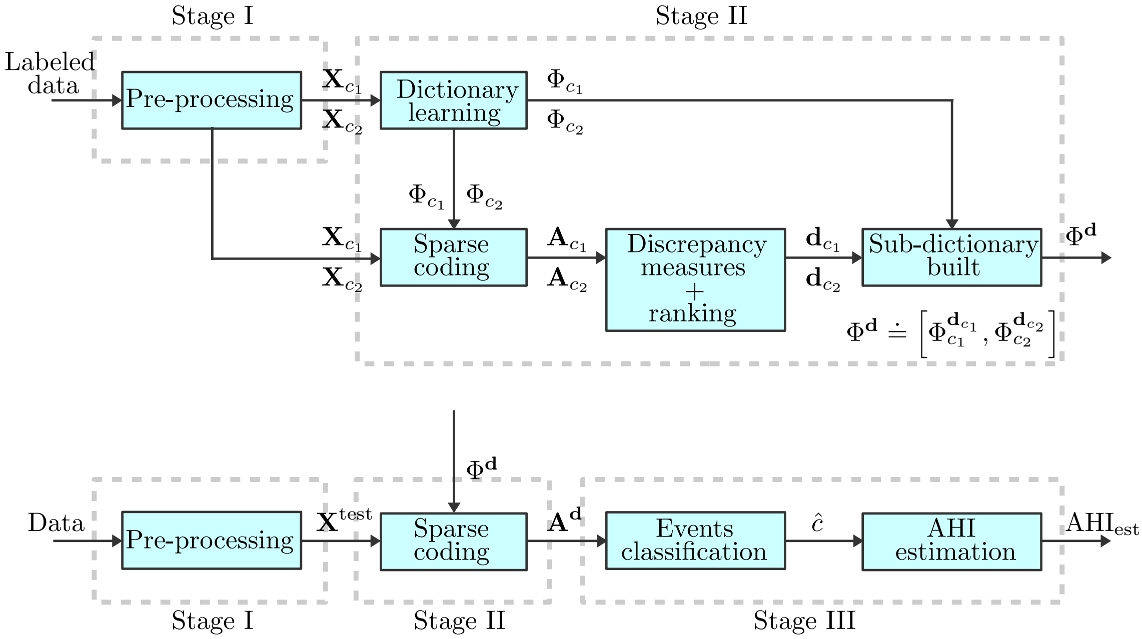

Figure 5: Block diagram of the proposed system during training (top) and testing (bottom).

studies[2].

The SaO_2 signals are mainly degraded by patient movements, baseline wander, disconnections and the limited resolution of pulse oximeters, among others factors. When a disconnection occurs, the recording during the time interval where the sensor signal is blocked is lost. In order to overcome this inconvenient, the values of blood oxygen saturation during such an interval are linearly interpolated. To denoise the signals, a wavelet processing technique [40] is used. The denoising process is performed by zeroing the approximation coefficients at level 8, as well as the coefficients of the first three detail levels of the discrete dyadic wavelet transform with mother wavelet Daubechies 2. The signals are then synthesized using the modified wavelet coefficients by inverse discrete dyadic wavelet transform. The application of this wavelet decomposition technique has the effect of a band-pass filter where the baseline wander and both the low frequency noise and the high frequency noise, as well as the quantization noise are eliminated. Figure 6 shows a small fragment of the original raw SaO_2 signal (top) and its wavelet-filtered version (bottom). Labels of apnea-hypopnea events (dashed lines) introduced by the medical experts are also added. These labels were generated by medical experts using the airflow information and thus are not aligned to the desaturations, i.e. there is a variable delay between the start time of an event and the corresponding desaturation.

The application of the sparse representation technique requires an appropriate segmentation of the signals. Segments of length $N = 128$ (corresponding to 128 seconds of the signal recording) with a 75% overlapping between two consecutive segments are taken. It is appropriate to point out that although several overlapping percentages were tested, the best system performances were yielded by a 75% overlapping. This redundancy prevents apnea-hypopnea events from being undetected. In this segmentation process, the time intervals where a disconnection occurs are discarded. The segments of pulse oximetry signals are then simultaneously arranged as column vectors $\mathbf{x}_i \in \mathbb{R}^N$ and labeled with ones (c_1) and minus ones (c_2), where a one corresponds to apnea-hypopnea events, and a minus one to the lack of it. Finally a signal matrix $\mathbf{X}$ is built by stacking side-by-side the column vectors $\mathbf{x}_i$, i.e. the signal matrix is defined as $\mathbf{X} \doteq [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$.

As mentioned above, the entire dataset used in this work contains 995 complete studies, 41 of which were not taken into account for performing the experiments since the size of the signal vectors differs from the corresponding vector of class labels. Among the remaining 954 studies, a subset of 667 (70%) studies were randomly selected and fixed for learning the dictionary and training the classifier. The remaining 287 (30%) studies were left out for the final test. The SaO_2 signals were filtered using wavelet filters and segmented as explained previously into column vectors of size 128. After performing the filtering and segmentation process, a signal matrix $\mathbf{X}^{\text{train}}$ of size 128×455515 is assembled by joining two previously constructed signal matrices, one for each class, $\mathbf{X}^{\text{train}} \doteq [\mathbf{X}^{\text{train}}_{c_1}\ \mathbf{X}^{\text{train}}_{c_2}]$, which contain 183163 and 272352

[2]https://physionet.org/physiobank/

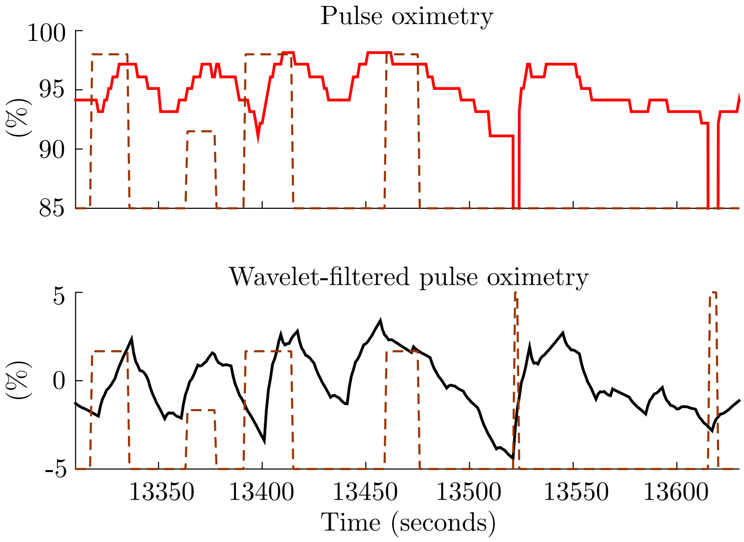

Figure 6: A small fragment of a pulse oximetry signal (top) and its wavelet-filtered version (bottom). Dashed lines represent labels of apnea-hypopnea events established by the medical expert.

segments, respectively. On the other hand, for each study included into the testing dataset, a testing matrix $\mathbf{X}^{\text{test}}$ is built.

6.2 Sparse coding and sub-dictionary construction

In our experiments, the learning of the dictionaries is performed by using the traditional KSVD method [14]. Optimized MATLAB codes for dictionary learning using KSVD as well as for sparse coding using the OMP algorithm are freely available for academic and personal use at the Ron Rubinstein's personal web page[3]. At the beginning, the atoms assigned to conform the initial dictionary are randomly selected from the input signal matrix for training without tacking into account any information about the classes. If the signal's space dimension is fixed, which should be the effect of constructing dictionaries with different over-completeness degree?. To answer this question, three types of dictionaries denoted by $\Phi 1$ of size 128×128, $\Phi 2$ of size 128×256 and $\Phi 4$ of size 128×512, corresponding to redundancy factors of 1, 2 and 4, respectively, were built. First the dictionary $\Phi 1$ was constructed by joining two sub-complete dictionaries of sizes 128×64 denoted by $\Phi 1_{c_1}$ and $\Phi 1_{c_2}$ learned using a large number of training segments (a total of 100,000 segments for each of the classes) belonging to the classes c_1 and c_2, respectively. Following the same idea, redundant dictionaries denoted by $\Phi 2$ (256 atoms) and $\Phi 4$ (512 atoms) were appropriately built. At the dictionary learning stage the number of non-zero elements was selected and fixed as a percentage value of 12.5 of the atoms conforming the dictionary. Also a total of 30 iterations of the KSVD algorithm were performed.

Once the dictionary has already been trained, the sparse representation vectors $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ corresponding to the input signals $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ are obtained by applying the OMP algorithm. In such a procedure, the nearest integer number to a percentage value of 12.5 of M is selected and fixed. The reason for having chosen this percentage value is because it presented the best trade-off between representativity and discriminability of the segments. Thus, sparsity values of $q = 16$, $q = 32$, and $q = 64$ are selected to represent the input signals for training in terms of the full dictionaries $\Phi 1$, $\Phi 2$, and $\Phi 4$, respectively.

Histograms are typically used to approximate data distributions. In this work we make use of histograms of the atom's activations to approximate the PDFs. The discretization process was performed by using a Δ value of 0.5. The detection of the most discriminative atoms is obtained by maximizing the discrepancy between the conditional PMFs of the atom's activations given the classes. This objective is achieved using the proposed DCAF measure as well as those denoted by KL, J, JS and F. The application of different discrepancy measures to the sparse vectors allows for the selection of different "discriminative atoms", which implies the construction of discriminative sub-dictionaries which are essentially different. The construction of sub-dictionaries, here denoted by $\Phi 1^{\mathbf{d}}$, $\Phi 2^{\mathbf{d}}$, and $\Phi 4^{\mathbf{d}}$, is performed by selecting atoms from $\Phi 1$, $\Phi 2$, and $\Phi 4$, respectively. Once the most discriminative atoms are detected, the sub-dictionary is built and consequently the feature vectors are obtained by applying the OMP algorithm. Finally each feature vector is assigned to be the input of the ELM classifier.

[3] `http://www.cs.technion.ac.il/~ronrubin/software.html`

6.3 Events detection and AHI estimation

Multilayer perceptron (MLP) neural networks trained for signal classification have proved to be a tool which provides quite good performances for OSAH syndrome detection [4], however, the process of training this class of neural network becomes very costly mainly in terms of time. For this reason, in this work we propose the use of extreme learning machine (ELM) [41] which is a type of single-hidden layer feedforward neural networks (SLFNs), instead of using MLP neural networks. Theoretically, this algorithm (ELM) results in providing a good generalization performance at extremely fast learning speed. The experimental results based on a few artificial and real benchmark function approximation and classification problems including large complex applications show that ELM can produce good generalization performance in most cases and can learn thousands times faster than conventional popular learning algorithms for feedforward neural networks [42].

Basic ELM classifier's MATLAB codes are available for download on the Guang-Bin Huang's web page[4]. To train such a classifier, the main parameters to be fixed are the number of neurons in the hidden layer as well as the activation function of the neurons. In our experiments, the number of neurons in the hidden layer of the ELM corresponds to four times the feature vector dimension. Also the well know sigmoid activation function, which is the most common activation function in the nodes of the hidden and/or output layer, is chosen.

In order to evaluate the performance of the proposed classifier in the detection of individual apnea-hypopnea events (a local approach), or more specifically, in the identification of persons suspected of suffering from moderate-severe OSAH syndrome (a global approach), three performance measures are used. For the identification of single segments containing apnea-hypopnea events, the sensitivity ($\mathrm{SE_{AH}}$) represents the total number of correctly classified segments of signals for which any apnea-hypopnea event occurred. Following the same idea, for the detection of individual segments of signals "not containing" any apnea-hypopnea event, the specificity ($\mathrm{SP_{AH}}$) is defined as the total number of correctly classified segments for which any apnea-hypopnea is not present. The accuracy ($\mathrm{AC_{AH}}$) is finally defined as follows:

$$\mathrm{AC_{AH}} \doteq \frac{1}{n}\sum_{i=1}^{n}\delta(c_i,\hat{c}_i), \tag{9}$$

where n represents the total number of segments, c_i and $\hat{c}_i$ denote the corresponding class label of the i^{th}-segment and the corresponding prediction of the classifier, respectively, and $\delta(x,y)$ represents the delta function whose output is true (one) if the condition $x = y$ is satisfied and false (zero) otherwise.

The differences in performance obtained for the event detection between each discrepancy measure were evaluated in order to test whether or not they are statistically significant. The test was performed assuming statistical independence of the classification errors for the different studies and approximating the error's Binomial distribution by means of a normal distribution. This assumptions are reasonable due to the large number of $\mathrm{SaO_2}$ signal segments available for each study (about 1100 segments per study, totaling 301306 segments).

The estimated AHI index ($\mathrm{AHI_{est}}$) is defined as the average number of predicted events per hour of study. This new index is used for OSAH syndrome detection. In this case, the sensitivity ($\mathrm{SE_{OSAH}}$) is defined as the ratio of persons with OSAH syndrome for whom the final test is positive, and the specificity ($\mathrm{SP_{OSAH}}$) is defined as the ratio of health patients for whom the final test is negative. Also the area under the ROC curve (AUC) derived from a receiver operating characteristic (ROC) analysis [43] is used. A ROC analysis consists of computing the values of the sensitivity and specificity across all the possible detection threshold (DT) values. Then, the ROC curve is built by performing a plot of 1-specificity versus sensitivity values. This curve has been widely used by medical physicians for evaluating diagnostic tests [44]. A comparison between two different methods can be effectively done by finding the "optimal" (in certain sense) cut-off point of the curve and evaluating their corresponding performances. Finally, the accuracy $\mathrm{AC_{OSAH}}$ is defined as follows:

$$\mathrm{AC_{OSAH}} \doteq \frac{1}{m}\sum_{i=1}^{m}\delta(\mathrm{AHI}_{\mathrm{est}}^{(i)} > \mathrm{DT}, \mathrm{AHI}^{(i)} > 15), \tag{10}$$

where m corresponds to the total number of studies coming from the testing dataset and "DT" is the detection threshold value which adjusts over-estimation of the events produced in the segmentation process. The value of DT results in the best cut-off point of the ROC curve. This point, which maximizes simultaneously sensitivity and specificity, corresponds to the minimum euclidean distance (d_{min}) to the point (0;1) of the ROC curve.

[4] `http://www.ntu.edu.sg/home/egbhuang/elm_codes.html`

7 Results and discussion

In this section results of the performed experiments are presented and discussed. This section is mainly separated into two sub-sections, namely *i*) the performance tunning section and *ii*) the optimal system performance section.

7.1 Performance tunning

This section presents results of the exploratory experiments performed to find optimal configurations of the proposed system. As explained in Section 6.2, three different full dictionaries called $\Phi 1$, $\Phi 2$ and $\Phi 4$ were learned by applying the standard KSVD algorithm. In this process, it is expected that most dictionary atoms would capture high frequency oscillations and normal respiration cycles in SaO_2 signals. It is important to point out however that, typical desaturations in signals associated to apnea-hypopnea events should be encoded by some atoms. Secondly, the sparse matrices $\mathbf{A}1$, $\mathbf{A}2$ and $\mathbf{A}4$ were obtained by applying the OMP algorithm. As described in Section 6.2 several measures were used to quantify the discriminative degree of individual atoms of each one of the studied dictionaries. Finally, the dictionary atoms were ranked in decreasing order of magnitude according to their discriminative power. Figure 7 shows the waveforms of the first seven ranked atoms of the dictionary $\Phi 1$ according to our measure (first row) as well as the first seven ranked atoms of such a dictionary according to all other discrepancy measures (rows from two to five). It can be seen that the most discriminative atom selected by DCAF (dashed waveform) provides information about two well-defined desaturations in the signal. It is also important to point out that, this atom corresponds to the most discriminative one when using J divergence, or eventually when using the JS divergence. Moreover, one can clearly note that no highly discriminative atoms were taken when using Fisher score.

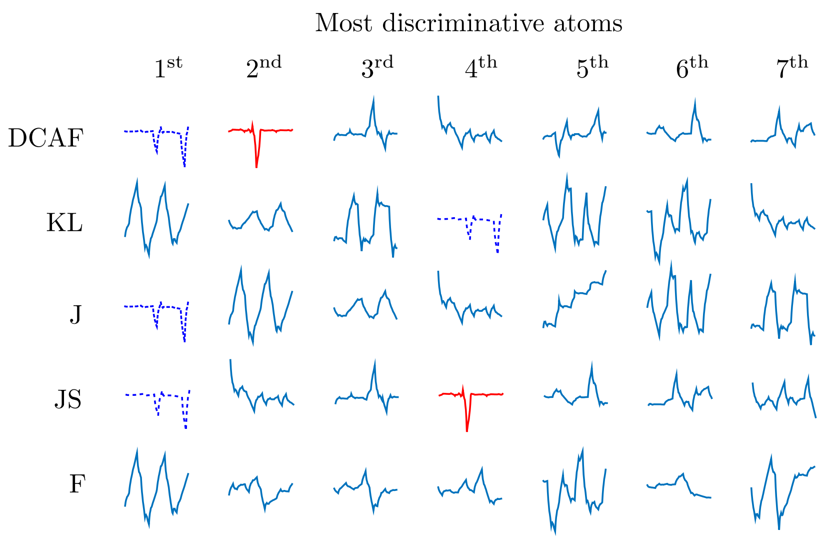

Figure 7: Waveforms corresponding to the first seven ranked atoms according to each one of the evaluated measures.

Discriminative sub-dictionaries called $\Phi 1^{\mathbf{d}}$, $\Phi 2^{\mathbf{d}}$ and $\Phi 4^{\mathbf{d}}$ were built by stacking side-by-side the first p ranked atoms from $\Phi 1$, $\Phi 2$ and $\Phi 4$, respectively, according to their discriminative degree. It is appropriate to mention that, the evaluation of several discrepancy measures leads to the construction of different discriminative sub-dictionaries. However, optimal values of p (sub-dictionary size) and q (sparsity level) are parameters that need to be tuned. In order to find optimal values of such hyper-parameters, a grid search was performed.

The performance of our system was first tested by performing a "random selection" of the dictionary atoms. The involved results were fixed and appropriately used as reference. The random selection of the atoms was performed ten times. Additionally, for each one of the atoms random selection, 60 iterations of the grid search were performed. Thus, the accuracy rate's variations introduced by the classifier were minimized. Figure 8 shows three images corresponding to averaged accuracy rates for each one of the evaluated dictionaries. Averaged accuracy rates (reference values) obtained by using the dictionary $\Phi 1$ for the detection of apnea-hypopnea events are shown on the left of this figure. It can be seen that sparse representations in terms of $\Phi 1$, using the smallest sub-dictionary size and the highest sparsity degree, result in better performance than the ones obtained by using all other configurations of Φ1 and the over-complete dictionaries Φ2 and Φ4. In this way, two regions can be distinguished corresponding to a high performance region and a low performance one. The first one, which is or our interest, is yielded by simultaneously employing a small sub-dictionary size (10%) and a high sparsity degree (5%).

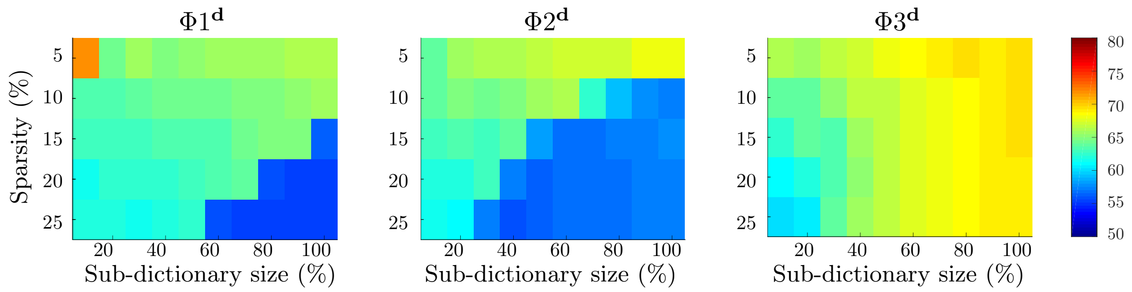

Figure 8: Averaged accuracy rates obtained by varying the percentages of the sub-dictionary size and the sparsity level according to a random ranking of the atoms.

Next, DCAF and four other discrepancy measures were used for appropriately constructing discriminative sub-dictionaries. Then, a grid search of hyper-parameters was performed by analyzing the performance that yields our system when using each one of the sub-dictionaries. Figure 9 shows five images corresponding to DCAF (upper-left) and the other four discrepancy measures. These images represent the differences between accuracy rates obtained by using discriminative measures and the reference one (random selection) for Φ1. Also, each pixel of these images correspond to particular percentages of sub-dictionary size and sparsity level. It can be observed that, independently of the discriminative measure, small percentages of sub-dictionary size yield good performances. It is appropriate to point out however that, the effect of the dimension (sub-dictionary size) in the performance of the system is more important than the one induced by using discriminative measures.

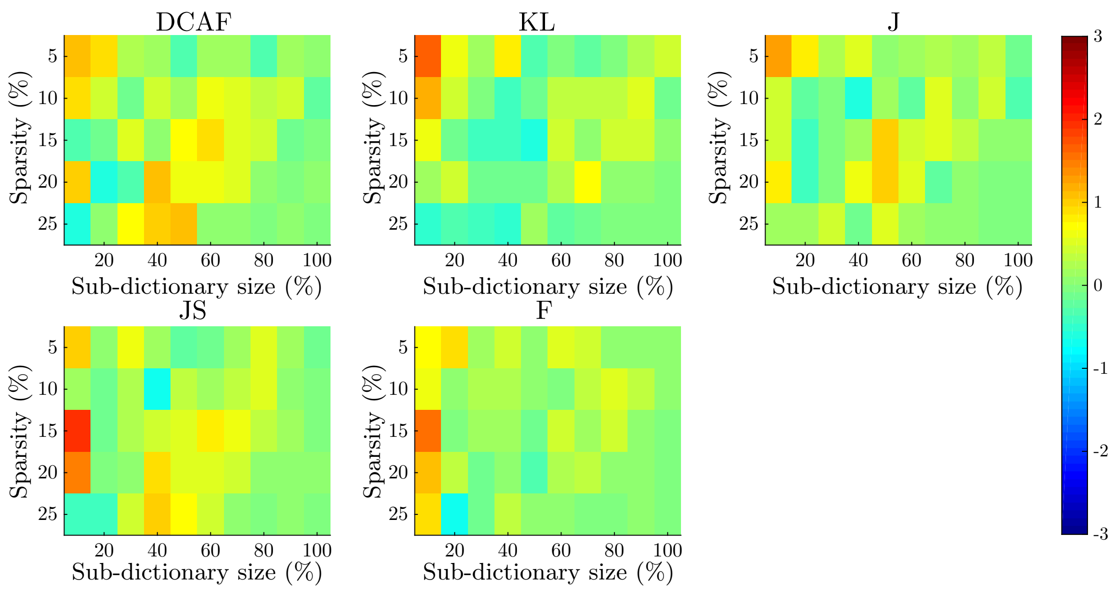

Figure 9: Five images representing differences between accuracy rates yielded by DCAF and all other discrepancy measures and random selection for Φ1.

Analogously, Figures 10 and 11 show five images which correspond to DCAF (upper-left) and all other discrepancy measures. The images depicted in Figures 10 and 11 represent the differences between accuracy rates obtained by using these measures and the reference one for dictionaries Φ2 and Φ4, respectively.

If we compare the results shown in Figures 9, 10, and 11, then it can be conclude that the proposed system presents the best performance, in terms of accuracy rate in the detection of apnea-hypopnea events, when using the full dictionary Φ1. Although similar results were obtained applying the proposed DCAF measure and those traditional ones (see Figure 9), it is important to point out that the use of

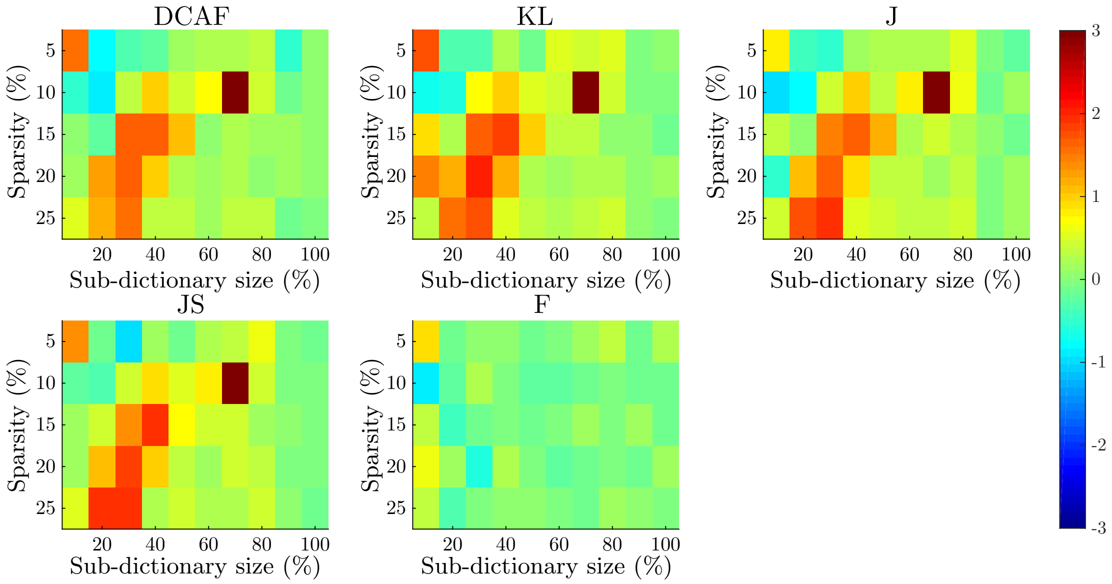

Figure 10: Five images representing differences between accuracy rates yielded by DCAF and all other discrepancy measures and random selection for Φ2.

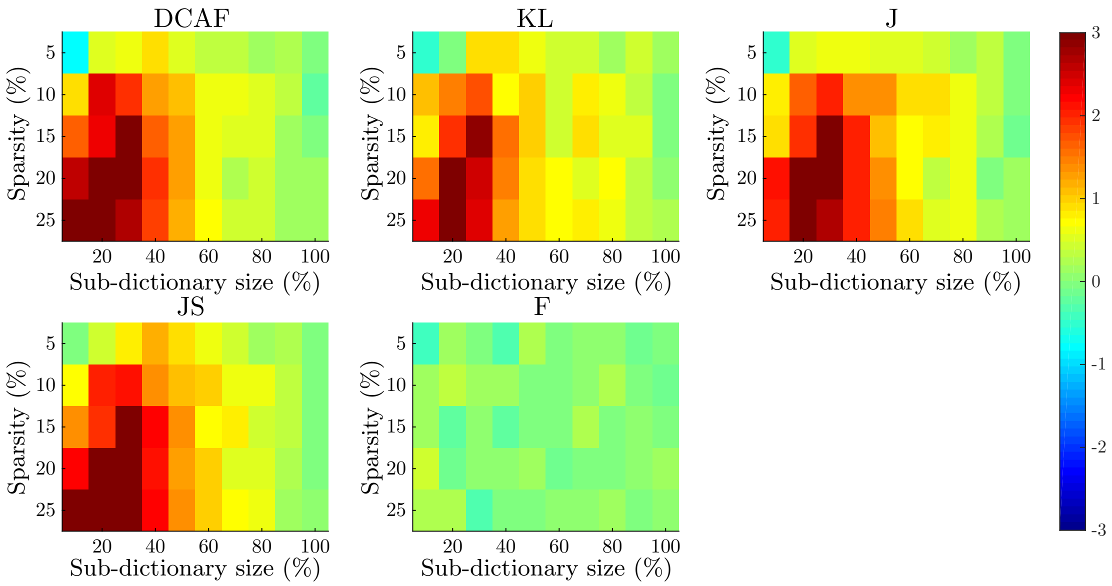

Figure 11: Five images representing differences between accuracy rates yielded by DCAF and all other discrepancy measures and random selection for Φ4.

discrepancy measures resulted in a significantly high improvement with respect to a "random" selection of the atoms. As discussed above, the dimension reduction in the sub-dictionary size as well as high sparse levels yielded high accuracy rates. This is the reason for which a small sub-dictionary size (10%) and high sparse level (5%) were chosen to perform the final test.

System performance changes were analyzed by performing a comparison between averaged accuracy rates obtained by using discriminative sub-dictionaries and the ones obtained by using full dictionaries. Table 1 shows averaged accuracy percentages obtained by taken into account fixed discriminative sub-dictionary sizes (10%) while allowing the sparsity level to change (rows from 3 to 7). The last row of this table presents averaged accuracy percentages yielded by using full dictionaries for different sparsity levels. It can be observed that, in all of cases, discriminative sub-dictionaries outperform full dictionaries in the detection of apnea-hypopnea events.

The impact of sparsity degree in the performance of our system is illustrated in Table 2. These results

Table 1: Averaged accuracy rates for sub-dictionary sizes of 10% regarding to each one of the evaluated full dictionaries.

	$\Phi 1^{\mathbf{d}}(128 \times 12)$		$\Phi 2^{\mathbf{d}}(128 \times 24)$		$\Phi 4^{\mathbf{d}}(128 \times 50)$	
Measure	Max	Avg	Max	Avg	Max	Avg
DCAF	72.62	64.68	65.20	63.15	65.19	64.21
KL	73.20	64.91	65.44	63.53	65.42	63.66
J	72.82	64.88	64.50	62.82	65.39	63.68
JS	72.55	64.10	65.02	63.18	65.87	64.01
F	72.23	65.21	64.57	63.04	65.64	62.71
Full dictionary	66.39	59.77	68.13	59.57	69.28	69.21

were yielded by averaging accuracy rates obtained for a sparsity level of 5% and considering all possible sub-dictionary sizes (from 10% to 90%). For example, the second row shows averaged accuracy rates obtained by means of discriminative sub-dictionaries whose atoms were taken from $\Phi 1$, $\Phi 2$ and $\Phi 4$ by using DCAF measure.

Table 2: Averaged accuracy rates by considering a sparsity level of 5% regarding to all possible sub-dictionary sizes.

Measure	$\Phi 1$	$\Phi 2$	$\Phi 4$
DCAF	66.41	66.51	67.95
KL	66.49	66.72	67.98
J	66.60	66.56	67.98
JS	66.41	66.57	68.15
F	66.53	66.54	67.58

7.2 Optimal system performance

Optimal system configurations were selected and fixed to perform the final test. In the previous section it was found that discriminative sub-dictionaries constructed by taken atoms from the dictionary $\Phi 1$ yields better performances than the ones constructed by selecting atoms from the dictionaries $\Phi 2$ and $\Phi 4$. Additionally, it was found that a discriminative sub-dictionary composed by only 12 atoms (10%) and a sparsity level of one (5%) yield in the best accuracy rate of our system.

In order to overcome the variance introduced by ELM predictors, 60 repetitions of the testing process were performed. Table 3 shows percentage values of minimum (Min), maximum (Max), average (μ) and standard deviation (σ) corresponding to obtained accuracy rates in the detection of apnea-hypopnea events. Although, DCAF perform similarly to the four other discrepancy measures, its performance is achieved with a relatively low computational cost. Additionally, results show that performances obtained by using discriminative measures for constructing sub-dictionaries always outperform the ones yielded by making use of randomly constructed sub-dictionaries.

Table 3: Averaged accuracy rates for a sub-dictionary percentage of 10 for the detection of apnea-hypopnea events.

Measure	Min	Max	μ	σ
DCAF	71.72	73.14	72.57	0.345
Kullback-Leibler	72.06	73.78	73.26	0.390
Jeffrey	71.77	73.31	72.66	0.319
Jensen-Shannon	71.79	73.11	72.55	0.295
Fisher	71.01	72.77	72.18	0.325
Random Selection	70.01	71.51	70.91	0.372

We have also evaluated the statistical significance of the results presented in Table 3 by computing the probability that using each one of the evaluated measures, including random selection (RS), yields in better classification performances than the others. In order to perform this test, we assumed the statistical independence of the classification errors for each study. Also it was possible to approximate the error's binomial probability distribution by a normal distribution due to a wide availability of signals (301,306). Table 4 summarizes the results of the performed statistical significance tests by considering a p-value of 0.01. It can be seen that DCAF and three other discrepancy measures (KL, J and JS divergences) are significant different respect to random selection. Also, no significant difference was found between F score and random selection. Additionally is was found that DCAF does not perform significantly better that the KL, J and JS divergences.

Table 4: A summary of the performed statistical significance tests.

	RS	DCAF	KL	J	JS	F
RS	-	✓	✓	✓	✓	✗
DCAF	-	-	✗	✗	✗	✗
KL	-	-	-	✗	✗	✗
J	-	-	-	-	✗	✗
JS	-	-	-	-	-	✗
F	-	-	-	-	-	-

To determine the severity degree of OSAH syndrome, a ROC curve analysis was successfully performed by considering a detection AHI of 15. This index was selected in order to identify patients suspected of suffering from moderate-severe OSAH syndrome. Table Table 5 shows the minimum operating (cut-off) point of the ROC curves and maximum percentages of sensitivity, specificity and accuracy as well as maximum values of area under the ROC curve for AHI diagnostic threshold values of 15 (Figure 12 (left)). It can be seen that DCAF resulted in a maximum area under the ROC curve of 0.9250 and sensitivity and specificity percentages of 81.88 and 87.32, respectively. These are the maximum performance measures at which the minimum cut-off point of the ROC curve is attained. If we compare the performances attained between all of the evaluated measures, then the maximum SE and AUC value is yielded by J divergence. Also, JS divergence outperformed all the others in terms of ACC and DCAF resulted in the minimum cut-off point of the ROC curve.

Table 5: Maximum cut-off points for testing accuracy for a sub-dictionary percentage of 10 for the detection of apnea-hypopnea events.

Measure	d_{min}	SE	SP	ACC	AUC
DCAF	0.2211	81.88	87.32	84.60	0.9250
Kullback-Leibler	0.2242	81.46	87.39	84.43	0.9271
Jeffrey	0.2311	80.86	87.04	83.95	0.9283
Jensen-Shannon	0.2267	80.75	88.03	84.39	0.9244
Fisher	0.2280	80.66	87.91	84.29	0.9252

We additionally performed a ROC curve analysis of the averaged performances of DCAF and all the other discrepancy measures (Figure 12 (right)). A random selection was additionally included in our results in order to be able to compare performance changes. Table 6 the averaged minimum operating (cut-off) point of the ROC curves and averaged maximum percentages of sensitivity, specificity and accuracy as well as averaged maximum values of AUC values for the same OSAH syndrome diagnostic threshold. The result show that DCAF outperforms all the other discrepancy measures in terms of minimum optimal operating cut-off point of the ROC curve as well as in terms of sensitivity and accuracy rate. Also KL divergence resulted in the best averaged area under the curve ROC and the maximum averaged specificity was yielded by JS divergence. A significant performance improvement was observed when using DCAF or any of the other discrepancy measures compared to random selection.

Several applications exist where it is desirable to maximize the sensitivity. For instance, if the primary purpose of the test is "screening", i.e. detection of early disease in a large numbers of apparently healthy persons, then a high sensitivity is generally desired. With this in mind, if a sensitivity of 98% is chosen in the ROC curves in Figure 12, for all used measures, the method achieves a specificity close to 45%. This fact shows that the analysis of pulse oximetry signals by means of the proposed method could be potentially applied as an efficient diagnostic screening tool in clinical practice.

In a previous work [4] it was shown that the MDCS method using DCAF to select discriminative atoms in a given dictionary, provides good accuracy rates in the detection of apnea-hypopnea events. In that work, a comparative analysis of the performances yielded by MDCS and other methods [45, 46, 47] has shown that MDCS outperforms all the others. It was also observed that the computational cost of MDCS is slightly higher than those required by the other three methods. On the other hand, in this work we show that MDCS using DCAF for selecting discriminative atoms performs similarly than MDCS using several other traditional discrepancy measures. It is important to highlight that DCAF is very easy to compute and yields competitive performance rates in the detection of apnea-hypopnea events at a low computational cost.

Table 6: Averaged cut-off points for testing accuracy for a sub-dictionary percentage of 10 for the detection of apnea-hypopnea events.

Measure	d_{min}	SE	SP	ACC	AUC
DCAF	0.2211	81.88	87.32	84.60	0.9250
Kullback-Leibler	0.2242	81.46	87.39	84.43	0.9271
Jeffrey	0.2311	80.86	87.04	83.95	0.9283
Jensen-Shannon	0.2267	80.75	88.03	84.39	0.9244
Fisher	0.2280	80.66	87.91	84.29	0.9252
Random Selection	0.2396	80.85	85.60	83.23	0.9222

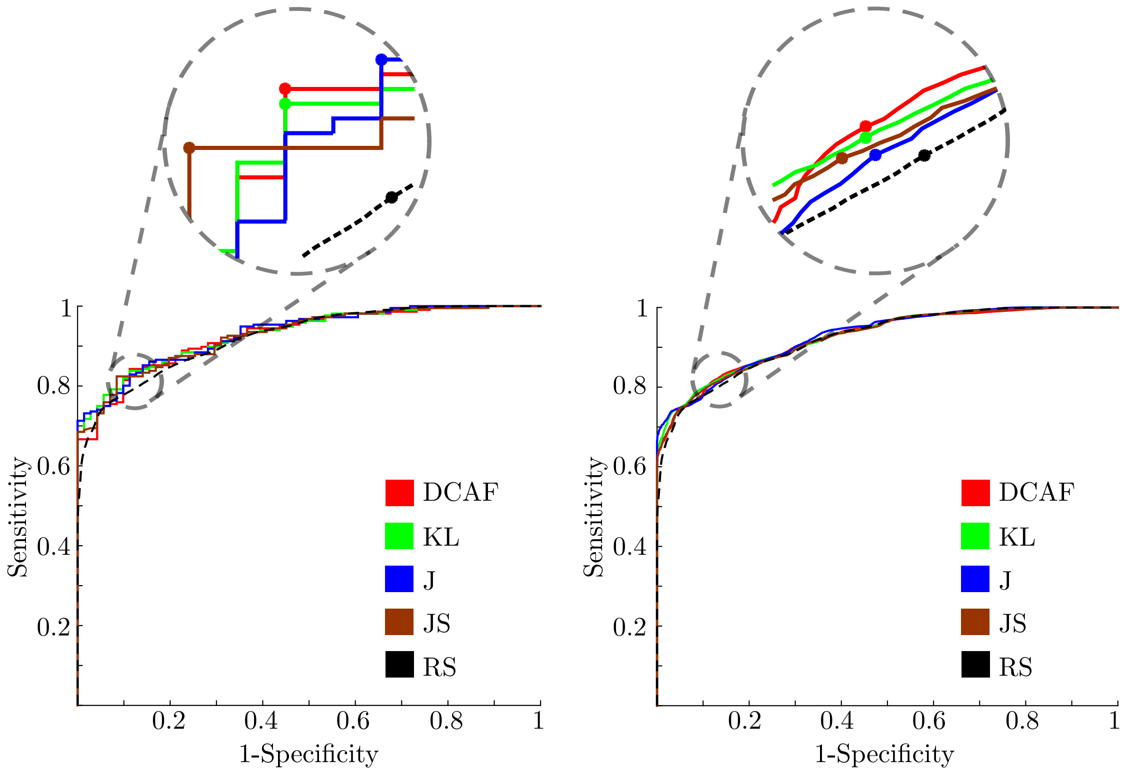

Figure 12: ROC curves corresponding to the performance measures described in Tables 5 and 6.

8 Conclusions

Sparse representations of signals constitute a powerful technique which yields high accuracy rates in the detection of apnea-hypopnea events. In this work the difference of conditional activation frequency (DCAF) measure was successfully used for accurately pointing out discriminative atoms in a full dictionary. Additionally, we compared the performance of the DCAF with four widely used discrepancy measures. It was found that the DCAF and three other discrepancy measures (KL, J y JS divergences) outperform the random selection of atoms, unlike F score. Additionally, DCAF is cheaper to compute. Discriminative sub-dictionaries were successfully constructed by taking the best ranked atoms of full dictionaries according to their discriminative power. Results show that sparse representations of signals in terms of discriminative sub-dictionaries result in better performances than the ones obtained in terms of full dictionaries in the detection of apnea-hypopnea events by using only pulse oximetry signals. In this context, it was found that more sparse solutions almost always yielded in better performances. Additionally, it was observed that larger dictionary over-completeness worsens the performance of the system.

Future research lines include more analysis of the DCAF measure, the study of its properties and an extension of such a measure to multi-class problems, among others.

9 Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO and PIP 2012-2014 Nro. 114 20110100284-KA4, by the Air Force Office of Scientific Research, AFOSR / SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral through projects CAI+D PIC Nro. 504 201501 00098 LI (2016) and PIC Nro. 504 201501 00036 LI (2016). The authors would like to thank Dr. Luis D. Larrateguy, who is a specialist in sleep-related disorders, for his valuable comments and suggestions.

References

[1] N. Saito and R. R. Coifman, "Local discriminant bases and their applications," *Journal of Mathematical Imaging and Vision*, vol. 5, no. 4, pp. 337–358, 1995.

[2] S. Tabibian, A. Akbari, and B. Nasersharif, "Speech enhancement using a wavelet thresholding method based on symmetric Kullback–Leibler divergence," *Signal Processing*, vol. 106, pp. 184–197, 2015.

[3] M. Sánchez-Gutiérrez, E. M. Albornoz, H. L. Rufiner, and J. G. Close, "Post-training discriminative pruning for rbms," *Soft Computing*, pp. 1–15, 2017.

[4] R. E. Rolón, L. D. Larrateguy, L. E. Di Persia, R. D. Spies, and H. L. Rufiner, "Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection," *Biomedical Signal Processing and Control*, vol. 33, pp. 358–367, 2017.

[5] V. Peterson, H. L. Rufiner, and R. D. Spies, "Generalized sparse discriminant analysis for event-related potential classification," *Biomedical Signal Processing and Control*, vol. 35, pp. 70–78, 2017.

[6] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[7] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[8] A. Gupta, S. Parameswaran, and C.-H. Lee, "Classification of electroencephalography (EEG) signals for different mental activities using Kullback Leibler (KL) divergence," pp. 1697–1700, 2009.

[9] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[10] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Trans. Inf. Theor.*, vol. 37, no. 1, pp. 145–151, 2006.

[11] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[12] X. Zhang and Q. Ding, "Respiratory rate estimation from the photoplethysmogram via joint sparse signal reconstruction and spectra fusion," *Biomedical Signal Processing and Control*, vol. 35, pp. 1–7, 2017.

[13] Y. Zhou, X. Hu, Z. Tang, and A. C. Ahn, "Sparse representation-based ECG signal enhancement and QRS detection," *Physiological Measurement*, vol. 37, no. 12, pp. 2093–2110, 2016.

[14] M. Aharon, M. Elad, and A. Bruckstein, "KSVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311–4322, Nov. 2006.

[15] D. S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[16] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2691–2698, June 2010.

[17] M. J. Sateia, "International classification of sleep disorders-third edition: Highlights and modifications," *Chest*, vol. 146, pp. 1387–1394, Nov. 2014.

[18] N. A. Dewan, F. J. Nieto, and V. K. Somers, "Intermittent hypoxemia and OSA: implications for comorbidities," *Chest*, vol. 147, no. 1, pp. 266–274, 2015.

[19] W. Kukwa, E. Migacz, K. Druc, E. Grzesiuk, and A. M. Czarnecka, "Obstructive sleep apnea and cancer: effects of intermittent hypoxia?," *Future Oncology (London, England)*, vol. 11, no. 24, pp. 3285–3298, 2015.

[20] M. Torres, R. Laguna-Barraza, M. Dalmases, A. Calle, E. Pericuesta, J. M. Montserrat, D. Navajas, A. Gutierrez-Adan, and R. Farré, "Male fertility is reduced by chronic intermittent hypoxia mimicking sleep apnea in mice," *Sleep*, vol. 37, no. 11, pp. 1757–1765, 2014.

[21] T. Young, L. Evans, L. Finn, and M. Palta, "Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women," *Sleep*, vol. 20, no. 9, pp. 705–706, 1997.

[22] J. Durán, S. Esnaola, R. Rubio, and A. Iztueta, "Obstructive sleep apnea-hypopnea and related clinical features in a population-based sample of subjects aged 30 to 70 yr," *American Journal of Respiratory and Critical Care Medicine*, vol. 163, pp. 685–689, 2001.

[23] R. Thurnheer, K. E. Bloch, I. Laube, M. Gugger, M. Heitz, and Swiss Respiratory Polygraphy Registry, "Respiratory polygraphy in sleep apnoea diagnosis. Report of the Swiss respiratory polygraphy registry and systematic review of the literature," *Swiss Medical Weekly*, vol. 137, no. 5-6, pp. 97–102, 2007.

[24] E. García-Díaz, E. Quintana-Gallego, A. Ruiz, C. Carmona-Bernal, A. Sánchez-Armengol, G. Botebol-Benhamou, and F. Capote, "Respiratory polygraphy with actigraphy in the diagnosis of sleep apnea-hypopnea syndrome," *Chest*, vol. 131, no. 3, pp. 725–732, 2007.

[25] A. Yadollahi, E. Giannouli, and Z. Moussavi, "Sleep apnea monitoring and diagnosis based on pulse oximetry and tracheal sound signals," *Medical & Biological Engineering & Computing*, vol. 48, no. 11, pp. 1087–1097, 2010.

[26] M. Elad, *Sparse and Redundant Representations.* Springer-Verlag New York, 2010.

[27] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[28] J. Tropp and A. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[29] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, "Signal processing and compression with wavelet packets," in *Wavelets and Their Applications*, pp. 363–379, Springer, Dordrecht, 1994.

[30] M. S. Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America A*, vol. 16, no. 7, p. 1587, 1999.

[31] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[32] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 2443–2446, 1999.

[33] Z. Jiang, Z. Lin, and L. Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2651–2664, Nov. 2013.

[34] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, 1989.

[35] W. Gersch, F. Martinelli, J. Yonemoto, M. D. Low, and J. A. Mc Ewan, "Automatic classification of electroencephalograms: Kullback-Leibler nearest neighbor rules," *Science (New York, N.Y.)*, vol. 205, no. 4402, pp. 193–195, 1979.

[36] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications," in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. K. Saul, and P. B. Schölkopf, eds.), pp. 1385–1392, MIT Press, 2004.

[37] C. C. Aggarwal, *Data classification: algorithms and applications.* CRC Press, 2014.

[38] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl, "The Sleep Heart Health Study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[39] B. K. Lind, J. L. Goodwin, J. G. Hill, T. Ali, S. Redline, and S. F. Quan, "Recruitment of healthy adults into a study of overnight sleep monitoring in the home: experience of the Sleep Heart Health Study," *Sleep and Breathing = Schlaf and Atmung*, vol. 7, no. 1, pp. 13–24, 2003.

[40] F. Lestussi, L. Di Persia, and D. Milone, "Comparison of on-line wavelet analysis and reconstruction: with application to ECG," *5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2011)*, 2011.

[41] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[42] J. Tang, C. Deng, and G. B. Huang, "Extreme Learning Machine for Multilayer Perceptron," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, 2016.

[43] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, no. 2, pp. 109–121, 1979.

[44] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian Journal of Internal Medicine*, vol. 4, no. 2, pp. 627–635, 2013.

[45] E. Chiner, J. Signes-Costa, J. M. Arriero, J. Marco, I. Fuentes, and A. Sergado, "Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies?," *Thorax*, vol. 54, pp. 968–971, Nov. 1999.

[46] J.-C. Vázquez, W. H. Tsai, W. W. Flemons, A. Masuda, R. Brant, E. Hajduk, W. A. Whitelaw, and J. E. Remmers, "Automated analysis of digital oximetry in the diagnosis of obstructive sleep apnoea," *Thorax*, vol. 55, pp. 302–307, Apr. 2000.

[47] G. Schlotthauer, L. E. Di Persia, L. D. Larrateguy, and D. H. Milone, "Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry," *Medical Engineering and Physics*, vol. 36, pp. 1074–1080, Aug. 2014.