

Kernel Recursive Least Squares Dictionary Learning Algorithm

Ghasem Alipoor¹

alipoor@hut.ac.ir

Karl Skretting²

karl.skretting@uis.no

¹Electrical Engineering Department, Hamedan University of Technology, Hamedan 6516913733, Iran

²Department of Electrical and Computer Engineering, IDE, University of Stavanger, N-4036, Stavanger, Norway

Abstract

An online dictionary learning algorithm for kernel sparse representation is developed in the current paper. In this framework, the input signal nonlinearly mapped into the feature space is sparsely represented based on a virtual dictionary in the same space. At any instant, the dictionary is updated in two steps. In the first step, the input signal samples are sparsely represented, in the feature space, using the dictionary that has been updated based on the previous data. In the second step, the dictionary is updated. In this paper, a novel recursive dictionary update algorithm is derived, based on the recursive least squares (RLS) approach. This algorithm gradually updates the dictionary, upon receiving one or a mini-batch of training samples. An efficient implementation of the algorithm is also formulated. Experimental results over two datasets in different fields show the superior performance of the proposed algorithm in comparison with its counterparts. In particular, the classification accuracy obtained by the dictionaries trained using the proposed algorithm gradually approaches that of the dictionaries trained in batch mode. Moreover, in spite of lower computational complexity, the proposed algorithm overdoes all existing online kernel dictionary learning algorithms.

Keywords: *Sparse Representation; Online Dictionary Learning; Kernel Methods; Recursive Least Squares; Classification*

1 Introduction

Data factorization methods have met with considerable success in discovering latent features of the signals encountered in wide-ranging applications. In this way, the representation bases, which make up the columns of the basis matrix or dictionary, are learned from the available samples of the target environment. An example is the sparse representation (SR) in which the dictionary is intended to best represent the data with a small number of atoms, much smaller than the dimension of the signal space. It has been shown that, in addition to a more informative representation of signals, imposing sparsity constraints on the representation coefficients can improve the generalization performance and the computational efficiency [1, 2, 3]. Furthermore, the sparse representation is more robust to noise, redundancy, and missing data. These features are mainly attributed to the fact that the intrinsic dimension of natural signals is usually much smaller than their apparent dimension and hence SR in an appropriate dictionary can extract these intrinsic features more efficiently. SR has been a successful strategy and has received considerable attention and achieved state-of-the-art results in many applications, e.g. vision, signal and image processing, and pattern recognition [1, 4, 2, 3].

SR entails two main problems. The first problem is sparse coding or Sparse Approximation (SA), in which, a data sample is represented using a given dictionary, based on an optimization criterion. Many optimization criteria and numerous solutions have been proposed for this purpose, mostly with matching or basis pursuit algorithms [5, 6, 7]. The second and more challenging problem is designing an effective and efficient dictionary [2]. Dictionaries can be either pre-determined or data-driven. It is well-accepted that data-driven dictionaries, learned from available training data samples via a procedure called Dictionary Learning (DL), generally can better extract the inherent features of signals and hence produce better results in many applications. DL is generally a combinatorial and highly non-convex optimization problem with two outputs, namely the learned dictionary and the representation coefficients. Nevertheless, this problem is convex with respect to each of these two arguments and is usually solved by alternating between two steps of SA (fixing the dictionary) and dictionary update (fixing the representation coefficients). Two main algorithms for DL are the Method of Optimal Directions (MOD) [8] and the K-Singular Value Decomposition (K-SVD) [9].

Classical SR-based methods represent data samples by linear combination of the atoms of the dictionary in the Euclidean space. These linear models are incapable of coping with nonlinearity encountered in many real applications. One intuitive remedy for this shortcoming is utilizing the kernel methods that have shown considerable success, in various fields [10, 11]. Many Kernel

SR (KSR) and KSR-based Classification (KSRC) algorithms have also been derived. To this end, data samples are first nonlinearly mapped into a higher (or even infinite) dimensional Reproducing Kernel Hilbert Space (RKHS), known as the feature space, and are then sparsely represented based on a dictionary in the same space. A few examples of the recent applications of the KSRC are medical applications [12, 13], image quality assessment [14], image classification [15], computer vision [16], domain adaption [17] and fault diagnosis [18].

Several SA algorithms have been straightforwardly extended to their kernel counterparts [19, 20, 21]. A large number of studies reported on the KSR have only considered this problem and have not touched the DL problem, where the dictionary is represented by a subset of the transformed training exemplars [22, 23, 20]. The Kernel DL (KDL) problem has been first addressed in [24] where in the update step, an approximate solution of the stochastic gradient descent approach is used to optimize the dictionary, atom by atom. The SA and DL problems in the spaces of symmetric positive definite (SPD) matrices were addressed in [25, 26, 27], using the fact that the SPD matrixes form a Riemannian manifold and then embedding this manifold into an RKHS. As a result of using the gradient descent in the original input space, the application of the abovementioned methods is restricted to specific kernels.

The milestone in KDL is [the Kernel Method of Optimal Direction \(KMOD\) algorithm proposed in \[21\]](#), in which Van Nguyen et al. have shown that the solution of the KDL, formulated as a Least-Squares (LS) minimization problem, can be represented by a linear combination of the training data samples. In particular, having L training data samples $\{x_l \in \mathbb{R}^N, l = 1, 2, \dots, L\}$ that make up the columns of the matrix $X \in \mathbb{R}^{N \times L}$, the optimum dictionary has the form

$$D^* = \phi(X)B \in \mathbb{R}^{F \times Q}, \quad (1)$$

where Q is the number of dictionary atoms and F is the dimension of the feature space or the length of each atom in that space. ϕ is a non-linear mapping from \mathbb{R}^N to the feature space \mathcal{F} and the coefficient matrix $B \in \mathbb{R}^{L \times Q}$ is optimized based on kernelized versions of K-SVD and MOD algorithms. Hence, the complex problem of learning a dictionary in the high (or even infinite) dimensional feature space can be simplified to finding the optimum matrix B , while staying in the original space. To the best of our knowledge, this simplification was adopted in all other similar studies [28, 29, 27, 30, 31].

Batch learning algorithms, in both linear and kernelized scenarios, incur a significant burden on computational and storage; as they generally require access to the entire batch of training data in each iteration. This problem

is more severe in the case of KSR, where, as a result of the representer theorem, the dictionary is represented by all training data samples. This issue was addressed in [31], in which a hierarchical DL approach was proposed that learns the matrix B in the coefficient space, independent of any computations involving the kernel matrix; but the Kernel SA stage still depends on the entire kernel matrix. Additionally, the Linearized Kernel Dictionary Learning (LKDL) method proposed in [32] does not solve this problem, as the complexity and the required storage still quadratically depend on the number of input samples. Moreover, both K-SVD and MOD algorithms, and their kernelized versions, suffer from the dilemmas of matrix inversion, namely complexity, and singularity.

To address the essential shortcomings of the batch algorithms, in some other fields, many algorithms have been proposed for online DL, but in the Euclidean space [33, 34, 35]. These algorithms can better cope with the dynamic change of the problem in non-stationary environments and have appealing applications in real-time problems, such as prediction and tracking problems. They have also been found useful in batch scenarios where the amount of the training data is too large to be accessed and processed at once. Among the proposed online linear DL algorithms, the current study more relates to the RLS DL algorithm proposed in [34] which is essentially an iterative version of the MOD algorithm. Nonetheless, a limited number of studies have been reported on the online KDL and this field of research is still infant. Reference [36] studies the problem of online dictionary learning on the SPD manifolds. The proposed algorithm is an online version of the batch DL algorithm proposed in [25] and hence is restricted to the Stein kernel. A few [online](#) KDL algorithms have also been proposed in [30, 37, 38] that they all rely on the representer theorem (1). Akin to other online kernel methods that use this theorem, it is necessary to duly control the size of the involved data samples; the set that we refer to it along with all corresponding variables as the profile.

The sparse KDL ([sKDL](#)) algorithm was proposed in [30] in which a row-sparsity constraint is imposed on the coefficient matrix B to restrict the dictionary to be represented based on a few training samples. This algorithm is then extended to online scenarios, where the profile size is limited by selecting the incoming samples based on the reconstruction error. A similar row-sparsity constraint idea is used in [the Online KDL with Row-Sparsity \(OKDLRS\) algorithm developed in \[37\]](#), where a new model is proposed to simultaneously perform profile selection and kernel dictionary learning. The formulated non-convex and non-smooth optimization problem is solved by proximal alternating linearized minimization. Two slightly different [Online KDL algorithms with a Fixed Budget \(OKDLFB\)](#) are also proposed in [38]

based on the Stochastic Gradient Descent (SGD) method. The difference between these two algorithms is in the space, whether feature space or parameter space, in which the dictionary update is performed.

In the current study, an online kernel dictionary learning algorithm is developed that recursively solves the LS minimization problem, with the aid of the Matrix Inversion Lemma (MIL). This algorithm is an extension of the RLS DL algorithm proposed in [34] to the feature spaces; however with very different concepts and challenges. This extension is done in a way similar to that of the work reported in [39] that kernelizes the RLS adaptive filtering algorithm for regression, but in a clearly different context and with absolutely different derivations. On the other hand, our algorithm can be viewed as an online extension to the batch KMOD algorithm of [21]. The derived algorithm is formulated efficiently and various strategies are also examined for restricting the size of the profile. The proposed algorithm differs from all previous online KDL algorithms in that, instead of the gradient method, our algorithm is based on the RLS approach. It is well-known that RLS algorithms are an order of magnitude faster in convergence than the gradient-based ones [40, 41]. Furthermore, our algorithm is not restricted to a specific kernel.

In summary, the main contributions of this study are as follows:

1. A novel online KDL algorithm is developed that is the first one that is based on the RLS method.
2. A relatively efficient implementation of the algorithm is derived.
3. The performance of the proposed algorithm in classification is studied and compared with those of other KDL algorithms over two disparate practical datasets.

After the introduction and the literature review presented in this section, the remaining sections of the paper are organized as follows. The original batch KDL problem is formulated in section 2 and its solution is derived therein. In section 3, this algorithm is extended to the online mode, by recursively updating the dictionary; all relevant aspects are also covered in this section. Experimental results are reported in section 4 and the paper is then concluded in the last section.

Notation review

In this paper, we assume the following sizes: $s < Q \leq L \leq L_{max} \leq L_{tot}$ and $F \leq \infty$. s is the number of non-zero elements in the SA of each data vector, Q is the number of atoms in the dictionary D , L is the number of training

vectors that are (still) remembered, L_{tot} is the total number of available training vectors. F is the dimensionality of the feature space that depends on the kernel used and can be rather large, e.g., for polynomial kernels, or even infinite, as for RBF kernels. Furthermore, N is the dimensionality of the feature space or the number of elements in each data vector, and M is the number of training input data samples in each mini-batch. As we will see, L can increase to L_{max} as subsequent training vectors are processed. To make the described algorithm more intelligible, the dimensions of all the variables involved in the algorithm are summarized in Table 1. In this table, all variables are shown without any subscript i ; moreover, these variables may be used in the text with different subscripts or diacritical marks.

Table 1: Dimensions of the variables involved in deriving the algorithm (Variables that are not described in the third column are the intermediate variables defined throughout the derivations.)

Variable	Dimension	Description
\mathbf{x}	$N \times M$	a mini-batch of M input vectors in the original space
X	$N \times L$	input matrix in the original space
ϕ, \mathbf{r}	$F \times M$	input and error vectors in the feature space
Φ, R	$F \times L$	input and error matrixes in the feature space
D	$F \times Q$	virtual dictionary in the feature space
$\mathbf{w}, \mathbf{u}, \mathbf{h}$	$Q \times M$	\mathbf{w} is (are) the coefficient vector(s).
W, U	$Q \times L$	W is the coefficient matrix.
C, Ψ	$Q \times Q$	$C = (WW^T + \xi I)^{-1}$ and Ψ is the dictionary Gram matrix.
\mathbf{k}, \mathbf{v}	$L \times M$	\mathbf{k} is (are) the kernel vector(s).
λ	$L \times 1$	forgetting factor vector
K, Λ	$L \times L$	kernel and diagonal forgetting factor matrixes
$\sigma_{\mathbf{x}}, \alpha$	$M \times M$	$\sigma_{\mathbf{x}} = k(\mathbf{x}, \mathbf{x})$
λ	1×1	forgetting factor

2 Kernel Least Squares Dictionary Learning

The purpose of a dictionary learning algorithm for KSR is to find an optimum dictionary for sparse representation of the input sample vectors, in a feature space induced by a kernel function. This problem can be solved iteratively, at each instant, in two steps.

- In the first step, having the dictionary updated in the preceding instant, SA of the mapped input vector is found. For this end, we use the KORMP algorithm, which is a kernelized version of the Order Re-

ursive Matching Pursuit (ORMP) algorithm, as described in [7].

- In the second step, having the coefficients for sparse representation of previous samples, the dictionary is updated. For this purpose, we develop a recursive algorithm based on the LS criterion.

The recursive solution to the LS problem is derived along the same lines as the RLS and the Kernel RLS (KRLS) adaptive filtering algorithms [39] and the RLS DL algorithm [34]. In this section, the objective function of the KDL is defined and the direct solution to the optimum dictionary is derived.

2.1 LS Solution

At each instant, the objective is to find a (virtual) dictionary in the feature space

$$D = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_Q] \in \mathbb{R}^{F \times Q} \quad (2)$$

for sparse representation of the input vectors up to the current time. In particular, we aim at a virtual dictionary D that the input vectors mapped into the feature space can be represented by a linear combination of a small number of atoms of that dictionary.

At each time instant, the memory consists of L input sample vectors $\{\mathbf{x}_j\}_{j=1}^L$ which can be represented as columns of a matrix

$$X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L] \in \mathbb{R}^{N \times L}. \quad (3)$$

The input vectors can, at least virtually, be mapped into the feature space $\phi_j = \varphi(\mathbf{x}_j)$ and these feature vectors can be represented as columns of a matrix

$$\Phi = [\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2), \dots, \varphi(\mathbf{x}_L)] = [\phi_1, \phi_2, \dots, \phi_L] \in \mathbb{R}^{F \times L}. \quad (4)$$

Moreover, we assume that, for each feature vector, the coefficients for a sparse representation in the feature space, i.e. $\{\mathbf{w}_j\}_{j=1}^L$, are available. The coefficient matrix is therefore defined as:

$$W = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_L] \in \mathbb{R}^{Q \times L}. \quad (5)$$

These coefficients are typically found by SA, e.g. using the KORMP algorithm, based on a dictionary from an earlier step/time in the DL process; perhaps one that is close to the optimal dictionary D^* .

Assuming that the approximation coefficients are available, we look only at the dictionary update step in the iterative two-step DL procedure. This

problem can be described as the following regularized (non-weighted) LS minimization problem

$$\begin{aligned} D^* &= \arg \min_D \sum_{j=1}^L \|\mathbf{r}_j\|_2^2 + \gamma \|D\|_F^2 \\ &= \arg \min_D \sum_{j=1}^L \|\phi_j - D\mathbf{w}_j\|_2^2 + \gamma \|D\|_F^2. \end{aligned} \quad (6)$$

$\|\cdot\|_F$ denotes the Frobenius norm, also known as the trace norm, and $\mathbf{r}_j = \phi_j - D\mathbf{w}_j$ is the approximation error of the j th input sample, in the feature space, using the dictionary D . Furthermore, γ is a regularization parameter to trade-off between the representation error and the energy of the atoms. Note that the first 2-norm is in the feature space. The regularization effect of γ is more easily seen from equation (8); adding a non-zero value is vital when WW^T is nearly non-invertible. Furthermore, a small non-zero γ is helpful for making the normalization step (described in subsection 3.4) needed less frequently.

We may also define

$$R = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_L] \in \mathbb{R}^{F \times L} \quad (7)$$

and express the LS problem (6) by matrixes in the Frobenius norm as:

$$\begin{aligned} D^* &= \arg \min_D \|R\|_F^2 + \gamma \|D\|_F^2 \\ &= \arg \min_D \|\Phi - DW\|_F^2 + \gamma \|D\|_F^2. \end{aligned} \quad (8)$$

The solution of (8) is

$$D^* = (\Phi W^T)(WW^T + \gamma I)^{-1} = (\Phi W^T)C, \quad (9)$$

where $C = (WW^T + \gamma I)^{-1}$ and I is the identity matrix. Alternatively

$$D^* = \Phi U^T, \quad (10)$$

where $U = CW = (WW^T + \gamma I)^{-1}W \in \mathbb{R}^{Q \times L}$. We see that each dictionary atom is a linear combination of the columns of Φ defined in (4), i.e. the data vectors mapped into the feature space. This is in accordance with model (1) which was first proposed in [21]; in the current work, $B = U^T \in \mathbb{R}^{L \times Q}$ is found adaptively.

2.2 Weighted LS Solution

Each term in (6) can be weighted with a given weight λ_j . Adjusting λ_j appropriately can control the memory span of the algorithm, make the algorithm less dependent on the initial dictionary, and improve both convergence properties as well as the representation ability of the resulting dictionary. All these weights can be collected into a column vector $\boldsymbol{\lambda}$ of length L or along the diagonal in a square diagonal matrix $\Lambda = \text{diag}(\boldsymbol{\lambda}) \in \mathbb{R}^{L \times L}$.

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_L]^T \in \mathbb{R}^{L \times 1}, \quad (11)$$

$$\Lambda = \text{diag}(\boldsymbol{\lambda}) \in \mathbb{R}^{L \times L}. \quad (12)$$

The regularized weighted LS (WLS) problem is defined as:

$$D^* = \arg \min_D \|(\Phi - DW) \Lambda^{0.5}\|_F^2 + \xi \|D\|_F^2, \quad (13)$$

where $\Lambda^{0.5}$ is a matrix whose elements are square roots of the elements of Λ . For now, ξ can be set to γ . It is noted that

$$\|(\Phi - DW) \Lambda^{0.5}\|_F^2 = \sum_{j=1}^L \{\lambda_j \|(\phi_j - D\mathbf{w}_j)\|_2^2\}. \quad (14)$$

The WLS solution is

$$D = (\Phi \Lambda W^T) (W \Lambda W^T + \xi I)^{-1} = (\Phi \Lambda W^T) C = \Phi U^T, \quad (15)$$

where C and U matrixes are defined as:

$$C = (W \Lambda W^T + \xi I)^{-1} \in \mathbb{R}^{Q \times Q}, \quad (16)$$

$$U = CW\Lambda = (W \Lambda W^T + \xi I)^{-1} W\Lambda \in \mathbb{R}^{Q \times L}. \quad (17)$$

Finally, the reconstruction in the feature space is

$$\tilde{\Phi} = DW = \Phi U^T W = \Phi \Lambda W^T (W \Lambda W^T + \xi I)^{-1} W. \quad (18)$$

Note that it is not generally possible to go from $\tilde{\Phi}$ to \tilde{X} , i.e. the reconstruction in the original input space. However, this representation is still meaningful and sensible; at least for applications in which signal reconstruction is not required, e.g. feature extraction and classification applications.

2.3 Kernel Sparse Approximation

One can perform Kernel sparse approximation without forming the dictionary in the feature space. For instance, the ORMP algorithm, derived in [7], can be extended to the kernel space straightforwardly. To this end, the input variables of the ORMP algorithm in the signal domain, including the dictionary D (F in [7]) and the signal vector \mathbf{x} , should be replaced with the corresponding variables in the feature space. More particularly, the kernelized ORMP algorithm, denoted as the KORMP algorithm, uses the dictionary Gram matrix Ψ , the inner product of the dictionary and data vectors that is indicated by \mathbf{h} , and $\sigma_x^2 = k(\mathbf{x}, \mathbf{x})$ as input variables to solve, in the feature space, for the approximation coefficient vector \mathbf{w} .

The dictionary Gram matrix Ψ can be expressed as:

$$\Psi = D^T D = U \Phi^T \Phi U^T = U K U^T \in \mathbb{R}^{Q \times Q}, \quad (19)$$

where $K = \Phi^T \Phi \in \mathbb{R}^{L \times L}$ is a symmetric positive definite kernel matrix with elements $K_{ij} = K_{ji} = \phi_i^T \phi_j = \langle \phi_i, \phi_j \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$. Usually, the simple kernel function $k(\cdot)$ is given rather than the related, and possibly more complicated, mapping function $\varphi(\cdot)$. Let the new data vector be \mathbf{x} , without subscript, to distinguish it from the data vectors in memory as stored in X . The inner product vector \mathbf{h} can be represented as:

$$\mathbf{h} = D^T \varphi(\mathbf{x}) = U \Phi^T \varphi(\mathbf{x}) = U \Phi^T \phi = U \mathbf{k} \in \mathbb{R}^{Q \times 1}, \quad (20)$$

where $\mathbf{k} = \Phi^T \phi \in \mathbb{R}^{L \times 1}$, and $k_j = \phi_j^T \phi = \langle \phi_j, \phi \rangle = k(\mathbf{x}_j, \mathbf{x})$. Moreover,

$$\sigma_x^2 = \phi^T \phi = \langle \phi, \phi \rangle = k(\mathbf{x}, \mathbf{x}). \quad (21)$$

Note that Ψ is independent of \mathbf{x} , while \mathbf{h} and σ_x^2 depend on \mathbf{x} . Using these variables, the KORMP algorithm returns the sparse coefficients \mathbf{w} , with s non-zero elements, and also the squared error, in the feature space, $\mathbf{r}^T \mathbf{r} = \|\mathbf{r}\|_2^2$.

3 Online Kernel Dictionary Learning

The common practice in online dictionary learning algorithms, such as the KRLS DL algorithm proposed here, is to repeat a process of SA followed by an update of the dictionary, for each new training vector \mathbf{x} (or a mini-batch of training vectors) that is presented to the algorithm. On the other hand, in kernel methods, the size of the involved matrixes, such as U , W , and K matrixes in the KRLS DL algorithm, grow unboundedly with time.

This makes it necessary to restrict the size of data by storing only the most informative samples and even removing some previously-stored samples from the memory. Therefore, the dictionary update step, performed by updating the involved matrixes or the profile, may require:

- adding one new training vector or a (mini-) batch of new training vectors
- removing one or a (mini-) batch of the previously stored training vectors
- dictionary normalization, so that all atoms have unit ℓ_2 -norm.

The iteration always starts by SA of new input data samples which returns the new coefficient vector(s) \mathbf{w} . The matrixes at the start are denoted by subscript $(\cdot)_i$, resulting after the previous iteration i . The step we consider is iteration $i + 1$ and the resulting matrixes are thus denoted by a subscript $(\cdot)_{i+1}$. Intermediate matrixes may be denoted by the hat symbol $(\hat{\cdot})$. For simplicity vectors for step $i + 1$ are usually denoted without a subscript, as \mathbf{x} and \mathbf{w} . The involved matrixes are updated depending on what action we want to do. In the following sections, the effect of each action on the involved matrixes is described.

3.1 Adding data vectors

One of the main steps in KRLS DL algorithm is to add the new training data vector and perhaps also apply a forgetting factor on previous data. Using a forgetting factor $\lambda \leq 1$, the weights of the previous approximation errors are reduced by this factor, while the new approximation is added with weight 1. Having $\lambda = 1$ is the dictionary update step without the forgetting factor. Note that this step increases the size of some matrixes used in the KRLS DL algorithm. Specifically, one term is added to the summation (14) and L is increased by one and also

$$\boldsymbol{\lambda}_{i+1} = [\lambda \boldsymbol{\lambda}_i^T, 1]^T \quad (22)$$

and

$$X_{i+1} = [X_i, \mathbf{x}] \quad \text{and} \quad W_{i+1} = [W_i, \mathbf{w}]. \quad (23)$$

Another option is to add a mini-batch, a set of some few M data vectors, instead of a single vector. This can be done by letting \mathbf{x} and \mathbf{w} in (23) be matrices of size $N \times M$ and $Q \times M$, respectively, and replacing the scalar element 1 in (22) with the M -length row vector I_M^T .

Referring to (16), to make it possible to derive a simple recursive update equation for C_i , here ξ_i is defined as:

$$\xi_i = \gamma \prod_{\lambda_j \in \Omega_i} \lambda_j, \quad (24)$$

where Ω_i is the set of forgetting factors that correspond to all input data samples that have engaged in dictionary learning, up to time instant i . Note that when no input sample is pruned (as described in subsection 3.2), $\prod_{\lambda_j \in \Omega_i} \lambda_j = \boldsymbol{\lambda}_i(1)$, i.e. the first element in the $\boldsymbol{\lambda}_i$ vector which equals the product of all previously used λ values. With this definition, C_i can be recursively updated as:

$$C_{i+1}^{-1} = \lambda C_i^{-1} + \mathbf{w}\mathbf{w}^T. \quad (25)$$

The term $\mathbf{w}\mathbf{w}^T$ is a $Q \times Q$ matrix; when $M \leq Q$, its rank can be up to M .

Applying Matrix Inversion Lemma (MIL):

$$(A + EFG)^{-1} = A^{-1} - A^{-1}E(F^{-1} + GA^{-1}E)^{-1}GA^{-1} \quad (26)$$

on (25) with $A = \lambda C_i^{-1}$, $E = \mathbf{w}$, $F = I_M$, and $G = \mathbf{w}^T$ gives

$$C_{i+1} = \lambda^{-1}C_i - \lambda^{-1}C_i\mathbf{w}(I_M + \mathbf{w}^T\lambda^{-1}C_i\mathbf{w})^{-1}\mathbf{w}^T\lambda^{-1}C_i.$$

As $(\lambda A)^{-1} = \lambda^{-1}A^{-1}$, this gives

$$C_{i+1} = \lambda^{-1}(C_i - C_i\mathbf{w}(\lambda I_M + \mathbf{w}^T C_i \mathbf{w})^{-1}\mathbf{w}^T C_i)$$

and

$$C_{i+1} = \lambda^{-1}(C_i - \mathbf{u}\alpha\mathbf{u}^T), \quad (27)$$

where,

$$\mathbf{u} = C_i\mathbf{w} \in \mathbb{R}^{Q \times M} \quad \text{and} \quad \alpha = (\lambda I_M + \mathbf{w}^T \mathbf{u})^{-1} \in \mathbb{R}^{M \times M}. \quad (28)$$

As C_i is symmetric positive semi-definite (PSD), $\mathbf{w}^T \mathbf{u} = \mathbf{u}^T \mathbf{w} = \mathbf{w}^T C_i \mathbf{w}$ and also α are $M \times M$ symmetric PSD matrixes. Hence, assuming that \mathbf{w} has full rank, these two matrixes are also full-rank.

Other relevant matrixes can also be recursively updated. The symmetric (and PSD) kernel matrix K , contains the results of evaluating the kernel function over all pairs of training data vectors in matrix X . By adding new data vectors, this matrix grows and its size increases from K_i of size $L_i \times L_i$

to K_{i+1} of size $L_{i+1} \times L_{i+1}$, where $L_{i+1} = L_i + M$. To add the effect of the new input vector(s), matrix K can be updated as:

$$K_{i+1} = [\Phi_i, \phi]^T [\Phi_i, \phi] = \begin{bmatrix} \Phi_i^T \Phi_i & \Phi_i^T \phi \\ \phi^T \Phi_i & \phi^T \phi \end{bmatrix} = \begin{bmatrix} K_i & \mathbf{k} \\ \mathbf{k}^T & \sigma_{\mathbf{x}}^2 \end{bmatrix}, \quad (29)$$

where $\phi = \varphi(\mathbf{x}) \in \mathbb{R}^{F \times M}$, $\mathbf{k} = \Phi_i^T \phi \in \mathbb{R}^{L_i \times M}$, and $\sigma_{\mathbf{x}}^2 = \phi^T \phi = k(\mathbf{x}, \mathbf{x}) \in \mathbb{R}^{M \times M}$.

Matrix $U = CW\Lambda \in \mathbb{R}^{Q \times L}$, that is defined in (17), also grows as W grows and L increases from L_i to $L_{i+1} = L_i + M$. Matrix U is used to find vector \mathbf{h} in (20) for SA and can be updated as (refer to appendix A.1):

$$U_{i+1} = \begin{bmatrix} U_i - \mathbf{u}\alpha \mathbf{v}^T, & \mathbf{u}\alpha \end{bmatrix}, \quad (30)$$

Note that λ is scalar, $\alpha \in \mathbb{R}^{M \times M}$, and we have used $\alpha \mathbf{u}^T \mathbf{w} = I_M - \lambda \alpha$ from (28) and defined:

$$\mathbf{v}^T = \mathbf{u}^T W_i \Lambda_i, \quad \mathbf{v} = \Lambda_i W_i^T \mathbf{u} \in \mathbb{R}^{L \times M}. \quad (31)$$

If $\varphi(\cdot)$ function is accessible, it is possible to update the dictionary from equation (15) as (refer to appendix A.2):

$$D_{i+1} = D_i + \mathbf{r}\alpha \mathbf{u}^T, \quad (32)$$

since $\Phi_i \mathbf{v} = \Phi_i \Lambda_i W_i^T C_i \mathbf{w} = \Phi_i U_i^T \mathbf{w} = D_i \mathbf{w} = \tilde{\phi}$ and $\mathbf{r} = \phi - \tilde{\phi}$.

Updating Ψ , which is PSD of size $Q \times Q$, is more complicated; referring to appendix A.3, it can be done as:

$$\Psi_{i+1} = \Psi_i + \mathbf{u}\alpha \tilde{\mathbf{u}}^T + \tilde{\mathbf{u}}\alpha \mathbf{u}^T + \mathbf{u}\alpha(\mathbf{v}^T K_i \mathbf{v} - \mathbf{v}^T \mathbf{k} - \mathbf{k}^T \mathbf{v} + \sigma_{\mathbf{x}}^2)\alpha \mathbf{u}^T, \quad (33)$$

where we have defined:

$$\tilde{\mathbf{u}} = U_i(\mathbf{k} - K_i \mathbf{v}) \in \mathbb{R}^{Q \times M}. \quad (34)$$

Note that, for $M = 1$, all these update equations involve only matrix-by-vector multiplications. In this case, by recasting the initial matrix-by-matrix multiplication to only matrix-by-vector multiplications, the complexity is reduced from $O(QL^2)$ to $O(L^2)$, since the largest matrix involved is the $L \times L$ matrix K_i .

3.2 Removing data vectors

It may be required to remove one or a mini-batch of data sample(s), say M vectors, from the profile; e.g. to make sure that the size of the profile does not exceed a predefined bound. These data vectors should be selected properly. The approach we use here to update the profile is to first remove the effect of these vectors in matrixes C , D , U , and Ψ . Then we may choose, if we want, to zero out the corresponding M columns in X , W , and Φ ; and the corresponding M columns and rows in matrixes K and Λ , or to remove them.

We assume that the indices of M data samples that are selected to be removed are collected in vector \mathbf{m} . Defining $\boldsymbol{\lambda}_m = \boldsymbol{\lambda}_i(\mathbf{m})$, we start by setting the corresponding M elements in $\hat{\boldsymbol{\lambda}}$ to 0, i.e. $\hat{\boldsymbol{\lambda}}(\mathbf{m}) = \mathbf{0}$, and also $\hat{\Lambda}(\mathbf{m}, \mathbf{m}) = \mathbf{0}$; the other elements of $\hat{\boldsymbol{\lambda}}$ and $\hat{\Lambda}$ are as in $\boldsymbol{\lambda}_i$ and Λ_i , respectively. Note that $\boldsymbol{\lambda}_m \in \mathbb{R}^M$ and $\hat{\Lambda}(\mathbf{m}, \mathbf{m}) \in \mathbb{R}^{M \times M}$. At first, nothing is done with X_i and W_i (and Φ_i), the effect of columns \mathbf{m} is canceled by this change in $\hat{\Lambda}$. Referring to (16), setting $\hat{\boldsymbol{\lambda}}_m$ to zero thus gives

$$\hat{C}^{-1} = C_i^{-1} - \mathbf{w}_m \boldsymbol{\lambda}_m \mathbf{w}_m^T, \quad (35)$$

where $\mathbf{w}_m = W_i(:, \mathbf{m}) \in \mathbb{R}^{Q \times M}$ contains the M coefficient vectors that should be removed. The special definition for the regularization factor (24), makes it possible to derive this recursive update equation.

Assuming that \hat{C}^{-1} is (still) an invertible matrix, using MIL (26) with $A = C_i^{-1}$, $F = -I_M$ and $E = G^T = \mathbf{w}_m \boldsymbol{\lambda}_m^{0.5}$ results in (refer to appendix A.4):

$$\hat{C} = C_i + \mathbf{u}_m \alpha_m \mathbf{u}_m^T, \quad (36)$$

in which

$$\mathbf{u}_m = C_i \mathbf{w}_m \quad \text{and} \quad \alpha_m = (\boldsymbol{\lambda}_m^{-1} - \mathbf{w}_m^T \mathbf{u}_m)^{-1}. \quad (37)$$

The new U matrix will be

$$\hat{U} = \hat{C} W_i \hat{\Lambda} = C_i W_i \hat{\Lambda} + \mathbf{u}_m \alpha_m \mathbf{u}_m^T W_i \hat{\Lambda} = C_i W_i \hat{\Lambda} + \mathbf{u}_m \alpha_m \mathbf{v}_m^T, \quad (38)$$

where $\mathbf{v}_m^T = \mathbf{u}_m^T W_i \hat{\Lambda}$ or if we prefer $\mathbf{v}_m = \hat{\Lambda} W_i^T \mathbf{u}_m$. Note that \mathbf{v}_m is defined similar to equation (31), but elements \mathbf{m} are set to zero and $C_i W_i \hat{\Lambda}$ is defined similar to U_i in (17), with columns \mathbf{m} equal to zero; thus \hat{U} equals $U_i + \mathbf{u}_m \alpha_m \mathbf{v}_m^T$, with columns \mathbf{m} set to zero. Equation (38) can be rewritten as

$$\hat{U} = U_i - U_{\downarrow m} + \mathbf{u}_m \alpha_m \mathbf{v}_m^T, \quad (39)$$

where $U_{\downarrow m} \in \mathbb{R}^{Q \times L}$ is a matrix for which M columns whose indices are specified by \mathbf{m} are the corresponding columns on U_i and all other elements are zero. Note that columns \mathbf{m} in \hat{U} are all-zero.

With this notation (refer to appendix A.5),

$$\hat{D} = D_i - [\phi_m \boldsymbol{\lambda}_m - \Phi_i \mathbf{v}_m \alpha_m] \mathbf{u}_m^T. \quad (40)$$

Since $\hat{\Lambda} = \Lambda_i - \Lambda_{\downarrow m}$, where $\Lambda_{\downarrow m}$ is defined similar to $U_{\downarrow m}$ (refer to appendix A.6),

$$\Phi_i \mathbf{v}_m = \hat{\phi}_m - \phi_m + \phi_m \boldsymbol{\lambda}_m \alpha_m^{-1}. \quad (41)$$

We have used the fact that $\boldsymbol{\lambda}_m \mathbf{u}_m^T \mathbf{w}_m = I_M - \boldsymbol{\lambda}_m \alpha_m^{-1}$. Inserting $\Phi_i \mathbf{v}_m$ from (41) into (40) yields to (refer to appendix A.7):

$$\hat{D} = D_i - \mathbf{r}_m \alpha_m \mathbf{u}_m^T. \quad (42)$$

One can compare (42) with (32).

The new Ψ can also be updated as (refer to appendix A.8):

$$\begin{aligned} \hat{\Psi} &= \Psi_i - (\mathbf{u}_m \hat{\mathbf{u}}_m^T + \hat{\mathbf{u}}_m \mathbf{u}_m^T) \\ &\quad + \mathbf{u}_m [\boldsymbol{\lambda}_m \sigma_m^2 \boldsymbol{\lambda}_m - \boldsymbol{\lambda}_m \mathbf{k}_m^T \mathbf{v}_m \alpha_m - \alpha_m \mathbf{v}_m^T \mathbf{k}_m \boldsymbol{\lambda}_m \\ &\quad + \alpha_m \mathbf{v}_m^T K_i \mathbf{v}_m \alpha_m] \mathbf{u}_m^T, \end{aligned} \quad (43)$$

where $\hat{\mathbf{u}}_m \triangleq U_i [\mathbf{k}_m \boldsymbol{\lambda}_m - K_i \mathbf{v}_m \alpha_m] \in \mathbb{R}^{Q \times M}$. The same result can be achieved using (42) and the fact that $\hat{\Psi}_i = \hat{D}^T \hat{D}$.

Finally, the size of the involved matrixes X_i , W_i , Λ_i , Φ_i and K_i can be reduced by discarding the columns \mathbf{m} , or the rows \mathbf{m} . This prevents the size of these matrixes from growing unboundedly. Again, for special case of $M = 1$, update equations involve only matrix-by-vector multiplications and the complexity is of order $O(L^2)$.

3.3 Profile abstraction

In the context of KSR, there are two complementary concepts of dictionary sparsification. In kernel methods, the size of the network over which the signal is expanded, or the number of past samples, increases with the size of the data. This phenomenon is also clear from the algorithm derived in section 2, where the size of the matrixes X , W , K , and U grow linearly with time. In other words, at each time instant, the algorithm needs to store all previous data samples. In the context of kernel methods, this set of stored data is usually called “*dictionary*”. By contrast, to distinguish it from the “*representation dictionary*” used in the context of sparse representation,

we use the term “*feature space profile*”, or “*profile*” to refer to these data samples along with all corresponding matrixes and vectors that collectively represent the representation dictionary. In practice, redundancy among the input data makes it possible to reduce the size of this profile, at the cost of a negligible effect on the performance of the model. In the context of kernel methods, this procedure is termed “*sparsification*”, but we call this procedure “*profile abstraction*”, or “*abstraction*”. Abstraction is essentially an active data selection problem in which the aim is to select the most informative or the most salient data points. This can be accomplished based on two different approaches, namely “*growing*” and “*pruning*”.

3.3.1 Profile Growing

In the growing procedure, the algorithm starts with an empty, or an initially filled, profile and gradually adds important data samples into the profile, as iterations go on and new learning vectors arrive. This is generally carried out greedily by selecting, based on a measure, the most informative data and discarding the others. Note that, when any sample data is included in the profile, the definition of most of the previously defined variables should be revised accordingly. The update equations for this step were derived in subsection 3.1. Furthermore, the set Ω_i used in defining the ξ value in (24) is the set of forgetting factors that correspond to all input data samples that have been selected in this step.

Coherency and *representation error* are two criteria that have been widely used, in the context of kernel learning, for profile abstraction [40]. However, we use a closely linked criterion known as the *projection* measure [30], based on which the new data sample \mathbf{x} is added to the profile if the projection of ϕ into the space spanned by the columns of Φ_i is small enough. More specifically, the new data is accepted to be informative if:

$$\|\Phi_i(\Phi_i^T\Phi_i)^{-1}\Phi_i^T\phi\|_F^2 < \delta\|\phi\|_F^2 \quad (44)$$

or

$$\mathbf{k}^TK_i^{-1}\mathbf{k} < \delta\sigma_x^2, \quad (45)$$

where δ is a pre-set threshold.

3.3.2 Profile Pruning

The pruning procedure is based on excluding less important elements from the profile, to restrict the size of the profile and hence harness the complexity

of the algorithm. This can be done by removing a properly-selected mini-batch of M data samples, when the size of the profile exceeds the pre-set threshold L_{max} so as the size of profile reduces to $L_{max} - M$. Again, the indices of the data sample that should be removed are stored in vector \mathbf{m} . In our problem, removing the stored input vectors necessitates updating some matrixes, as described in subsection 3.2. However, pruning does not affect the set Ω_i ; as this set corresponds to the input samples that have been included in the profile. This special definition is necessary for deriving (35).

One idea for selecting the least important input sample is to discard data samples in the order of their *entrance*. Specifically, the effect of the first M data samples, that are the oldest input vectors in X_i , can be removed. A more reasonable idea is to discard data samples with the least *contribution* in representation. In the proposed approach, this contribution should be calculated in the feature space. At each time instant, the input data samples mapped into the feature space, i.e. $\hat{\Phi}_i$, is approximated based on the profile dictionary D_i using the coefficient matrix W_i as:

$$\hat{\Phi}_i = D_i W_i = (\Phi_i U_i^T) W_i. \quad (46)$$

Defining $B_i = U_i^T W_i \in \mathbb{R}^{L \times L}$, this gives:

$$\hat{\Phi}_i = D_i W_i = \Phi_i B_i. \quad (47)$$

Column j of B_i thus gives the coefficients for the linear combination of the mapped data samples within the profile, i.e. the columns in Φ_i , as used in the approximation of the column j in $\hat{\Phi}_i$. In other words, each row of B_i represents the contribution weights of the corresponding data sample in representation of all retained data samples. Based on this criterion, denoting the j th row of the matrix B_i by \check{b}_j , the indexes of the candidate data samples is found as those with the minimum value of $\|\check{b}_j\|_2$. But, to consider also the order of the entrance, we decided to select the M data samples based on the proposed contribution criterion, only from the first half of the data samples within the profile.

However, when a data sample is discarded from the profile, zeroing out the corresponding coefficient vector of W_{i+1} may render $(\boldsymbol{\lambda}_m^{-1} - \mathbf{w}_m^T \mathbf{u}_m)$ singular. Note that, unlike the α parameter defined in (28), the denominator of α_m in (37) may be singular. In this situation, the matrix $W_{i+1} \hat{\Lambda} W_{i+1}^T$ is singular and hence it does not have an inverse. Moreover, as the coefficient vectors are generally sparse, removing \mathbf{w} may cause some rows of the matrix W_{i+1} to be all-zero and hence render the matrix $W_{i+1} \hat{\Lambda} W_{i+1}^T$ non-invertible. Therefore, to make certain that removing the coefficients vectors does not cause the W_{i+1} matrix to become singular, two conditions must be checked, before removing:

$(\boldsymbol{\lambda}_m^{-1} - \mathbf{w}_m^T \mathbf{u}_m)$ is non-singular and all rows of W_{i+1} at least have a non-zero element. In other words, we decide to select the first M coefficient vectors, based on the proposed contribution criterion, for which these two conditions are satisfied.

3.4 Dictionary Normalization

In some situations, it is preferable to normalize the dictionary so that all atoms have unit ℓ_2 -norm. This restrains the approximation from trivial solutions. However, the profile update steps do not preserve the norm of the dictionary columns and hence it may be necessary to re-normalize the dictionary, even if periodically. The virtual dictionary D can be simply normalized by rescaling the Gram matrix Ψ . Defining a diagonal matrix S whose j th diagonal element is the square root of the j th diagonal element of the matrix Ψ , this rescaling can be formulated as:

$$\Psi_n = S^{-1} \Psi S^{-1}, \quad (48)$$

where S^{-1} is the inverse of S . This normalization necessitates rescaling some other matrixes, in particular W , C , and U matrixes must be rescaled as:

$$W_n = SW, \quad C_n = S^{-1}CS^{-1} \quad \text{and} \quad U_n = S^{-1}U. \quad (49)$$

3.5 Final Algorithm

In the proposed algorithm, the dictionary is represented by the profile that is the collection of X_i , K_i , W_i , C_i , U_i , and Ψ_i matrixes and $\boldsymbol{\lambda}_i$ vector. The algorithm starts with initializing the profile. Assuming that at start $L = Q$, the columns of X_i can be filled randomly or, if available, with Q data samples and then K_i is calculated using the kernel function. By initializing X_i in this way, W_i , C_i , and U_i are initialized with $Q \times Q$ identity matrixes and $\boldsymbol{\lambda}_i$ is formed with the Q first values of the λ parameter, that we assume to be all 1. After initializing, the profile is sequentially updated at each iteration, according to the remaining steps summarized in Algorithm 1.

Algorithm 1: KRLS DL Algorithm

Input: Training data samples $\mathbf{x} \in \mathbb{R}^N$, that are sequentially fed into the algorithm, in mini-batches of M vectors

Output: The updated profile, represented by $X_i, K_i, W_i, C_i, U_i, \Psi_i$, and λ_i

- 0 **Initialization:** Initialize the profile, as described in subsection 3.5.
While new data is available, iterate over the next four steps.
 - 1 **New Data:** Get the new M data samples \mathbf{x} and check for the informativeness of each data sample independently, e.g. using (45).
If the informativeness of at least one data sample is satisfied, discard the non-informative vectors and go to the next step; otherwise, go to Step 1 on the next iteration.
 - 2 **SA:** Calculate \mathbf{w} by SA, e.g. using the KORMP algorithm.
 - 3 **Profile Growing:** Grow the profile, using equations (22), (23), (27), (29), (30), and (33).
 - 4 **Profile Pruning:** If the profile size is greater than L_{max} , prune the profile, using equations (36), (38), and (43), and then discard the columns \mathbf{m} of X_i, W_i, K_i , and \hat{U} , the rows \mathbf{m} of K_i , as well as the elements \mathbf{m} of λ_i .
 - 5 **Normalization:** Normalize the dictionary based on the procedure described in subsection 3.4, if necessary, but normally not.
-

It is worth noting that the proposed algorithm still involves matrix inversion in equations (28) and (37). But now, the matrixes that must be inverted are of size $M \times M$, which can be selected very small. More importantly, for the case $M = 1$, which is quite a usual case in practice, matrix inversion is replaced by division. With regard to the profile abstraction, our tests showed that by just checking the informativeness of the new data vectors, based on the growing procedure, the size of the profile may grow well beyond a reasonable bound. Therefore, it may be necessary to implement this algorithm in a fixed-budget fashion, i.e. restricting the size of the profile to a predefined limit, say L_{max} , based on the pruning procedure. When a new batch of input vectors are included in the profile or old vectors are excluded from the profile, corresponding matrixes are updated based on the formulas derived in subsections 3.1 and 3.2, respectively. Moreover, since in practice the dictionary changes slowly over the successive iterations, one can skip the last step, i.e. normalization, for most iterations and employ it periodically. In the simulation tests reported in the next section, we decide to apply this step every time the profile is pruned.

4 Simulation Tests

To evaluate the performance of the proposed online dictionary learning algorithm, some numerical tests were conducted for the KSR-based classification task in two different applications. This evaluation was carried out in comparison to existing online KDL algorithms. To the best of our knowledge, hitherto three online KDL algorithms have been proposed for general kernels [30], [37] and [38], whose algorithms are denoted as sKDL, OKDLRS and OKDLFB, respectively. Furthermore, the batch KMOD algorithm proposed in [21] is considered as the benchmark. The main results of these tests are reported in this section. It is noted that batch DL algorithms are not alternatives to online DL algorithms. As explained in section 1, the computational cost and storage cost of batch algorithms increase prohibitively, as size of the training set increases. This can also be seen from the results reported later in Table 2 for moderately sized datasets used in the experiments.

4.1 Datasets

The simulation tests were conducted on the following datasets:

USPS: United States Postal Service (USPS) digit recognition dataset contains 16×16 -pixel images of handwritten digits from 0 to 9 in 8-bit grayscale, created by scanning digits on the envelopes. Therefore, a ten-class classification task is defined for digit recognition. In the version used in the current study ¹, each class contains 1100 images and all images are first vectorized to 256-dimensional data samples, i.e. $N = 256$.

ISOLET: ISolated LETter (ISOLET) recognition dataset contains speech-specific features extracted from spoken letters. To generate this dataset, 150 subjects spoke the name of each letter of the English alphabet twice; that is there are 300 samples for each letter. 617 features, including spectral, contour, sonorant, pre-sonorant, and post-sonorant features, were extracted from each utterance; therefore in this case $N = 617$. All feature values are continuous, real-valued attributes scaled into the range -1.0 to 1.0 . This dataset ² is used for spoken letter recognition which is a 26-class classification task.

EEG: EEG in schizophrenia dataset [42] comprised 14 patients with paranoid schizophrenia and 14 healthy controls. Therefore, the aimed problem is a 2-class classification task. Data were acquired with the sampling frequency of $250Hz$ using the standard 10 – 20 EEG montage with 19 EEG

¹https://cs.nyu.edu/~roweis/data/usps_all.mat

²<https://datahub.io/machine-learning/isolet/r/isolet.csv>

channels. The signals of healthy and patient cases are of length 925 and 845 seconds, respectively, but only the first 169 seconds of signals, in both classes, were used in the current study. Each 19-channel signal is divided into non-overlapping segments of 1-second length, and each segment is regarded as an independent data sample. Hence, each class contain 2366 data samples. Four frequency features, known as delta ($0-4Hz$), theta ($4-8Hz$), alpha ($8-13Hz$), and beta ($13-31Hz$), were extracted from each channel. This procedure results in 4 features over each channel, i.e., totally $N = 76$ features are extracted from each data sample.

DistNet One feeder named “Persian Gulf” from the “Malayer 1” distribution substation of the “Hamedan power distribution system” in Iran, which is a critical feeder, was simulated to generate simulated data for fault detection. The simulation was done in Electromagnetic Transient Program (EMTP) software under various fault conditions and factors. In addition to the no-fault condition, different short-circuit faults, including Single-Line-to-Ground (SLG), Line-to-Line (LL), Double-Line-to-Ground (LLG), and Triple-Line (LLL), were considered. 400 cases were conducted for each condition, where for each case, the main factors of the grid and fault (including the fault location, the load value of each consumer, the fluctuation of the voltage of the main source, the inception angle, and the impedance of the fault) were assigned randomly, within reasonable ranges. One cycle (201 samples) of the three-phase voltage and current signals at the beginning of the feeder was recorded with the sampling frequency of $10KHz$. These signals were concatenated and are used for this study. In summary, generated data samples are of $N = 1206$ dimensions and are employed for a 5-class classification task.

4.2 Test Scenario

All evaluations are performed using 5-fold cross-validation. One dictionary is trained for each class and each test sample is assigned to the class whose dictionary results in the minimum representation error. For online DL learning algorithms, training samples, i.e. all samples within other 4 groups over each fold, are exposed to the algorithm in mini-batches of M samples. By receiving a mini-batch of data from each class, the corresponding profile is updated. The learning procedure is proceeding for a predefined number of batches. If the total number of available batches of training samples is less than that number, the learning procedure sequentially iterates over the previously seen data vectors, for some epochs. As the profiles are sequentially trained based on the training data samples, the updated dictionaries are intermittently used to classify all test samples, i.e. the samples in the hold-out

group. The total number of tests is set to 20. Ideally, by increasing the number of training batches and exposing more data samples, the classification accuracy should increase and reach the accuracy obtained by the batch KMOD DL algorithm.

For all tests reported herein, a quadratic polynomial kernel is used. The forgetting factor λ linearly increases from 0.98 to 1 for the first 80% of batches and is then kept constant for the remaining batches. Moreover, the number of dictionary atoms and the maximum size of the profile are set to $Q = 30$ and $L_{max} = 200$, respectively. The batch size for both growing and pruning steps were set to 10, and γ regularization parameter is set to 0.1. To provide a fair comparison among different DL algorithms, the test samples were always approximated using the KORMP algorithm [7] with sparsity level $s = 5$, regardless of the approximation algorithm proposed in the DL algorithm. Each algorithm is implemented with its own parameter values that result in higher accuracy. As an indication of the computational complexity of the algorithms, the average training time over all mini-batches is also reported. It should be noted that for the online DL algorithms the training time depends on the profile size that gradually increases to L_{max} , as mini-batches arrive. All simulation tests were conducted using MATLAB 2019b in 64bit Windows operating system on a 3.2GHz Intel CPU and 16GB memory while only one core is enabled ³.

4.3 Results

Classification accuracies averaged over $k = 5$ folds as a function of the number of mini-batches for four datasets are shown in Figure 1, where the performance of the proposed algorithm is compared with that of all other existing KDL algorithms. The averaged accuracy achieved by the batch KMOD algorithm is also reported; this can be considered as a benchmark. Furthermore, the average training time of each class per each mini-batch for the online kernel algorithms along with the averaged execution time of each class per each iteration of the batch algorithm, are tabulated in Table 2. To provide a more detailed analysis of the computational cost for the proposed algorithm, the average time of two main tasks of the training procedure is reported in this table. The growing task represents steps 2 and 3 of the algorithm, i.e., SA and profile growing, while the pruning task accounts for steps 4 and 5; i.e., profile pruning and normalization. The total average training time of the KRLS algorithm hence equals the sum of these two values.

These results clearly show the outperformance of the proposed algorithm.

³Simulation codes are available at https://github.com/G-Alipoor/KRLS_DLA.

In particular, during the training progress, the proposed algorithm is always of higher classification accuracy. Moreover, the accuracy achieved by the proposed algorithm almost monotonically increases toward that of the batch algorithm; this observation shows the high scalability of the KRLS DL algorithm. This improvement is achieved despite a smaller execution time. The decrease in the training time is an indication of the smaller computational complexity of the KRLS DL algorithm that is in turn a direct effect of the efficient derivation of the algorithm.

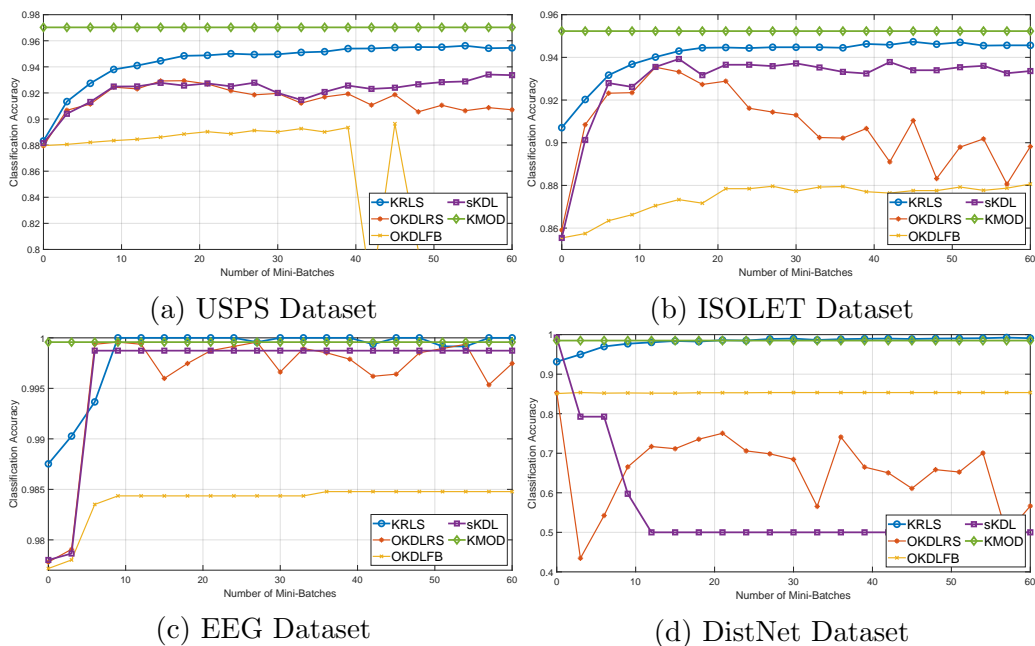


Figure 1: The averaged classification accuracy of DL algorithms

Table 2: Average training time (milliseconds)

Algorithm	Dataset	OKDLRS	OKDLFB	sKDL	KMOD	KRLS	
						Growing	Pruning
	USPS	2856.6	141.5	87.2	4651.3	2.0	2.3
	ISOLET	2868.3	117.5	37.3	114.1	2.3	2.5
	EEG	2910.5	5.1	0.7	69860	1.8	2.4
	DistNet	354.3	55.2	.5	174.9	5.4	5.1

In another test the robustness of the dictionaries trained with the proposed algorithm is evaluated. For this aim, a randomly selected portion of the data points of all test samples are removed by replacing their values with

0. It should be noted that this destruction is only applied to the test data samples, that is dictionaries are trained with the un-changed data samples. The averaged classification accuracies for different percentages of missing data points are summarized in figure 2. These results show the robustness of the dictionaries trained with the KRLS DL algorithm against noise, even in severe cases when most data points are removed.

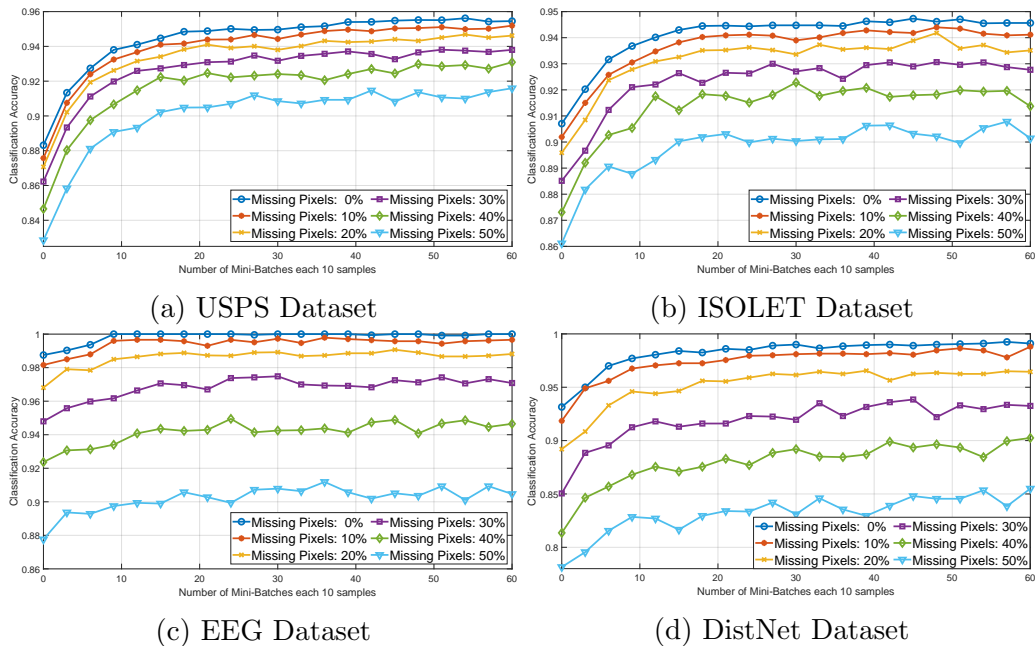


Figure 2: Effect of missing pixels on the classification accuracy of the KRLS DL algorithm

5 Conclusion

Online dictionary learning algorithms have appealing properties in many applications where it is not practically possible to utilize batch algorithms. An online kernel dictionary learning algorithm was developed in this work whose objective is to sequentially optimize a dictionary in a feature space induced by a Mercer kernel that minimizes a weighted recursive least squares function of the sparse representation error in the same feature space. This recursive solution was derived using the matrix inversion lemma. The proposed algorithm is an online version of the KMOD batch dictionary learning algorithm and at the same time can be viewed as an extension of the RLS dictionary learning algorithm to kernel space. In spite of the possibly apparent similar-

ity, this work addressed an absolutely different problem with its own concepts and challenges. Moreover, the derived algorithm is the first LS-based kernel dictionary learning algorithm that inherits the good property of a higher convergence rate from the LS approach.

An efficient implementation of the solution is also derived which makes the implementation only-matrix-by-vector for batch size of 1 sample and thus the final algorithm is of considerably lower computational complexity. Experimental tests conducted on [four practical datasets in quite different applications, i.e., image recognition, sound classification, disorder diagnosis in medical science, and fault detection in power distribution networks](#), showed the outperformance of the proposed algorithm, in terms of both classification accuracy and computational burden. Overall, we think that the proposed algorithm is a considerable progress in the field of online kernel dictionary learning algorithms.

A Appendix: Some Mathematical Derivations

Detailed derivations for equations (30), (32), (33), (36), and (40)-(43) are provided in this appendix.

A.1 Equation (30)

$$\begin{aligned}
U_{i+1} &= C_{i+1}W_{i+1}\Lambda_{i+1} \\
&= \lambda^{-1} (C_i - \mathbf{u}\alpha\mathbf{u}^T)[W_i \lambda\Lambda_i, \mathbf{w}] \\
&= \lambda^{-1} [C_iW_i\lambda\Lambda_i - \mathbf{u}\alpha\mathbf{u}^TW_i\lambda\Lambda_i, C_i\mathbf{w} - \mathbf{u}\alpha\mathbf{u}^T\mathbf{w}] \\
&= \lambda^{-1} [\lambda U_i - \mathbf{u}\alpha\mathbf{u}^TW_i\lambda\Lambda_i, \mathbf{u} - \mathbf{u}(I_M - \lambda\alpha)] \\
U_{i+1} &= \begin{bmatrix} U_i - \mathbf{u}\alpha\mathbf{v}^T, \mathbf{u}\alpha \end{bmatrix},
\end{aligned}$$

A.2 Equation (32)

$$\begin{aligned}
D_{i+1} &= \Phi_{i+1}U_{i+1}^T = [\Phi_i, \phi] \begin{bmatrix} U_i^T - \mathbf{v}\alpha\mathbf{u}^T \\ \alpha\mathbf{u}^T \end{bmatrix} \\
&= \Phi_i U_i^T - \Phi_i \mathbf{v}\alpha\mathbf{u}^T + \phi\alpha\mathbf{u}^T = D_i + (\phi - \Phi_i \mathbf{v})(\alpha\mathbf{u}^T) \\
&= D_i + \mathbf{r}\alpha\mathbf{u}^T,
\end{aligned}$$

A.3 Equation (33)

$$\begin{aligned}
\Psi_{i+1} &= D_{i+1}^T D_{i+1} = U_{i+1} K_{i+1} U_{i+1}^T \\
&= [U_i - \mathbf{u}\alpha\mathbf{v}^T, \mathbf{u}\alpha] \begin{bmatrix} K_i & \mathbf{k} \\ \mathbf{k}^T & \sigma_x^2 \end{bmatrix} \begin{bmatrix} U_i^T - \mathbf{v}\alpha\mathbf{u}^T \\ \alpha\mathbf{u}^T \end{bmatrix} \\
&= (U_i - \mathbf{u}\alpha\mathbf{v}^T)K_i(U_i^T - \mathbf{v}\alpha\mathbf{u}^T) \\
&\quad + (U_i - \mathbf{u}\alpha\mathbf{v}^T)\mathbf{k}\alpha\mathbf{u}^T + \mathbf{u}\alpha\mathbf{k}^T(U_i^T - \mathbf{v}\alpha\mathbf{u}^T) + \mathbf{u}\alpha\sigma_x^2\alpha\mathbf{u}^T \\
&= U_i K_i U_i^T - \mathbf{u}\alpha\mathbf{v}^T K_i U_i^T - U_i K_i \mathbf{v}\alpha\mathbf{u}^T + \mathbf{u}\alpha\mathbf{v}^T K_i \mathbf{v}\alpha\mathbf{u}^T \\
&\quad + U_i \mathbf{k}\alpha\mathbf{u}^T - \mathbf{u}\alpha\mathbf{v}^T \mathbf{k}\alpha\mathbf{u}^T + \mathbf{u}\alpha\mathbf{k}^T U_i^T - \mathbf{u}\alpha\mathbf{k}^T \mathbf{v}\alpha\mathbf{u}^T + \mathbf{u}\alpha\sigma_x^2\alpha\mathbf{u}^T \\
&= \Psi_i + \mathbf{u}\alpha(\mathbf{k}^T - \mathbf{v}^T K_i)U_i^T + U_i(\mathbf{k} - K_i \mathbf{v})\alpha\mathbf{u}^T \\
&\quad + \mathbf{u}\alpha(\mathbf{v}^T K_i \mathbf{v} - \mathbf{v}^T \mathbf{k} - \mathbf{k}^T \mathbf{v} + \sigma_x^2)\alpha\mathbf{u}^T \\
\Psi_{i+1} &= \Psi_i + \mathbf{u}\alpha\tilde{\mathbf{u}}^T + \tilde{\mathbf{u}}\alpha\mathbf{u}^T + \mathbf{u}\alpha(\mathbf{v}^T K_i \mathbf{v} - \mathbf{v}^T \mathbf{k} - \mathbf{k}^T \mathbf{v} + \sigma_x^2)\alpha\mathbf{u}^T,
\end{aligned}$$

A.4 Equation (36)

$$\begin{aligned}
\hat{C} &= C_i - C_i \mathbf{w}_m \boldsymbol{\lambda}_m^{0.5} \left(-I_M + \boldsymbol{\lambda}_m^{0.5} \mathbf{w}_m^T C_i \mathbf{w}_m \boldsymbol{\lambda}_m^{0.5} \right)^{-1} \boldsymbol{\lambda}_m^{0.5} \mathbf{w}_m^T C_i \\
&= C_i + C_i \mathbf{w}_m \left(\boldsymbol{\lambda}_m^{-1} - \mathbf{w}_m^T C_i \mathbf{w}_m \right)^{-1} \mathbf{w}_m^T C_i \\
\hat{C} &= C_i + \mathbf{u}_m \alpha_m \mathbf{u}_m^T,
\end{aligned}$$

A.5 Equation (40)

$$\begin{aligned}
\hat{D} = \Phi_i \hat{U}^T &= \Phi_i [U_i^T - U_{\downarrow m}^T + \mathbf{v}_m \alpha_m \mathbf{u}_m^T] \\
&= \Phi_i U_i^T - \Phi_i U_{\downarrow m}^T + \Phi_i \mathbf{v}_m \alpha_m \mathbf{u}_m^T \\
&= D_i - [\phi_m \boldsymbol{\lambda}_m - \Phi_i \mathbf{v}_m \alpha_m] \mathbf{u}_m^T.
\end{aligned}$$

A.6 Equation (41)

$$\begin{aligned}
\Phi_i \mathbf{v}_m = \Phi_i \hat{\Lambda} W_i^T \mathbf{u}_m &= \Phi_i \hat{\Lambda} W_i^T C_i \mathbf{w}_m \\
&= \Phi_i \Lambda_i W_i^T C_i \mathbf{w}_m - \Phi_i \Lambda_{\downarrow m} W_i^T C_i \mathbf{w}_m \\
&= \hat{\phi}_m - \phi_m (\boldsymbol{\lambda}_m \mathbf{u}_m^T \mathbf{w}_m) \\
&= \hat{\phi}_m - \phi_m + \phi_m \boldsymbol{\lambda}_m \alpha_m^{-1}.
\end{aligned}$$

A.7 Equation (42)

$$\begin{aligned}
\hat{D} &= D_i - [\phi_m \boldsymbol{\lambda}_m - \hat{\phi}_m \alpha_m + \phi_m \alpha_m - \phi_m \boldsymbol{\lambda}_m] \mathbf{u}_m^T \\
&= D_i - [\phi_m - \hat{\phi}_m] \alpha_m \mathbf{u}_m^T \\
&= D_i - \mathbf{r}_m \alpha_m \mathbf{u}_m^T.
\end{aligned}$$

A.8 Equation (43)

$$\begin{aligned}
\hat{\Psi} = \hat{U}K_i\hat{U}^T &= [U_i - U_{\downarrow m} + \mathbf{u}_m\alpha_m\mathbf{v}_m^T]K_i[U_i^T - U_{\downarrow m}^T + \mathbf{v}_m\alpha_m\mathbf{u}_m^T] \\
&= U_iK_iU_i^T - U_iK_iU_{\downarrow m}^T + U_iK_i\mathbf{v}_m\alpha_m\mathbf{u}_m^T - U_{\downarrow m}K_iU_i^T \\
&\quad + U_{\downarrow m}K_iU_{\downarrow m}^T - U_{\downarrow m}K_i\mathbf{v}_m\alpha_m\mathbf{u}_m^T + \mathbf{u}_m\alpha_m\mathbf{v}_m^TK_iU_i^T \\
&\quad - \mathbf{u}_m\alpha_m\mathbf{v}_m^TK_iU_{\downarrow m}^T + \mathbf{u}_m\alpha_m\mathbf{v}_m^TK_i\mathbf{v}_m\alpha_m\mathbf{u}_m^T \\
&= \Psi_i - U_i\mathbf{k}_m\boldsymbol{\lambda}_m\mathbf{u}_m^T + U_iK_i\mathbf{v}_m\alpha_m\mathbf{u}_m^T - \mathbf{u}_m\boldsymbol{\lambda}_m\mathbf{k}_m^TU_i^T \\
&\quad + \mathbf{u}_m\boldsymbol{\lambda}_mK_i(m, m)\boldsymbol{\lambda}_m\mathbf{u}_m^T - \mathbf{u}_m\boldsymbol{\lambda}_m\mathbf{k}_m^T\mathbf{v}_m\alpha_m\mathbf{u}_m^T \\
&\quad + \mathbf{u}_m\alpha_m\mathbf{v}_m^TK_iU_i^T - \mathbf{u}_m\alpha_m\mathbf{v}_m^T\mathbf{k}_m\boldsymbol{\lambda}_m\mathbf{u}_m^T \\
&\quad + \mathbf{u}_m\alpha_m\mathbf{v}_m^TK_i\mathbf{v}_m\alpha_m\mathbf{u}_m^T \\
\hat{\Psi} &= \Psi_i - (\mathbf{u}_m\hat{\mathbf{u}}_m^T + \hat{\mathbf{u}}_m\mathbf{u}_m^T) \\
&\quad + \mathbf{u}_m[\boldsymbol{\lambda}_m\sigma_m^2\boldsymbol{\lambda}_m - \boldsymbol{\lambda}_m\mathbf{k}_m^T\mathbf{v}_m\alpha_m - \alpha_m\mathbf{v}_m^T\mathbf{k}_m\boldsymbol{\lambda}_m \\
&\quad + \alpha_m\mathbf{v}_m^TK_i\mathbf{v}_m\alpha_m]\mathbf{u}_m^T,
\end{aligned}$$

References

- [1] K. Huang and S. Aviyente, "Sparse representation for signal classification," *Advances in neural information processing systems*, vol. 19, pp. 609–616, 2006.
- [2] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.
- [3] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: algorithms and applications," *IEEE access*, vol. 3, pp. 490–530, 2015.
- [4] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.
- [5] 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.
- [6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

- [7] K. Skretting and J. H. Husøy, “Partial search vector selection for sparse signal representation,” in *Nordic Signal Processing Symposium (NORSIG)*, pp. 1–6, IEEE, 2008.
- [8] 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. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 5, pp. 2443–2446, IEEE, 1999.
- [9] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [10] J. Shawe-Taylor, N. Cristianini, *et al.*, *Kernel methods for pattern analysis*. Cambridge university press, 2004.
- [11] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” *The annals of statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.
- [12] N. Moradi and N. Mahdavi-Amiri, “Kernel sparse representation based model for skin lesions segmentation and classification,” *Computer methods and programs in biomedicine*, vol. 182, p. 105038, 2019.
- [13] J. Tong, K. Li, W. Lin, X. Shudong, A. Anwar, and L. Jiang, “Automatic lumen border detection in ivus images using dictionary learning and kernel sparse representation,” *Biomedical Signal Processing and Control*, vol. 66, p. 102489, 2021.
- [14] Z. Zhou, J. Li, Y. Quan, and R. Xu, “Image quality assessment using kernel sparse coding,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1592–1604, 2020.
- [15] G. Zhang, J. Yang, Y. Zheng, Z. Luo, and J. Zhang, “Optimal discriminative feature and dictionary learning for image set classification,” *Information Sciences*, vol. 547, pp. 498–513, 2021.
- [16] C. Fan, C. Hu, and B. Liu, “Linearized kernel dictionary learning with group sparse priors for action recognition,” *The Visual Computer*, vol. 35, no. 12, pp. 1797–1807, 2019.
- [17] W. Lei, Z. Ma, Y. Lin, and W. Gao, “Domain adaption based on source dictionary regularized rkhs subspace learning,” *Pattern Analysis and Applications*, pp. 1–20, 2021.

- [18] K. Gao, T. Wang, K. Peng, Z. Wang, Q. He, and R. Peng, “A recursive model of residual life prediction for human beings with health information from activities of daily living and memory,” *Systems Science & Control Engineering*, vol. 9, no. 1, pp. 529–541, 2021.
- [19] H. Li, Y. Gao, and J. Sun, “Fast kernel sparse representation,” in *2011 International Conference on Digital Image Computing: Techniques and Applications*, pp. 72–77, IEEE, 2011.
- [20] J. Yin, Z. Liu, Z. Jin, and W. Yang, “Kernel sparse representation based classification,” *Neurocomputing*, vol. 77, no. 1, pp. 120–128, 2012.
- [21] H. Van Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, “Design of non-linear kernel dictionaries for object recognition,” *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5123–5135, 2013.
- [22] L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, and F.-Z. Li, “Kernel sparse representation-based classifier,” *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1684–1695, 2011.
- [23] X.-T. Yuan, X. Liu, and S. Yan, “Visual classification with multitask joint sparse representation,” *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.
- [24] S. Gao, I. W.-H. Tsang, and L.-T. Chia, “Kernel sparse representation for image classification and face recognition,” in *European conference on computer vision*, pp. 1–14, Springer, 2010.
- [25] M. T. Harandi, C. Sanderson, R. Hartley, and B. C. Lovell, “Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach,” in *European Conference on Computer Vision*, pp. 216–229, Springer, 2012.
- [26] P. Li, Q. Wang, W. Zuo, and L. Zhang, “Log-Euclidean kernels for sparse representation and dictionary learning,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1601–1608, 2013.
- [27] Y. Wu, Y. Jia, P. Li, J. Zhang, and J. Yuan, “Manifold kernel sparse representation of symmetric positive-definite matrices and its applications,” *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3729–3741, 2015.
- [28] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, “Multiple kernel sparse representations for supervised and unsupervised learning,” *IEEE transactions on Image Processing*, vol. 23, no. 7, pp. 2905–2915, 2014.

- [29] Z. Chen, W. Zuo, Q. Hu, and L. Lin, “Kernel sparse representation for time series classification,” *Information Sciences*, vol. 292, pp. 15–26, 2015.
- [30] C. O’Brien and M. D. Plumbley, “Sparse kernel dictionary learning,” in *Proceedings of the 11th IMA International Conference on Mathematics in Signal Processing*, 2016.
- [31] V. Abrol, P. Sharma, and A. K. Sao, “Greedy dictionary learning for kernel sparse representation based classifier,” *Pattern Recognition Letters*, vol. 78, pp. 64–69, 2016.
- [32] A. Golts and M. Elad, “Linearized kernel dictionary learning,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 726–739, 2016.
- [33] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *Journal of Machine Learning Research*, vol. 11, no. 1, 2010.
- [34] K. Skretting and K. Engan, “Recursive least squares dictionary learning algorithm,” *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2121–2130, 2010.
- [35] Y. Naderahmadian, M. A. Tinati, and S. Beheshti, “Generalized adaptive weighted recursive least squares dictionary learning,” *Signal Processing*, vol. 118, pp. 89–96, 2016.
- [36] S. Zhang, S. Kasiviswanathan, P. C. Yuen, and M. Harandi, “Online dictionary learning on symmetric positive definite manifolds with vision applications,” in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [37] H. Liu and F. Sun, “Online kernel dictionary learning for object recognition,” in *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 268–273, IEEE, 2016.
- [38] J. Lee and S.-J. Kim, “Online kernel dictionary learning on a budget,” in *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 1535–1539, IEEE, 2016.
- [39] Y. Engel, S. Mannor, and R. Meir, “The kernel recursive least-squares algorithm,” *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

- [40] W. Liu, J. C. Principe, and S. S. Haykin, *Kernel adaptive filtering: a comprehensive introduction*, vol. 1. Wiley Online Library, 2010.
- [41] S. S. Haykin, *Adaptive Filter Theory*, vol. 1. Pearson, fifth ed., 2014.
- [42] E. Olejarczyk and W. Jernajczyk, “EEG in schizophrenia,” *RepOD*, 2006. <https://doi.org/10.18150/repod.0107441>.