![The University of Edinburgh logo]

Edinburgh Research Explorer

Dictionary Subselection Using an Overcomplete Joint Sparsity Model

OPEN ACCESS

Dictionary Subselection Using an Overcomplete Joint Sparsity Model

Mehrdad Yaghoobi[†], Laurent Daudet[‡], Michael E. Davies[†]

[†] Institute for Digital Communications (IDCom), The University of Edinburgh, EH9 3JL, UK

[‡] Paris Diderot University / IUF, Institut Langevin, 1, rue Jussieu 75005 Paris, France.

{m.yaghoobi-vaighan, mike.davies}@ed.ac.uk, laurent.daudet@espci.fr

Abstract—Many natural signals exhibit a sparse representation, whenever a suitable describing model is given. Here, a linear generative model is considered, where many sparsity-based signal processing techniques rely on such a simplified model. As this model is often unknown for many classes of the signals, we need to select such a model based on the domain knowledge or using some exemplar signals. This paper presents a new exemplar based approach for the linear model (called the *dictionary*) selection, for such sparse inverse problems. The problem of dictionary selection, which has also been called the *dictionary learning* in this setting, is first reformulated as a joint sparsity model. The joint sparsity model here differs from the standard joint sparsity model as it considers an overcompleteness in the representation of each signal, within the range of selected subspaces. The new dictionary selection paradigm is examined with some synthetic and realistic simulations.

I. INTRODUCTION

The sparse signal model is one the most successful low-dimensional signal models for modern signal processing applications [1]. In this model, any considered signal $\mathbf{y} \in \mathbb{R}^m$, can be represented as the sum of a few elementary functions, called the *atoms*, plus some noise $\mathbf{n} \in \mathbb{R}^m$, as follows,

$$\mathbf{y} = \mathbf{Dx} + \mathbf{n},$$

where $\mathbf{D} \in \mathbb{R}^{m \times p}$, called the *dictionary*, is the collection of the atoms and $\mathbf{x} \in \mathbb{R}^p$ is a sparse vector. In this setting, $\mathbf{y}$ is often called a sparse signal in $\mathbf{D}$. The additive noise is used to consider the inaccuracy of the measurement device or the model mismatch. While choosing an overcomplete dictionary, *i.e.* $p > m$, gives us a flexibility to choose sparser representation, the extra redundancy can be damaging in ducking failures coefficient recovery. Therefore, the success of sparse signal models depends on how well we choose a redundant $\mathbf{D}$, which is the main focus of this paper.

There is a lot of interest in building redundant dictionaries to make more flexible models and various techniques have already been proposed to design the dictionary using some domain knowledge, see for example [2], or learning the dictionary using a given set of exemplars [3], see [4] and [5] for a more complete review on different dictionary selection techniques. The advantage of the first approach is the possibility of incorporating already known signal structures and often fast implementation of the dictionary. The second approach does not need such prior information about the signals, but they often find an unstructured dictionary with a computationally expensive implementation. We will combine these two methods in this paper, by considering a large set of potentially good atoms $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, $n > p$, called *mother dictionary* here, and selecting a smaller set of atoms as the final dictionary $\mathbf{D}$. Fast implementation of such dictionaries are guaranteed, if the mother dictionary has such a property. For instance, scalar products of a given signal $\mathbf{x}$ with a family of Gabor atoms of length m can be implemented with a computational complexity of $\mathcal{O}(m \log m)$. Also, as we restrict the search space to the dictionaries with mother atoms, it can be learned using much less exemplars. In other words, restricting the dictionary to a subset of mother atoms, regularises the dictionary learning problem and reduces the necessary amount of training data.

As all the atoms of $\mathbf{D}$ exist in $\mathbf{\Phi}$, any sparse signal in $\mathbf{D}$, can be represented using $\mathbf{\Phi}$. The reader may ask, if we can use the large dictionary $\mathbf{\Phi}$, why we need to shrink it to find a dictionary which at best can only sparsify the signal to the same level. The answer to this question can be given by noting that, finding the sparse approximations have non-polynomial complexity, in a general setting. The success of practical sparse approximation algorithms depends on some internal structures of the dictionary, including mutual coherence [6], Restricted Isometry Property (RIP) [7] or the null-space property [8]. Dictionary size indirectly affects these properties such that larger dictionaries mostly make the sparse recovery more difficult. Roughly speaking, it is caused by the fact that by putting more atoms in the dictionary, the atoms become more similar. Such similarities between different atoms, indeed make it more challenging to find which set of atoms represents the signals more accurately, *i.e* the problem of exact (support) recovery. The approximation in such large dictionaries would also be noise sensitive, as small noise may cause the wrong atoms to become selected. Finally, in coding, the cost of indexing which atoms being used in the representation ($\mathbf{x}$), *a.k.a.* the binary significance map, grows by increasing the dictionary size.

A. Related Work

The problem of dictionary design by combining the atoms of a mother dictionary was considered in [9]–[11]. In this setting, an auxiliary sparse matrix combines the mother atoms, to generate a dictionary which fits the given learning samples. The size of dictionary is fixed here and as the learned dictionary is the multiplication of a sparse matrix and a structured matrix (with a possibly fast multiplication), we can implement such a dictionary in two steps, where each of them are cheaper than $\mathcal{O}(p^2)$. The dictionary selection problem can be interpreted as a particular case of sparse dictionary learning, when the sparse matrix can have only $p < n$ non-zero elements, with one non-zero on each row.

The problem of learning a dictionary, when the size of dictionary is not given, has been investigated in [12]. The dictionary selection problem has also a similar approach, by finding smaller size dictionaries from the given larger reference dictionaries. The difference is that the reference dictionary is fixed throughout the learning here, which allows us to handle significantly larger problems and find some computationally fast dictionaries.

The dictionary selection, which will be considered in this paper, is also related to the problem of subset selection in machine learning [13], [14], where the goal is to select the most relevant subset, which describes the whole set. [13] uses the fact that such a model selection can be formulated as a submodular cost minimisation. For such a

formulation, there exist some canonical solvers, which guarantee to find a neighbourhood solution. The derived neighbourhood is indeed not small, which motivated Das and Kempe [14] to present an alternative submodular formulation to reduce the approximation error.

B. Contributions

We here choose a different path to the mentioned dictionary selection techniques in previous section, by reformulating the problem as a generalised form of joint sparse representation problem [15], [16]. To the authors' knowledge, it is the first time that the dictionary selection problem is modelled in this way. In this model, representation of each signal is not only p-joint sparse, it is also k-sparse in the selected joint sparsity support. We here assume $p > m$, which makes the representation of each signal in the selected p-joint support, non-unique, where k-sparsity constraint can help to find the correct representation.

Based on the new signal model, we need to solve a quadratic objective. As the signal model and the objective include unbounded solutions, we need to investigate the conditions that the problem is *well-defined*. Such an analysis is useful for the convergence study of any algorithm for solving the problem. The boundedness and uniqueness of the solutions of the introduced optimisation problem are also investigated in this paper.

As the dictionary can be found using the active rows of the coefficient matrix of the introduced optimisation program, we need to practically solve a non-polynomial time complexity problem. We here introduce a technique, which is inspired by the Iterative Hard Thresholding (IHT) for sparse approximations [17], [18], to find such an active set of atoms. IHT is an iterative method that at each iteration, thresholds the coefficient vector, after updating in the gradient direction. The algorithm is equipped with a line-search technique to guarantee the monotonic decrease of the (positive) objective. With some numerical experiments, the new approach is shown to recover the exact dictionary, in a large range of sparsity/overcompleteness parameters.

C. Paper Organisation

We initially formulate the dictionary selection problem as an overcomplete joint sparse representation problem in Section II. We then introduce an iterative algorithm to solve the problem approximately in Section III and show some dictionary recovery results with the synthetic data simulation in Section IV. We also show some simulation results on the Curvelet and harmonic based sub-dictionary selections, respectively for the finger print and audio data in this section. The paper will be concluded in Section V.

II. MATHEMATICAL MODELING

Let $\mathbf{Y} = [\mathbf{y}_l]_{l \in [1, L]}$ be a matrix made by training samples $\mathbf{y}_l \in \mathbb{R}^m$ and $\mathbf{\Phi} = [\boldsymbol{\phi}_i]_{i \in \mathcal{I}}, |\mathcal{I}| = n$ be a mother dictionary of normalised atoms $\boldsymbol{\phi}_i \in \mathbb{R}^m$. We assume that the generative dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$, $m \le p$ is made by a subset selection of atoms in $\mathbf{\Phi}$, *i.e.* $\mathbf{D} = [\boldsymbol{\phi}_i]_{i \in \mathcal{J}}$ where $\mathcal{J} \subset \mathcal{I}$ and $|\mathcal{J}| = p < n$. We assume that each $\mathbf{y}_l$ is *approximately* generated by a k-sparse coefficient vector $\boldsymbol{\gamma}_l$,

$$\mathbf{y}_l \approx \mathbf{D}\boldsymbol{\gamma}_l,$$

with respect to the Euclidean metric, *i.e.* $\|\mathbf{y}_l - \mathbf{D}\boldsymbol{\gamma}_l\|_2 / \|\mathbf{y}_l\|_2 \ll 1$. We want to find a dictionary that fulfils the two (apparently contradictory) objectives: few elements in the dictionary, and sparsest decomposition for each signal. In other words $\mathbf{D}$ which is both *small* and *efficient*! The problem of *dictionary subselection* can

thus be defined as finding the index set $\mathcal{J}$ meeting those criteria, given $\mathbf{Y}$, $\mathbf{\Phi}$, p and k. Let $\mathbf{X} \in \mathbb{R}^{n \times L}$ be a coefficient matrix and $f_{\mathcal{J}}(i) : [1, p] \mapsto [1, n]$ be the mapping that assigns the corresponding atom index of $\mathbf{\Phi}$ to the ith component of $\boldsymbol{\gamma}_l$. By assigning $\{\mathbf{x}_l\}_{f_{\mathcal{J}}(i)} \leftarrow \{\boldsymbol{\gamma}_l\}_i$, $\forall i \in [1, p]$, $\forall l \in [1, L]$, while the other elements of $\mathbf{X}$ are set to zero, the generative model can be reformulated as,

$$\mathbf{Y} \approx \mathbf{\Phi}\mathbf{X}, \tag{1}$$

with respect to the canonical Euclidean metric, *i.e.* $\|\mathbf{Y} - \mathbf{\Phi}\mathbf{X}\|_F / \|\mathbf{Y}\|_F \ll 1$. As $\mathbf{X}$ is k-sparse in each column and p-row-sparse, i.e. only p rows of $\mathbf{X}$ have non-zero components, it lies in the intersection of the following sets,

$$\mathcal{K} := \left\{ \mathbf{\Theta} \in \mathbb{R}^{n \times L} : \ \|\boldsymbol{\theta}_l\|_0 \le k, \forall l \in [1, L] \right\} \tag{2}$$

$$\mathcal{P} := \left\{ \mathbf{\Theta} \in \mathbb{R}^{n \times L} : \ \|\mathbf{\Theta}\|_{0, \infty} \le p \right\} \tag{3}$$

where θ_l is the lth column of $\mathbf{\Theta}$, $\|\mathbf{\Theta}\|_{0, \infty} = \|\boldsymbol{\nu}\|_0$, with $\{\boldsymbol{\nu}\}_i := \|\boldsymbol{\theta}^{(i)}\|_\infty$ and $\theta^{(i)}$ is the ith row of $\mathbf{\Theta}$. In other words, sets $\mathcal{K}$ and $\mathcal{P}$ are the sets of n by L matrices which respectively have k non-zero elements on each column and p non zero rows. The class of signals which can be represented with some coefficient matrices in $\mathcal{K} \cap \mathcal{P}$, is called here the (k, p)-(overcomplete) joint sparse signals. We actually combine the coefficient matrix and the dictionary parameters, *i.e.* the index set of optimal dictionary, in a single matrix $\mathbf{X}$, where the optimal atom indices are specified by the locations of non-zero rows of $\mathbf{X}$.

The optimal dictionary $\mathbf{D}$, which can alternatively be indicated by $\mathcal{J}$, is defined as the solution of the following problem,

$$\min_{\mathbf{\Theta}} \|\mathbf{Y} - \mathbf{\Phi}\mathbf{\Theta}\|_F^2, \text{ s. t. } \mathbf{\Theta} \in \mathcal{K} \cap \mathcal{P}. \tag{4}$$

$\mathbf{D}$ can actually be found using the solution of (4), by selecting the atoms of $\mathbf{\Phi}$ which have been used at least once in the representation of $\mathbf{Y}$. This formulation has some similarities with the convex formulation of Friedman *et al.* [19], where they combine the convex ℓ_1 and ℓ_2 penalties to promote an overcomplete joint sparsity model. The alternative formulation (4), used in this study, has the benefit of being directly related to the size-p dictionary selection problem. Furthermore, the associated iterative algorithm, as presented in Section III offers a complexity that scales well with the dimension of the problem, that can be large in many practical problems.

A. Boundedness and Uniqueness of the Solutions

The constraint set $\mathcal{K} \cap \mathcal{P}$ is unbounded. This means that for any given finite value t, there exists at least a point $\mathbf{X} \in \mathcal{K} \cap \mathcal{P}$ such that $\max_i \|\mathbf{x}_i\|_\infty > t$. It is necessary to find a condition which guarantees the boundedness of the solution of (4). Such a condition is given in Lemma 1. To prove this lemma, we use the following proposition.

Proposition 1: Let $\mathcal{B}_r^\infty$ be an open ball centred at the origin, with the radius r, defined by $\mathcal{B}_r^\infty = \{\mathbf{A} \in \mathbb{R}^{m \times L}, \max_{i,j} |\{\mathbf{A}\}_{i,j}| < r\}$. For a given $\zeta \in \mathbb{R}_+$, if

$$\text{Null}(\mathbf{\Phi}) \cap \mathcal{K} \cap \mathcal{P} = \{\mathbf{0}\}, \tag{5}$$

there exists a finite radius $r \in \mathbb{R}_+$ such that, $\forall \mathbf{X}_{KP} \in (\mathcal{K} \cap \mathcal{P}) \backslash \mathcal{B}_r^\infty$,

$$\min_{\mathbf{X}_N \in \text{Null}(\mathbf{\Phi})} \|\mathbf{X}_{KP} - \mathbf{X}_N\|_F > \zeta,$$

Proof: $\text{Null}(\mathbf{\Phi})$ is a (linear) subspace of $\mathbb{R}^{m \times L}$ and $\mathcal{K} \cap \mathcal{P}$ is a union of subspaces [20], which intersect at the origin. The shortest distance between a given non-zero point $\mathbf{X}_{KP}$ in $\mathcal{K} \cap \mathcal{P}$ and $\text{Null}(\mathbf{\Phi})$

is non-zero, as $\mathbf{0}$ is the only point in $\text{Null}(\boldsymbol{\Phi}) \cap \mathcal{K} \cap \mathcal{P}$. This distance becomes larger, with $\alpha \mathbf{X}_{KP}$, for increasing $\alpha \in \mathbb{R}_+$. Therefore, there exists a radius r, which any point in $\mathcal{K} \cap \mathcal{P}$, located outside of $\mathcal{B}_r^\infty$, is at least ζ away from the closest point in $\text{Null}(\boldsymbol{\Phi})$. ∎

Lemma 1: Let the null space of the operator $\boldsymbol{\Phi}$, in the space $\mathbb{R}^{m \times L}$, be noted by $\mathcal{N}$. The set of all solutions of (4) is bounded if and only if $\mathcal{N} \cap \mathcal{K} \cap \mathcal{P} = \{\mathbf{0}\}$.

Proof: Let $\mathbf{X}$ be a solution of (4) and $\mathbf{X}_\mathcal{N}$ and $\mathbf{X}_\mathcal{R}$ respectively be the projection of $\mathbf{X}$ onto the null-space and range of $\boldsymbol{\Phi}$. As $\|\mathbf{X}\|_F^2 = \|\mathbf{X}_\mathcal{N}\|_F^2 + \|\mathbf{X}_\mathcal{R}\|_F^2$, we only need to show that $\|\mathbf{X}_\mathcal{N}\|_F$ and $\|\mathbf{X}_\mathcal{R}\|_F$ are bounded for any solution of (4)[1]. As the matrix $\mathbf{0} \in \mathcal{K} \cap \mathcal{P}$, any solution of (4) should then have smaller objective than this matrix. We can then have,

$$2\|\mathbf{Y}\|_F \geq \|\mathbf{Y}\|_F + \|\mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}\|_F$$
$$\geq \|\boldsymbol{\Phi}\mathbf{X}\|_F$$
$$\geq \sigma_{min}\|\mathbf{X}_\mathcal{R}\|_F,$$

where σ_{min} is the minimum (non-zero) singular value of $\boldsymbol{\Phi}$. This induces $\|\mathbf{X}_\mathcal{R}\|_F \leq 2\sigma_{min}^{-1}\|\mathbf{Y}\|_F$, which is the boundedness of $\|\mathbf{X}_\mathcal{R}\|_F$.

We respectively denote Λ and $\bar{\Lambda}$ as the support index of $\mathbf{X}$, *i.e.* $\mathbf{X}_\lambda \neq 0$, $\lambda \in \Lambda$, and its complement. The matrix $\mathbf{A}_\Lambda$ (respectively $\mathbf{A}_{\bar{\Lambda}}$) is a matrix which is equal to $\mathbf{A}$ on the support index (respectively on the complement of support index) and zero on the other indices. The solution $\mathbf{X}$ is zero on the indices specified by $\bar{\Lambda}$, *i.e.* $\mathbf{X}_{\bar{\Lambda}} = \mathbf{0}$. $\mathbf{X}_{\bar{\Lambda}} = \mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}} + \mathbf{X}_{\mathcal{R}_{\bar{\Lambda}}} = \mathbf{0}$, shows that $\mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}} = -\mathbf{X}_{\mathcal{R}_{\bar{\Lambda}}}$. On the other hand,

$$\|\mathbf{X}_{\mathcal{R}_{\bar{\Lambda}}}\|_F^2 = \|\mathbf{X}_\mathcal{R}\|_F^2 - \|\mathbf{X}_{\mathcal{R}_\Lambda}\|_F^2$$
$$\leq 4\sigma_{min}^{-2}\|\mathbf{Y}\|_F^2,$$

which assures the boundedness of $\mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}}$. We finally need to show that $\mathbf{X}_{\mathcal{N}_\Lambda}$ is also bounded. Momentarily assume that $\mathbf{X}_{\mathcal{N}_\Lambda}$ is unbounded. $\mathbf{X}_\Lambda = \mathbf{X}_{\mathcal{R}_\Lambda} + \mathbf{X}_{\mathcal{N}_\Lambda}$ is in $\mathcal{K} \cap \mathcal{P}$ and $\mathbf{X}_\mathcal{N} \in \mathcal{N}$. As $\mathbf{X}_\Lambda$ is unbounded when $\mathbf{X}_{\mathcal{N}_\Lambda}$ is unbounded, we can use Proposition 1 with $\zeta = \left(\|\mathbf{X}_{\mathcal{R}_\Lambda}\|_F^2 + \|\mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}}\|_F^2\right)^{\frac{1}{2}}$ as follows,

$$\zeta^2 < \|\mathbf{X}_\Lambda - \mathbf{X}_\mathcal{N}\|_F^2$$
$$= \|(\mathbf{X}_{\mathcal{R}_\Lambda} + \mathbf{X}_{\mathcal{N}_\Lambda}) - (\mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}} + \mathbf{X}_{\mathcal{N}_\Lambda})\|_F^2$$
$$= \|\mathbf{X}_{\mathcal{R}_\Lambda} - \mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}}\|_F^2$$
$$= \|\mathbf{X}_{\mathcal{R}_\Lambda}\|_F^2 + \|\mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}}\|_F^2,$$

which contradicts with the fact that $\zeta^2 = \|\mathbf{X}_{\mathcal{R}_\Lambda}\|_F^2 + \|\mathbf{X}_{\mathcal{N}_{\bar{\Lambda}}}\|_F^2$. Therefore the assumption of unboundedness of $\mathbf{X}_{\mathcal{N}_\Lambda}$ is incorrect, which complete the proof of boundedness of $\mathbf{X}$.

If (5) is not valid, we have a non-zero $\boldsymbol{\Delta} \in \mathbb{R}^{m \times L}$ in the null space of $\boldsymbol{\Phi}$, which is also overcomplete joint sparse. This means that any non-zero (k, p)-joint sparse solution $\mathbf{X}$, with the same support as $\boldsymbol{\Delta}$, generates another solution of (4) given by $\mathbf{X} + \lambda\boldsymbol{\Delta}$, for any $\lambda \in \mathbb{R}$, which lies in $\mathcal{K} \cap \mathcal{P}$, since $\mathbf{X}$ and $\boldsymbol{\Delta}$ share the support. Hence, λ can be chosen arbitrary large, which shows that the set of all solutions of (4) cannot be bounded. ∎

It is generally difficult to check (5) for a given mother dictionary. However, if the mother dictionary is in a general position, when the dimension of signal space nL is larger than the sum of the dimension of null space $L(n-m)$ and each subspace kL, which means $k < m$, the Lebesgue measure of the lhs of (5) is zero.

Although this lemma shows the boundedness of the solutions, it does not provide any explicit bound for the results. It means that if the $\text{Null}(\boldsymbol{\Phi})$ subspace is very close to one of the subspaces in $\mathcal{K} \cap \mathcal{P}$, ζ can become very large.

---

[1]We here show that Frobenius norm of $\mathbf{X}$ is bounded, which induces the boundedness of $\max_i \|\mathbf{x}_i\|_\infty$.

The reader may notice that from the *optimality* of $\mathbf{X}$ in the proof of Lemma 1, we only used the fact that the objective at $\mathbf{X}$ is less than the objective at $\mathbf{0}$. Therefore we can easily extend this lemma to $\boldsymbol{\Theta}$'s which are not the optimal solutions, but satisfy a similar condition and derive the boundedness of the search space.

Corollary 1: The set $\{\boldsymbol{\Theta} \in \mathcal{K} \cap \mathcal{P}, \|\mathbf{Y} - \boldsymbol{\Phi}\boldsymbol{\Theta}\|_F \leq \|\mathbf{Y}\|_F\}$ is bounded if (5) is true.

It is always useful to know when an optimisation problem like (4), has a unique solution. This is particularly useful in the dictionary design problem, as the other formulations often have multiple solutions. This is caused by the fact that any permutation of a dictionary is also a solution for the problem. This indeed makes the convexification of the problem much more challenging.

We can use a general theorem of the Union of Subspaces (UoS) model to show the injection of the mapping $\boldsymbol{\Phi}$. In the UoS signal model, we assume that, while a given class of signals spans the whole space, each signal (approximately) lies in a low-dimensional subspace. The UoS model is a general and useful model for many sampling and signal processing applications. Here, we only consider a *finite* union of subspaces model. For more information and extensions to other types of UoS model, please see [21]. [21] expresses that:

Theorem 1 (Th. 2.6 of [21]): Let $\mathcal{A}$ be a union of subspace, equipped with a proper measure. Almost all linear maps $\boldsymbol{\Phi} : \mathbb{R}^N \to \mathbb{R}^M$ are one to one on almost all elements of $\mathcal{A}$, whenever $k < M \leq k_{max} - 1$, where k and k_{max} respectively are the (maximum) dimension of each subspace and the maximum dimension of the union of two subspaces of $\mathcal{A}$.

In our setting, if $k < m$, almost all linear maps $\boldsymbol{\Phi}_d = \text{diag}\{\boldsymbol{\Phi}\} \in \mathbb{R}^{mL \times nL}$, with $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$, are one to one on almost all elements of the (k, p)-joint sparse matrices. $\boldsymbol{\Phi}_d$ is a diagonal matrix with $\boldsymbol{\Phi}$ on the main diagonal. Interested readers may notice that the derived condition, *i.e.* $k < m$, is indeed the sufficient condition for the lhs of (5) to have zero measure.

We now derive a sufficient condition for the uniqueness of the solution in a deterministic sense. It is indeed a particular case of the uniqueness results for the UoS model [20]. The one to one map of an operator $\boldsymbol{\Phi} : \mathbf{x} \to \boldsymbol{\Phi}\mathbf{x}$, *i.e.* invertible sampling, is defined as,

Definition 1 (Def. 2 in [20]): We call $\boldsymbol{\Phi}$ an *invertible* sampling operator for a union of subspace $\mathcal{A}$, if each $\mathbf{x} \in \mathcal{A}$ is uniquely determined by its sampling data $\boldsymbol{\Phi}\mathbf{x}$; that means for every $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathcal{A}$,

$$\boldsymbol{\Phi}\mathbf{x}_1 = \boldsymbol{\Phi}\mathbf{x}_2 \text{ implies } \mathbf{x}_1 = \mathbf{x}_2$$

With this definition we can derive a sufficient condition for the uniqueness of the solution of (4) as follows,

Lemma 2: Let $k \leq \frac{m}{2}$, $p \leq \frac{n}{2}$, $\mathcal{N} = \text{Null}\{\boldsymbol{\Phi}\}$ and $\mathcal{K}_{2k} := \{\mathbf{X} \in \mathbb{C}^{m \times L} : \|\mathbf{x}_l\|_0 \leq 2k, \forall l \in [1, L]\}$ and $\mathcal{P}_{2p} := \{\mathbf{X} \in \mathbb{C}^{m \times L} : \|\mathbf{X}\|_{0,\infty} \leq 2p\}$ The optimisation problem (4) has a unique solution if

$$\mathcal{N} \cap \mathcal{K}_{2k} \cap \mathcal{P}_{2p} = \{\mathbf{0}\} \tag{6}$$

Proof: Let the solution not be unique and we have $\mathbf{X}_1$ and $\mathbf{X}_2$ as two distinctive solutions of (4). We have $\boldsymbol{\Phi}\mathbf{X}_1 = \boldsymbol{\Phi}\mathbf{X}_2 = \mathbf{Y}$, which means $\boldsymbol{\Phi}(\mathbf{X}_1 - \mathbf{X}_2) = \mathbf{0}$. As $\mathbf{X}_1 - \mathbf{X}_2 \in \mathcal{K}_{2k} \cap \mathcal{P}_{2p}$ and $\mathbf{X}_1 - \mathbf{X}_2 \in \mathcal{N}$, it should be $\mathbf{0}$, which gives $\mathbf{X}_1 = \mathbf{X}_2$ and it contradicts with the fact that they are distinctive solutions. ∎

Remark 1: Note that Lemma 2 presents a *sufficient* condition for the uniqueness of the solution, which is different to the standard k-sparse and p-joint sparse UoS models. Similar to the general form of block-sparse model, this is caused by the fact that some of the sparsity patterns in $\mathcal{K}_{2k} \cap \mathcal{P}_{2p}$ cannot be divided to two disjoint sparsity patterns in $\mathcal{K} \cap \mathcal{P}$.

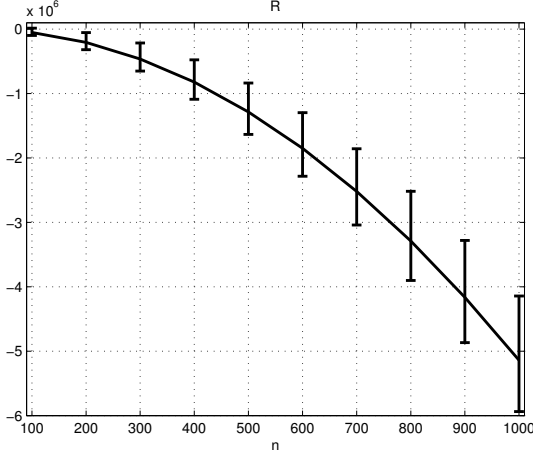*Remark 2:* The boundedness of the solutions of (4) needs a weaker

Fig. 1. $\mathcal{R}$ for different n, when $\delta = \frac{1}{4}$, $\rho = \frac{1}{10}$ and $t = 100$. f is shown by the solid line and the bounds for $\mathcal{R}$ are shown with error bars, see (10).

condition than its uniqueness. We can actually use the uniqueness condition of Lemma 2 to show the boundedness of the solution.

B. Number of Subspaces

It was mentioned that the introduced signal model is a UoS model, as fixing the support coefficient, generates a low-dimensional subspace of the $\mathbb{R}^{m \times L}$. We are restricting the set of matrices which are k-sparse on each column, to the matrices which are also p-joint sparse. Such a restriction reduces the number of admissible subspaces, which increases the robustness of the mapping $\mathbf{\Phi}$ on its domains. In practical applications, we need some robustness to the noise and model mismatches for a successful sparse recovery. This is indirectly related to the distance between each two distinct points, after mapping. If two points have some small distance after the mapping, the embedding is sensitive to the noise. A measure which characterises such a robustness is the restricted isometry constant δ for each UoS model [7]: a large δ ensures a more robust embedding. We refer the readers to [7] for more information about the definition and implication of the restricted isometry constant.

Based upon [21, Corollary 3.6], a necessary number of measurements to have a robust embedding with a particular δ, has a lower bound, which is proportional to $\ln(N_s)$ and inversely proportional to $c_\delta - \ln(\Delta_s)$, where N_s is the total number of subspaces, c_δ is a function of δ and Δ_s is the subspace separation of the proposed UoS [21, Eq. (18)]. Δ_s decreases by restricting the UoS to a subset of the original UoS. We therefore reduce the necessary number of training samples in this context, by decreasing N_s. In the following, we characterise the reduction in the number of subspaces, using the proposed UoS model, in the comparison with the k-sparse signal model.

When the matrix is k-sparse on rows, we have L times $\binom{n}{k}$ options to choose the support. The number of subspaces is thus $\binom{n}{k}^L$. If we also restrict the matrices to be p-joint sparse, we choose k positions for each row, within the selected p rows. We have therefore $\binom{p}{k}^L \binom{n}{p}$ subspaces. To quantify exponential reduction in the number of subspaces using the (k, p)-joint sparsity model, we approximate $\mathcal{R}$, defined as,

$$\mathcal{R} := \log_2 \frac{\binom{p}{k}^L \binom{n}{p}}{\binom{n}{k}^L}. \tag{7}$$

To find some upper and lower bounds for $\mathcal{R}$, we use the concept of *binary entropy* $\mathcal{H}$ from Information Theory, which is defined as

follows,

$$\mathcal{H}(\tau) \triangleq -\tau \log_2(\tau) - (1 - \tau) \log_2(1 - \tau) \tag{8}$$

where $0 \leq \tau \leq 1$ is the probability of a binary number. We can now bound $\binom{n}{k}$ as follows [22, eq. 12.40],

$$\frac{1}{n+1} 2^{n\mathcal{H}\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq 2^{n\mathcal{H}\left(\frac{k}{n}\right)}. \tag{9}$$

Using the similar bounds for $\binom{p}{k}$ and $\binom{n}{p}$, and after some simple algebraic manipulations, we can derive a bound for $\mathcal{R}$ as follows,

$$-L \log_2(p+1) - \log_2(n+1) + f(k, p, n, L)$$
$$\leq \mathcal{R} \leq L \log_2(n+1) + f(k, p, n, L), \tag{10}$$

where $f(k, p, n, L) \triangleq n\mathcal{H}\left(\frac{p}{n}\right) + Lp\mathcal{H}\left(\frac{k}{p}\right) - nL\mathcal{H}\left(\frac{k}{n}\right)$ If we replace the binary entropy in $f(k, p, n, L)$, we can derive an explicit formulation for f as follows,

$$f(k, p, n, L) = \left(p \log_2 \frac{n}{n-p} + p \log_2 \frac{n-p}{p} \right)$$
$$- L \left(n \log_2 \frac{n}{n-k} - p \log_2 \frac{p}{p-k} + k \log_2 \frac{n-k}{p-k} \right). \tag{11}$$

As (10) depends on many parameters, it is hard to figure out the reduction in the number of subspaces from $\mathcal{R}$. To demonstrate this better, we can fix $\delta = \frac{p}{n}$, $\beta = \frac{k}{p}$ and $t = \frac{L}{n}$, and plot f based on n, which is an approximation for $\mathcal{R}$, and showing the bounds of (10) with some error bars. If we choose $\delta = \frac{1}{4}$, $\beta = \frac{1}{10}$ and $t = 100$, the bounds for $\mathcal{R}$ are plotted as functions of n in Figure 1. As $2^{\mathcal{R}}$ is the ratio between the number of subspaces in the new model and k-sparsity model, we can see that ratio is significantly reduced for large n. In other words, the search space for the solution is now much smaller, which may boost the exact recovery using practical recovery algorithms, as we can see in the simulation section.

III. A PRACTICAL OPTIMISATION ALGORITHM

Although the objective of (4) is quadratic, the optimisation of (4) subject to the non-convex constraints $\mathcal{K}$ and $\mathcal{P}$, is not easy. Most of the efficient optimisation techniques cannot be used in this setting. A powerful technique, called the projected gradient, can be used when the projection onto the admissible set is available. In the space of real matrices $\mathbb{R}^{n \times L}$, the projection of a point $\mathbf{X} \in \mathbb{R}^{n \times L}$ onto a closed set $\kappa \subseteq \mathbb{R}^{n \times L}$ is defined by $\mathcal{P}_\kappa(\mathbf{X}) := \arg \min_{\mathbf{\Theta} \in \kappa} \|\mathbf{\Theta} - \mathbf{X}\|_\chi$, where $\| \cdot \|_\chi$ is the norm of the proposed space. We use the Hilbert - Schmidt, or Frobenius, norm here, as it is more related to the quadratic objective (3), *i.e.* using the same normed space, and we can analytically find the projection. In this setting, a projection onto $\mathcal{K}$ can be found by keeping the k largest coefficients of each column and letting the others be zero. The projection onto $\mathcal{P}$ can be found by keeping the p rows of $\mathbf{X}$ with the largest maximum absolute values and letting the other rows be zero. Sadly, the projection onto the intersection of $\mathcal{K}$ and $\mathcal{P}$ is not analytically possible, the projected gradient algorithm cannot be used in its canonical form. A property of the admissible sets $\mathcal{K}$ and $\mathcal{P}$ is that the consequent projections of a point in these sets provide a point in the intersection of them, which may indeed not necessarily be the projection onto $\mathcal{K} \cap \mathcal{P}$. The following lemma shows that alternating projection onto $\mathcal{K}$ and $\mathcal{P}$ converges in a single consecutive projections, *i.e.* two projections in total.

Lemma 3: Let $\mathbf{X}$ be a matrix in $\mathbb{R}^{n \times L}$. The following two statements hold,

$$\mathcal{P}_\mathcal{P} \mathcal{P}_\mathcal{K}(\mathbf{X}) \in \mathcal{P} \cap \mathcal{K}$$
$$\mathcal{P}_\mathcal{K} \mathcal{P}_\mathcal{P}(\mathbf{X}) \in \mathcal{P} \cap \mathcal{K}. \tag{12}$$

Proof: Projections of $\mathbf{X}$ onto $\mathcal{K}$ or $\mathcal{P}$ shrinks some of $\mathbf{X}$'s non-zero elements to zero and does not produce any further non-zero elements. This simply shows that the projection of a point in $\mathcal{K}$, onto $\mathcal{P}$, gives a new point which is still in $\mathcal{K}$. It assures the first statement. The second statement can be shown similarly. ∎

Remark 3: According to Lemma 3, although $\mathcal{P} \cap \mathcal{K}$ includes $\mathcal{P}_{\mathcal{P}}\mathcal{P}_{\mathcal{K}}(\mathbf{X})$ and $\mathcal{P}_{\mathcal{K}}\mathcal{P}_{\mathcal{P}}(\mathbf{X})$, for any $\mathbf{X}$, it does not induce that these operators are identical. We practically found that using $\mathcal{P}_{\mathcal{K}}\mathcal{P}_{\mathcal{P}}(\mathbf{X})$ works better for the purpose of this paper.

Remark 4: The sets $\mathcal{K}$ and $\mathcal{P}$ are non-convex and the projection onto each of these sets may thus be non-unique. In this case we can randomly choose one of the projections.

A. Overcomplete Joint Sparsity Dictionary Selection Algorithm (OJSDS)

We use a gradient based method which iteratively updates the current solution $\mathbf{X}^{[n]}$, in the negative gradient direction and maps onto a point in $\mathcal{K}\cap\mathcal{P}$, to approximately solve (4). If $\psi(\mathbf{X}) := \|\mathbf{Y} - \mathbf{\Phi X}\|_F^2$, the gradient of ψ can be found as follows,

$$\mathbf{G} := \frac{\partial}{\partial \mathbf{X}}\psi(\mathbf{X}) = 2\mathbf{\Phi}^H(\mathbf{\Phi X} - \mathbf{Y}). \tag{13}$$

An important part of the gradient descent methods, is how to select the step size. An efficient step size selection technique for unconstrained quadratic minimisation problems, with objectives like $\psi(\mathbf{X})$, is to use half of the spectral radius of linear operator, here $\mathbf{\Phi}$, as follows,

$$\mu = \frac{1}{2}\frac{\mathbf{G}^H\mathbf{\Phi}^H\mathbf{\Phi G}}{\mathbf{G}^H\mathbf{G}}$$

Such a step size is optimal for the first order gradient descent minimisation of the unconstrained problem with the quadratic objective $\psi(\mathbf{X})$. In a constrained minimisation scenario, we can choose a similar initial step size and shrink the size, if the objective increases. It thus needs an extra step to check that the objective is actually not increased after each update of the parameters. A more clever initial step size was selected in [18] for the sparse approximations of k-sparse signals. If the support of sparse coefficient vectors are fixed, *i.e.* the overall projection steps do not change the support, the update is only in the direction which changes current non-zero coefficients. When the problem size is shrunk to the space of current support, the problem is quadratic and the step size can be similarly calculated using the gradient matrix $\mathbf{G}$, constrained to the support, as follows,

$$\mu = \frac{1}{2}\frac{\mathbf{G}_S^H\mathbf{\Phi}^H\mathbf{\Phi G}_S}{\mathbf{G}_S^H\mathbf{G}_S}$$

where $\mathbf{G}_S \in \mathbb{R}^{m\times L}$ is the gradient matrix $\mathbf{G}$ masked by the support of $\mathbf{X}$, S, as follows,

$$\{\mathbf{G}_S\}_{i,j} = \begin{cases} \{\mathbf{G}\}_{i,j} & \{\mathbf{X}\}_{i,j} \neq 0 \\ 0 & \text{Otherwise} \end{cases}$$

A pseudo-code for the algorithm is presented in Algorithm 1. The condition which is checked in line 10, guarantees that the algorithm reduces the objective by updating the coefficients. As the dictionary selection algorithm 1 is based upon a gradient projection type technique, the learned dictionary may be more suitable for such greedy sparse approximation techniques. However, the simulation results show that the reference dictionary can be recovered using this algorithm, given a rich set of training samples. If the real signal is sparse and the dictionary satisfies the exact recovery conditions, the dictionary is thus optimal for any sparse recovery algorithms.

1: **initialisation:** $\mathbf{X}^{[0]} = \mathcal{P}_{\mathcal{K}}(\mathcal{P}_{\mathcal{P}}(\mathbf{\Phi}^H\mathbf{Y}))$, $S = \mathrm{supp}(\mathbf{X}^{[0]})$, $\rho < 1$, $\beta < 1$, $\epsilon \ll 1$, $t = 0$, $K \geq 1$ and $i = 0$
2: **while** $i < K$ & $t \neq 1$ **do**
3:   $\mathbf{G} = 2\mathbf{\Phi}^H(\mathbf{\Phi X}^{[i]} - \mathbf{Y})$
4:   $\mu = \frac{1}{2}\frac{\mathbf{G}_S^H\mathbf{\Phi}^H\mathbf{\Phi G}_S}{\mathbf{G}_S^H\mathbf{G}_S}$
5:   $\mathbf{Z} = \mathcal{P}_{\mathcal{K}}(\mathcal{P}_{\mathcal{P}}(\mathbf{X}^{[i]} - \mu\mathbf{G}))$
6:   **if** $\|\mathbf{X}^{[i]} - \mathbf{Z}\|_F^2 < \epsilon$ **then**
7:     $t = 1$
8:   **end if**
9:   **if** $t \neq 1$ **then**
10:     **while** $\mu > \frac{\rho}{2}\frac{\|\mathbf{X}^{[i]} - \mathbf{Z}\|_F^2}{\|\mathbf{\Phi}(\mathbf{X}^{[i]} - \mathbf{Z})\|_F^2}$ **do**
11:       $\mu = \beta.\mu$
12:       $\mathbf{Z} = \mathcal{P}_{\mathcal{K}}(\mathcal{P}_{\mathcal{P}}(\mathbf{X}^{[i]} - \mu\mathbf{G}))$
13:     **end while**
14:   **end if**
15:   $i = i + 1$
16:   $\mathbf{X}^{[i]} = \mathbf{Z}$
17:   $S = \mathrm{supp}(\mathbf{X}^{[i]})$
18: **end while**
19: $\mathbf{X}^* = \mathbf{X}^{[i-1]}$
20: **output:** $\mathbf{X}^*$

Algorithm 1: Alternating Projected Gradient for Dictionary Selection
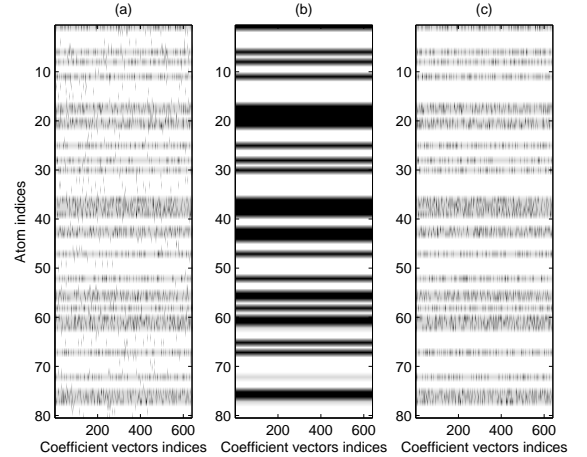

Fig. 2. Dictionary selection results using, (a) $\mathcal{K}$, (b) $\mathcal{P}$ and (c) $\mathcal{K} \cap \mathcal{P}$ as admissible sets. The black dots in each plot indicate non-zero coefficients. In plot (b), as dots are very populated, we observe solid horizontal lines. Gray horizontal lines are plotted as a guideline, for the correct dictionary.

In the following theorem, we prove that Algorithm 1 is numerically stable and the generated sequence has limit points.

Theorem 2: Let $\mathbf{X}^{[0]} \in \mathbb{R}^{m\times L}$ be a bounded initial point. The gradient based method of Algorithm 1, generates a bounded sequence of solutions, which accumulate.

Proof: As the algorithm reduces the objective at each iteration, the search space is a bounded subset of $\mathcal{K}\cap\mathcal{P}$, based upon Corollary 1. $\mathcal{K}\cap\mathcal{P}$ is a closed set, the search space is then a compact subset of $\mathbb{R}^{m\times L}$. The sequence generated by Algorithm 1, lives in a compact set, which is enough to have bounded accumulation points, based on the Bolzano-Weierstrass theorem. ∎
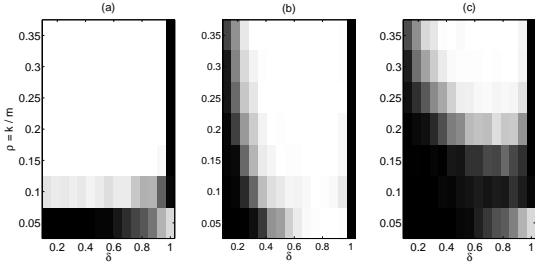
Fig. 3. Phase transition using, (a) $\mathcal{K}$, (b) $\mathcal{P}$ and (c) $\mathcal{K} \cap \mathcal{P}$ as admissible sets. The black area indicates successful recovery of the dictionary.

IV. SIMULATIONS

In the first experiment, a dictionary $\mathbf{\Phi} \in \mathbb{R}^{20 \times 80}$ was randomly generated using a normal zero mean distribution with unit variance and normalised to have unit ℓ_2-norm on each column. The target dictionary $\mathbf{D} \in \mathbb{R}^{20 \times 30}$ was generated by randomly selecting $p = 30$ atoms of $\mathbf{\Phi}$. A number $L = 320$ of k-sparse coefficient vectors (with $k = 4$), were generated by randomly selecting the support, with a uniform distribution of the magnitudes in $[0.2, 1]$ and random signs. A set of training matrix $\mathbf{Y}$ of length L were generated using the generative model and randomly generated sparse vectors. To recover the reference dictionary $\mathbf{D}$, given $\mathbf{Y}$, p and k, we used a gradient descent based algorithm similar to Algorithm 1, with three different admissible sets, and demonstrate the advantages of the proposed technique. We first used $\mathcal{K}$ from (2) and no constraint on the row-sparsity of the coefficient matrix $\mathbf{X}$ and showed the recovered support of the sparse matrix in the left panel (a) of Figure 2. If we only assume joint sparsity model and use $\mathcal{P}$ from (3) as the admissible set, we find the coefficient matrix whose support is shown in the middle panel (b) of the same figure. Using both constraint sets, as explained in Algorithm 1, provides a coefficient matrix whose support is shown in the right panel. The correct $\mathcal{J}$ is shown in these plots using some *grey lines*. It is clear that the proposed projected gradient onto both sets can correctly recover $\mathcal{J}$, where the other two methods have some errors in the recovery.

This experiment can be repeated for different $\delta = \frac{p}{n}$ and $\rho = \frac{k}{m}$ by selecting a range of p and k's, while keeping m and n fixed. If we repeat the simulations 100 times for each setting and calculate the average exact dictionary recovery, we can plot the phase transition for each methods. We have plotted such phase transitions in Figure 3, with k sparsity constraint in (a), p joint sparsity constraint in (b) and proposed constraint in (c). The black colour means high exact dictionary recovery. The area with exact recovery in (c) is larger than the same areas in (a) and (b) added together. This clearly demonstrates the relevance of the new framework.

In the next set of experiments, we will select a subset of the Curvelet [23] dictionary for the sparse representation of fingerprints. We chose a Curvelet transform for the image size 64 by 64. The mother dictionary $\mathbf{\Phi} \in \mathbb{R}^{4096 \times 10521}$ is roughly 2.59 times overcomplete, which we want to shrink to half size, *i.e.* $\mathbf{D} \in \mathbb{R}^{4096 \times 5260}$. This is indeed a large scale dictionary learning problem, which is difficult to solve in a standard dictionary learning setting. With the help of the proposed method, we can handle such a big dictionary selection process, as we need fewer training samples, only need to keep a sparse matrix, *i.e.* sparse representation matrix, in the memory and use the fact that the mother dictionary has a fast implementation. We assume the sparsity of each image patch is $k = 1052 \approx 0.1N$ and $L = 64$. We used two different settings here to choose the dictionary, a) p-joint sparsity model and b) (k, p)-overcomplete joint sparse model. The simulations were done in the Matlab environment,

on a 12-core, 2.6 GHz linux machine, which respectively took 72 and 90 seconds to learn $\mathbf{D}_p$ and $\mathbf{D}_{(k,p)}$. Another fingerprint image was used to test the selected dictionaries. The original image and the k sparse representation of the original image with $\mathbf{\Phi}$ are shown in the first row of Figure 4. The k-sparse representation with the learned $\mathbf{D}$'s are shown in the second row of this figure. The left image is the representation with learned $\mathbf{D}_p$, when the model was p-joint sparse and the right image is the same, but with $\mathbf{D}_{(k,p)}$, where the (k, p)-overcomplete joint sparse model was incorporated. As we can see the PSNR of the representation with the shrunk dictionary $\mathbf{D}_{(k,p)}$ is slightly better than the other. We can also see the bottom-right quarter of these images in Figure 5, in the same order. The sparse reconstructed images are actually denoised and the reconstructed image using $\mathbf{D}_{(k,p)}$ is more similar than $\mathbf{D}_p$ to the image reconstructed using $\mathbf{\Phi}$.

We setup a new experiment with audio signals to demonstrate the performance of the proposed dictionary selection algorithm in comparison with the fixed dictionaries and another dictionary learning methods. To this end, we used some recorded audio data from BBC radio 3 (mostly classical musics), and down-sampled the signals at a sampling rate of 32k samples per second, as there is very little energy above 16 kHz. We randomly selected a $\mathbf{Y} \in \mathbb{R}^{1025 \times 8196}$ from more than eight hours of recorded audio. A three times overcomplete mother dictionary was generated using a two times frequency oversampled DCT plus the Delta Dirac transform, *i.e.* identity matrix. The reason for such a selection is to incorporate the temporal and harmonic properties of the audio. There has been a question on how useful can be to combine such dictionaries and how many DCT atoms are necessary. We thus found a subset, *i.e.* $p = \frac{3}{2} * 1024$, of the mother atoms. If we run the proposed dictionary selection algorithm with $k = 128$, for $K = 1000$ iterations, and plot the frequency of appearance of the mother atoms in $\mathbf{X}$, we get the plot of Figure 6. The low-frequency DCT atoms have been used most, while high-frequency DCT atoms have not been selected in $\mathbf{D}$. Although there is no regular pattern for the selected delta Diracs, it is clear that some Dirac atoms close to the boundary of the window have been selected, *i.e.* close to the atom indices 2048 and 3072. If we plot the ℓ_2 errors of representing a set of test data $\widehat{\mathbf{Y}} \in \mathbb{R}^{1025 \times 8196}$, which is randomly selected from the same audio database, through out the iterations, we get the plots of Figure 7. The ℓ_2 errors corresponding to using the mother and two times overcomplete DCT dictionaries, are also shown for the reference with solid and dash-doted lines, respectively. The final SNR using the selected dictionary is $26.43dB$, which is slightly worse than using the mother dictionary, *i.e.* $26.53dB$, but is also significantly better than using the two times overcomplete DCT, *i.e.* $25.96dB$. For a comparison, we also ran the sparse dictionary learning [10], with the same training data samples. The reason for selecting this dictionary learning algorithm is that it has some similarities with the proposed framework here, and it provides a relatively fast dictionary, *i.e.* an extra sparse matrix-vector multiplication is also necessary. In the sparse dictionary learning, we used the same mother dictionary we used earlier, the objective multipliers $\lambda = \gamma = .01$ and ran the simulations for 1000 iterations. The ℓ_2 errors of using the sparse learned dictionary is shown by the dotted line in Figure 7. The final SNR is not as good as when we use other dictionaries. As one aim of the proposed dictionary selection method is to provide a fast dictionary, we also measured the average calculation time of the forward and backward applications of $\mathbf{D}$ on the same desktop machine as previous experiments, while using a single core of the processor, for a fair comparison. The application of *subselected* dictionary $\mathbf{D}$ and $\mathbf{D}^T$, using a regular matrix-vector multiplication and a fast operator implementation, are shown in Table I. This table shows that, using a fast selected dictionary, speeds up the

TABLE I
COMPUTATION TIME OF SINGLE USE SUBSELECTED DICTIONARY AND ITS
TRANSPOSE IN ms.

	$\mathbf{D}$	$\mathbf{D}^T$
Regular	288	285
Fast	230	124

TABLE II
COMPUTATION TIME OF DIFFERENT DICTIONARY LEARNING/SELECTION
ALGORITHMS (IN SECONDS) WITH DIFFERENT AUDIO SIGNAL SIZES, *i.e.*
256, 512 AND 1024.

	256	512	1024
OJSDS	1315	9194	35873
SparseDL [10]	4273	14122	48294
K-SVD [24]	5665	29447	-

practical sparse approximation algorithms, as applying the dictionary and its transposed, often are the most computationally expensive parts of such algorithms.

Learning a dictionary in the settings of previous experiment, using the canonical dictionary learning algorithms, *e.g.* K-SVD, MOD and MMDL [4], needs a large set of training samples and the learning computational time is generally high. To compare with the proposed method, we chose K-SVD, as it shows promising results, using a relatively small training set. We started with a small size problem and gradually increased the size, until the simulation was no longer tractable for us. We used $L = 8192$ training samples and repeated the previous simulations with $m = 256$, 512 and 1024. We have presented the running times of the learning with the proposed (OJSDS), Sparse and K-SVD dictionary learning algorithms in Table II, using a single core of the processor. We were unable to apply the K-SVD algorithm, when $m = 1024$, for its high computation time. We only used $L = 8\,m$ samples for OJSDS, as it does not need a large training set. The low-complexity of our algorithm, with respect to the other algorithms, is clearly the lowest. As we increase the number of training samples while using different window sizes in OJSDS, the computational cost increases faster, w.r.t. the problem size, than the others. However, when the size of training samples are equal, *i.e.* $8\,m = 8192$ with $m = 1024$, our proposed technique is still 25% faster than SparseDL.

We calculated the running time and the final SNR, using Normalised IHT (NIHT) [18], $k = \lceil m/8 \rceil$, 32 test signals and presented the results in Table III. This table shows that the learned dictionary with the proposed method is slightly slower than K-SVD dictionary, but it provides the highest SNR's, in the sparse approximations of audio signals. The reason for being slower than K-SVD is that the implementation of the mother dictionary, as an operator, is slower than, or in the same order of, a simple matrix-vector multiplication for small size problems. Comparing the first and third columns of Table III, we observe that, unlike the unstructured K-SVD dictionary, the subselected dictionaries are of the order $\mathcal{O}(m)$, *i.e.* linear complexity with respect to the signal size. It makes the subselected dictionaries computationally suitable for large scale problems.

V. SUMMARY AND FUTURE WORK

We presented a new technique for dictionary selection for the linear sparse representation, when a collection of possibly suitable atoms and some exemplar signals are available. The dictionary selection problem is reformulated as a more general form of the joint sparse approximation problem, when the number of active locations in sparse coefficients is larger than the size of signal space. As such
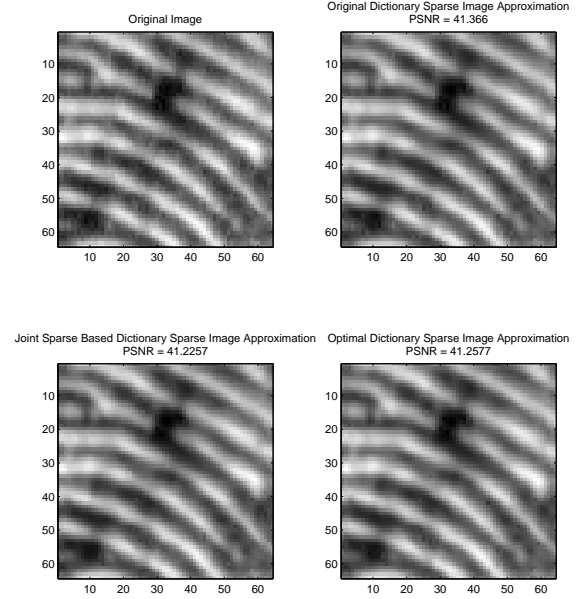

Fig. 4. The original image (top left), the k sparse representation of the original image with the dictionaries, $\mathbf{\Phi}$ (top right), $\mathbf{D}_p$ (bottom left) and $\mathbf{D}_{(k,p)}$ (bottom right).
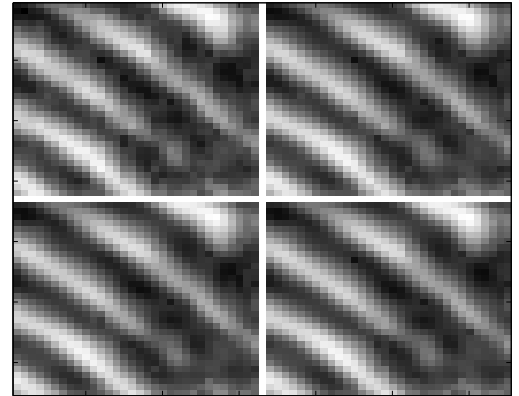

Fig. 5. The bottom-right quarter of the images shown in Figure 4, in the same order.

overcomplete joint sparsity framework has generally infinitely many solutions, the sparsity within the active set helps to regularise the problem. It was shown that the overcomplete joint sparse approximation problem is well-defined under some conditions on the null-space of the matrix generated by the given large set of atoms (mother dictionary). As the objective of the introduced program is continuously differentiable, we used a gradient mapping technique to approximately solve the problem. The introduced algorithm converges in a weak sense (convergence to a bounded non-empty set).

We presented some synthetic data simulation result to support this hypothesis that the introduced algorithm can recover the original dictionary. The phase plot of the dictionary recovery is compared with two other cases, when we use other sparsity models, namely k-sparse and p-joint sparse model. As the simulations with synthetic data
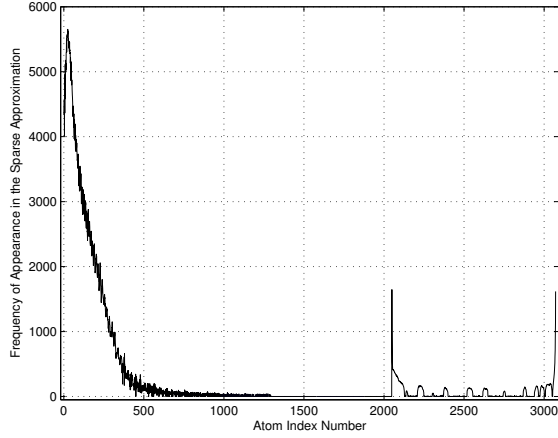
Fig. 6. The frequency of selected atoms, per 8192 trials. The first 2048 atoms are the two times frequency oversampled DCT and the last 1024 atoms are Dirac functions.
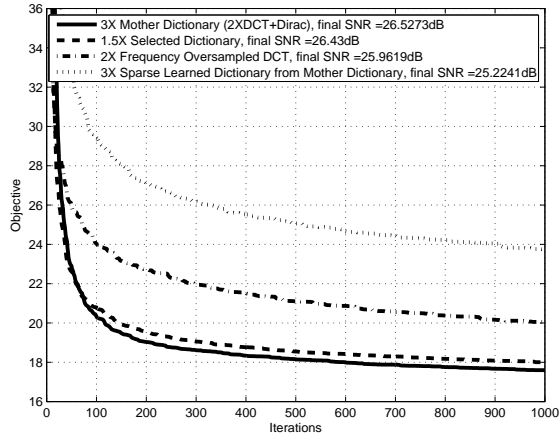

Fig. 7. The ℓ_2 norm error of representations of 8192 testing trials, using Normalised IHT [18] and different dictionaries. The dictionaries are: (a) three times overcomplete DCT+Dirac mother dictionary, (b) the one and a half times overcomplete selected dictionary, (c) a two times frequency oversampled DCT and (d) the leaned sparse dictionary using the mother dictionary of (a) and mothod of [10].

were promising, we also did some simulations to select a subset of a commonly used dictionary, *Curvelet* and Overcomplete DCT+Dirac dictionaries, to reduce the complexity of the sparse coding algorithm. The size of dictionary learning problem is such that it cannot be handled by the vast majority of current dictionary learning algorithms. As we do not need to keep the dictionary in the memory and as the dictionary-vector multiplications can be implemented efficiently, the learning in the new framework is relatively easy. The results show that we can roughly get the same image/audio quality for a specific class of image/audio signals, when we use a smaller dictionary than the mother dictionary.

The new overcomplete joint sparsity model seems an interesting extension of the previously investigated joint sparsity model. We have left the theoretical investigation of exact recovery and other sparse signal processing applications, for the future work.

TABLE III
COMPUTATION TIME (IN SECONDS) AND RECONSTRUCTED SNR OF SPARSE APPROXIMATION (IN dB), USING NIHT, DIFFERENT DICTIONARIES AND AUDIO SIGNAL LENGHTS, *i.e.* 256, 512 AND 1024. NIHT ITERATED 1000 TIMES.

	OJSDS	SparseDL [10]	K-SVD [24]
(256) Time	219	398	129
(256) SNR	16.47	16.00	15.56
(512) Time	422	954	344
(512) SNR	17.97	17.90	16.67
(1024) Time	821	2685	-
(1024) SNR	18.43	17.4	-

VI. ACKNOWLEDGEMENT

REFERENCES

[1] R. Baraniuk, V. Cevher, and M. Wakin, "Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 959–971, 2010.

[2] M. Yaghoobi, L. Daudet, and M. Davies, "Parametric dictionary design for sparse coding," *IEEE Transaction on Signal Processing*, vol. 57, no. 12, pp. 4800–4810, 2009.

[3] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[4] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045 – 1057, June 2010.

[5] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011.

[6] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[7] E. Candés and T. Tao, "Decoding by linear programming," *EEE Transaction on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[8] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. on Information Theory*, vol. 49, no. 12, pp. 3320 – 3325, 2003.

[9] R. Neff and A. Zakhor, "Matching pursuit video coding - part I: Dictionary approximation," *IEEE Trans. on Circuits, Systems for Video Technology*, vol. 12, no. 1, p. 1326, 2002.

[10] M. Yaghoobi and M. Davies, "Compressible dictionary learning for fast sparse approximation," in *IEEE Workshop on Statistical Signal Processing*, Aug. 31- Sept. 3 2009.

[11] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, 2010.

[12] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Parsimonious dictionary learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 2869–2872.

[13] A. Krause and V. Cevher, "Submodular dictionary selection for sparse representation," in *International Conference on Machine Learning (ICML)*, 2010.

[14] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *International Conference on Machine Learning (ICML)*, 2011.

[15] D. Malioutov, "A sparse signal reconstruction perspective for source localization with sensor arrays," Master's thesis, MIT, July 2003.

[16] S. Cotter, B. Rao, K. Engan, and K. Kreutz Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[17] K. Herrity, A. Gilbert, and J. Tropp, "Sparse approximation via iterative thresholding," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[18] T. Blumensath and M. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 298 – 309, April 2010.

[19] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[20] Y. Lu and M. Do, "A theory for sampling signals from a union of subspaces," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2334 –2345, june 2008.

[21] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872 –1882, april 2009.

[22] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Interscience, 1991.

[23] E. Candes, L. Demanet, D. Donoho, and L. Ying, "Fast discrete curvelet transforms," *Multiscale Model. Simul.*, vol. 5, pp. 861–899, 2006.

[24] M. Aharon, E. Elad, and A. Bruckstein, "K-SVD: an algorithm for designining of overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

Mike E. Davies (M00-SM11) received the B.A. (Hons.) degree in engineering from Cambridge University, Cambridge, U.K., in 1989 and the Ph.D. degree in nonlinear dynamics and signal processing from University College London (UCL), London, U.K., in 1993. He was awarded a Royal Society University Research Fellowship in 1993 which he held rst at UCL and then the University of Cambridge. He acted as an Associate Editor for IEEE TRANSACTIONS IN SPEECH, LANGUAGE AND AUDIO PROCESSING, between 20032007. Since 2006, Dr. Davies has been with the University of Edinburgh where he holds the Jeffrey Collins SFC funded chair in Signal and Image Processing.

He is currently Head of the Institute for Digital Communications and the Director of the Joint Research Institute in Signal and Image Processing, part of the Edinburgh Research Partnership in Engineering and mathematics. His current research focus is on sparse approximation, computational harmonic analysis, compressed sensing and their applications within signal processing. His other research interests include nonlinear and non-Gaussian signal processing, high-dimensional statistics, and information theory.

Mehrdad Yaghoobi (S'98-M'09) received the BSc. and MSc. in electrical and biomedical engineering in 1999 and 2002 from the University of Tehran and Sharif University of Technology, Iran, respectively. He received his PhD from the University of Edinburgh, UK, in 2010 in Signal Processing. He has been a research fellow (post-doc) since then in the Institute for Digital Communications (ID-Com) at the University of Edinburgh. His research has been funded by an EU FP7, FET-Open grant "SMALL", the EPSRC platform grant on "Sensor Signal Processing" and program grant "Signal Processing for the Networked Battlespace". He has served as a reviewer for many IEEE Transactions.

His current research interests include low-dimensional signal modelling, compressed sensing, analog to information conversion, inverse problems and super-resolution sensing.

Laurent Daudet (M04-SM10) studied at the Ecole Normale Suprieure in Paris, where he graduated in statistical and non-linear physics. In 2000, he received a PhD in mathematical modeling from the Universit de Provence, Marseille, France. After a Marie Curie post-doctoral fellowship at the C4DM, Queen Mary University of London, UK, he worked as associate professor at UPMC (Paris 6 University) in the Musical Acoustics Lab. He is now Professor at Paris Diderot University Paris 7, with research at the Langevin Institute for Waves and Images, where he currently holds a joint position with the Institut Universitaire de France. Laurent Daudet is author or co-author of over 150 publications (journal papers or conference proceedings) on various aspects of acoustics and audio signal processing, in particular using sparse representations.