Analysis SimCO Algorithms for Sparse Analysis Model Based Dictionary Learning

Jing Dong, Wenwu Wang, Senior Member, IEEE, Wei Dai, Member, IEEE, Mark D. Plumbley, Fellow, IEEE, Zi-Fa Han, and Jonathon Chambers, Fellow, IEEE

Abstract—In this paper, we consider the dictionary learning problem for the sparse analysis model. A novel algorithm is proposed by adapting the simultaneous codeword optimization (SimCO) algorithm, based on the sparse synthesis model, to the sparse analysis model. This algorithm assumes that the analysis dictionary contains unit $\ell_2$-norm atoms and learns the dictionary by optimization on manifolds. This framework allows multiple dictionary atoms to be updated simultaneously in each iteration. However, similar to several existing analysis dictionary learning algorithms, dictionaries learned by the proposed algorithm may contain similar atoms, leading to a degenerate (coherent) dictionary. To address this problem, we also consider restricting the coherence of the learned dictionary and propose Incoherent Analysis SimCO by introducing an atom decorrelation step following the update of the dictionary. We demonstrate the competitive performance of the proposed algorithms using experiments with synthetic data and image denoising as compared with existing algorithms.

Index Terms—Sparse representation, analysis model, SimCO, analysis dictionary learning.

I. INTRODUCTION

MANY problems in signal processing can be regarded as inverse problems, for example, denoising [1], inpainting [2] and super-resolution [3]. These problems aim to reconstruct original signals from their observed measurements. Some prior knowledge or assumptions about the signals are required due to the lack of information or the presence of noise in the observations. One assumption that has attracted extensive attention in the past decade is that the signals to be restored are sparse in some domain. Two signal models to capture the sparse property of the signals have been proposed, namely, the sparse synthesis model [4] and sparse analysis model [5], [6]. More recently, the sparse analysis model has been extended to a more generalized model, referred to as the sparsifying transform model [7].

J. Dong, W. Wang and M. D. Plumbley are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (emails: {j.dong, w.wang, m.plumbley}@surrey.ac.uk).

W. Dai is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (email: wei.dai1@imperial.ac.uk).

Z.-F. Han is with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. (email: zifahan@gmail.com).

J. Chambers is with the School of Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K. (email: Jonathon.Chambers@newcastle.ac.uk).

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K013070/1 and the MOD University Defence Research Collaboration in Signal Processing.

The authors confirm that all data underlying the findings are fully available without restriction. Details of the data and how to get access are available from the University of Surrey: http://dx.doi.org/10.15126/surreydata.00808101.

A. Sparse Synthesis Model

The most well-known model in sparse representation is the sparse synthesis model [4], [8], [9], [10]. This model assumes that a signal $y \in \mathbb{R}^m$ can be linearly represented with some atoms (columns) of a synthesis dictionary $D \in \mathbb{R}^{m \times d}$, where the dictionary is usually overcomplete with $d > m$. The number of atoms used to represent $y$ is much smaller than the total number of atoms in the dictionary, which reflects the sparsity of the signal $y$. Mathematically, this model can be written as $y = Da$ with $\|a\|_0 = s$, where the $\ell_0$-norm $\| \cdot \|_0$ counts the number of non-zero elements of its argument, and $a$ is the representation coefficient vector with $s$ being its sparsity.

The atoms corresponding to the non-zero elements of $a$ are used to synthesize the signal $y$ via their linear combination, which brings about the term “synthesis” in the name of this model.

One challenge related to this model is the sparse coding problem which aims to find the sparsest representation of a given signal $y$ with respect to a given dictionary $D$. In order to tackle this problem, greedy algorithms have been proposed, such as matching pursuit (MP) [11], orthogonal matching pursuit (OMP) [12], stagewise orthogonal matching pursuit (StOMP) [13] and subspace pursuit (SP) [14], as well as relaxation methods such as basis pursuit (BP) [15] and focal underdetermined system solver (FOCUSS) [16].

A second challenge is to design or learn an appropriate dictionary $D$ to represent a set of signals as sparsely as possible. Many analytical dictionaries have been developed, but dictionaries learned from a set of training signals have the potential to fit these signals better than the analytical dictionaries [4]. As a result, the dictionary learning problem for the sparse synthesis model has become one of the most popular topics in sparse representation. This problem aims to seek the dictionary $D$ that leads to the best set of representations for a given set of training signals. Many algorithms have been proposed to address this problem, for example, method of optimal directions (MOD) [9], K-SVD [4] and SimCO [10]. These algorithms typically alternate iteratively between an update of the coefficients and an update of the dictionary. For the update of the coefficients, sparse coding algorithms are often used while keeping the dictionary fixed. The main difference between synthesis dictionary learning algorithms is the way in which the dictionary is updated.

B. Sparse Analysis Model and Sparsifying Transform Model

In contrast to the synthesis model, the sparse analysis model uses an analysis dictionary $\Omega \in \mathbb{R}^{p \times m}$ with $p > m$ to
“analyze” the signal $y \in \mathbb{R}^m$. Specifically, it assumes that the product of $\Omega$ and $y$ is sparse, i.e. $x = \Omega y$ with $\|x\|_0 = p - l$, where $0 \leq l \leq p$ is the number of zeros in $x \in \mathbb{R}^p$. The matrix $\Omega$ is usually referred to as an analysis dictionary [17] or analysis operator [18], [19], with each row of $\Omega$ being an atom. The vector $x$ is the analysis representation of the signal $y$ with respect to $\Omega$. This model is also referred to as a co-sparse analysis model, and the number of zeros $l$ is called the co-sparsity of the signal $y$ with respect to $\Omega$ [6].

Let $\Lambda = \{i : x_i = 0\}$ denote the index set of the rows in $\Omega$ corresponding to the zero elements in $x$ (thus, $\text{card}(\Lambda) = l$) and let $\Omega_\Lambda$ denote the sub-matrix of $\Omega$ containing only the rows indexed by $\Lambda$. The set $\Lambda$ is called the co-support of $y$. For the analysis model, we have $\Omega_\Lambda y = 0$, meaning that the $l$ atoms indexed by $\Lambda$ are orthogonal to the signal $y$. From the subspace point of view, $y$ lies in the subspace which is orthogonal to the subspace spanned by the rows of $\Omega_\Lambda$. Even though the description of the sparse analysis model may seem similar to its synthesis counterpart, these two models differ significantly if the dictionaries are overcomplete [5].

If the signal $y$ is known, its analysis representation with respect to a given $\Omega$ can be obtained via multiplying $y$ by $\Omega$. However, when the observed signal is contaminated by noise, the clean signal $y$ has to be estimated first in order to get its analysis representation, which leads to the analysis pursuit problem [17]. Some algorithms like backward-greedy (BG) [17], optimized-backward-greedy (OBG) [17], and greedy analysis pursuit (GAP) [6] have been proposed to address this problem.

In a similar way to the dictionary in the synthesis model, the analysis dictionary $\Omega$ also plays an important role in the analysis representation of the signal $y$, and the dictionaries learned from a set of training signals show some advantages compared with pre-defined dictionaries [17]. In the past few years, the analysis dictionary learning (ADL) problem has begun to attract more attention [17], [18], [19]. In this paper, we focus on this ADL problem.

Recently, the so-called sparsifying transform model, which assumes that a signal can be approximately sparsified with an analysis transform operator, was introduced in [7]. This model can be regarded as a natural extension of the sparse analysis model. Learning a sparsifying transform has been investigated in [7], [20], [21]. These algorithms deal with the sparsification error in the transform domain rather than in the original signal domain as in the ADL algorithms [17], [18], by applying the transform operator to the training signals even if the signals contain noise. In the present paper, we intentionally ignore this subtle difference between these two formulations and regard the sparsifying transform learning algorithms as alternative solvers of the ADL problem, since the results of sparsifying transform learning can be regarded as dictionaries for the sparse analysis model.

C. Analysis Dictionary Learning

Several algorithms have been proposed for the ADL problem. The Analysis K-SVD algorithm [17] assumes that the training samples are noisy signals and minimizes the error between the training samples and the signals estimated using the learned dictionary. It applies the OBG [17] analysis pursuit algorithm to detect the co-support of each training signal with respect to an initial dictionary, and employs the singular value decomposition (SVD) to update the dictionary atoms one-by-one. After the update of all atoms, similar atoms, determined with inner-product of two atoms, are replaced by new randomly generated atoms. However, these new atoms cannot preserve the information of the atoms to be replaced because of the randomness. Besides, the computational complexity of Analysis K-SVD is quite high due to the involvement of the analysis pursuit problem [17].

The learning overcomplete sparsifying transforms (LOST) algorithm [20] minimizes the so-called sparsifying error which is defined in the transform or analysis domain rather than the original signal domain as in the formulation of Analysis K-SVD. As a result, the time consuming algorithm OBG is not used anymore. Two penalty terms are added to the objective function of LOST to apply two constraints on the learned dictionary respectively, i.e. the full column rank constraint and the constraint on the correlation between the atoms. The coefficients of the penalty terms play an important role in the performance of LOST, but selecting proper coefficients is a practical challenge [20].

Transform K-SVD proposed recently in [21] combines the sparsifying error formulation of LOST with the dictionary update approach of Analysis K-SVD. This algorithm uses the same method as in Analysis K-SVD to avoid similar atoms, but overcomes its computational complexity issue with a formulation used in LOST. We have found that Transform K-SVD performs well in recovering a reference dictionary, but its denoising performance is relatively limited, as shown in our simulations (see Section VI later). We will see that the denoising performance of (NA)AOL is also limited when the noise level is relatively high.

The analysis operator learning (AOL) algorithm reported in [18] addresses the ADL problem using a constrained optimization framework. In this algorithm, the $\ell_1$-norm is used as the co-sparsity measurement. It restricts the dictionary to be a uniform normalized tight frame (UNTF) which is the intersection of uniform normalized (UN) frames manifold and tight frames (TF) manifold. The AOL algorithm learns a dictionary by combining an $\ell_1$ optimization framework with projection of the dictionary onto the UNTF set. The algorithms (NL)AOL (noiseless AOL) and (NA)AOL (noise-aware AOL) have been developed for learning with noiseless and noisy training samples respectively. However, random dictionaries cannot be recovered by the AOL algorithms since the UNTF constraint limits the possible dictionaries to be learned (see Section VI later). We will see that the denoising performance of (NA)AOL is also limited when the noise level is relatively high.

The GeOmetric Analysis operator Learning (GOAL) algorithm [19] employs the $\ell_p$-norm ($0 \leq p \leq 1$) as the co-sparsity measurement, which is different from the algorithms reviewed above. Similar to the LOST algorithm, the objective function of GOAL also incorporates two additional penalty terms to address the full rank and the correlation constraints, leading to the difficulty of setting the weights for the penalties. The conjugate gradient method on manifolds [22] is applied for the optimization of the cost function in GOAL.
D. Contributions

In this work, we propose two new algorithms which can partly address the limitations of the ADL algorithms mentioned above. Firstly, we adapt the synthesis model based SimCO algorithm [10] to the analysis model and develop a new ADL algorithm which is referred to as the Analysis SimCO algorithm. In SimCO, the optimization method on manifolds is applied to update multiple dictionary atoms simultaneously, leading to a better performance compared with K-SVD where the atoms are updated one-by-one. Thus, we adapt the framework of SimCO to the ADL problem to enable the simultaneous update of multiple atoms via the optimization on manifolds. The preliminary results of this work have been presented in [23]. This dictionary update method is different from the methods used by the existing algorithms. Analysis K-SVD and Transform K-SVD only allow one atom to be updated in one iteration. In LOST, the dictionary is updated as a whole matrix by the standard conjugate gradient method. Compared with the AOL algorithms, the updated dictionary in our proposed method is more general without projection onto the UNTF set. Notice that the GOAL algorithm also employs our proposed method is different from that of GOAL due to the different co-sparsity measure based on $\ell_0$-norm, and the fewer penalty terms used. Besides, our proposed algorithm employs the gradient descent method on manifolds rather than the conjugate gradient method as in GOAL.

Secondly, we propose the Incoherent Analysis SimCO algorithm to avoid similar atoms appearing in the dictionaries learned by Analysis SimCO. In the Incoherent Analysis SimCO algorithm, a constraint restricting the correlations of two distinct atoms of the dictionary is considered and an atom decorrelation step is applied to enforce this constraint by rotating the highly-correlated atom pairs. In this way, the correlation of any two distinct atoms can be restricted to be below a given threshold explicitly. Compared with the methods used in existing ADL algorithms to avoid similar atoms, the decorrelation step applied in the Incoherent Analysis SimCO algorithm has some advantages. For example, this method avoids the coefficient selection problem of LOST since the constraint is tackled directly rather than applied as a penalty term of the objective function. Besides, the new atoms obtained by the decorrelation step are more likely to be closer to the atoms replaced than the atoms that are generated randomly in Analysis K-SVD and Transform K-SVD.

E. Notations

Bold capital letters are used to represent matrices. In particular, $I$ denotes the identity matrix whose dimension can be decided from the context. The notation $X_{i,:}$ is used to specify the $i$th row of the matrix $X$ and $X_{:,j}$ represents its $j$th column. Bold lowercase letters represent vectors. Scalars are either capital or lowercase letters. The norms $\| \cdot \|_2$ and $\| \cdot \|_F$ denote the $\ell_2$-norm and the Frobenius norm respectively. The notation $\| \cdot \|$ returns the absolute value of a scalar. The notation $\langle \cdot, \cdot \rangle$ is used to represent the canonical inner-product of two vectors.

F. Organization of the Paper

The remainder of the paper is organized as follows. In Section II the original SimCO algorithm is reviewed. In Section III we present our formulation for the ADL problem and the optimization framework. More details of our proposed Analysis SimCO algorithm and the discussions of its convergence and computational complexity are provided in Section IV. Section V introduces Incoherent Analysis SimCO where an atom decorrelation step is involved. Section VI provides experimental results of learning dictionaries with synthetic data and for image denoising. Conclusions are drawn in Section VII.

II. THE SIMCO ALGORITHM

The SimCO algorithm [10] was proposed to learn a synthesis dictionary from a set of signals so that the signals can each be represented by a few atoms of the dictionary. Let $Y \in \mathbb{R}^{m \times n}$ denote the matrix of the training signals, where each column of $Y$ is one training signal. In SimCO, the dictionary learning problem is formulated as

$$\arg \min_{D \in \mathcal{D}} f(D) = \arg \min_{D \in \mathcal{D}} \min_{A \in \mathcal{A}} \| Y - DA \|_F^2, \quad (1)$$

where the columns of $A \in \mathbb{R}^{d \times n}$ are the representation coefficient vectors and $D \in \mathbb{R}^{m \times d}$ is the dictionary to be learned. In this formulation, $D$ is assumed to contain unit $\ell_2$-norm columns, which is addressed by the constraint $D \in \mathcal{D}$ with $\mathcal{D}$ representing the set of all matrices that contain unit $\ell_2$-norm columns. The positions of the non-zero elements of the coefficient matrix $A$ are fixed, achieved with the constraint $A \in \mathcal{A}$.

To solve the optimization problem (1), SimCO follows the conventional two-stage optimization process – a sparse coding stage and a dictionary update stage. The sparse coding stage determines the sparse representations $A$ of the signals in $Y$ for a given dictionary $D$. Various sparse coding algorithms such as OMP [8] can be employed in this stage.

In the dictionary update stage, SimCO applies optimization methods on manifolds [24] to update the dictionary $D$ under the unit $\ell_2$-norm constraints on the columns of $D$. According to the updated $D$, the coefficient matrix $A$ is also updated, while the positions of the non-zero elements are kept unchanged. This framework is able to update multiple atoms and the corresponding coefficients simultaneously, which gives rise to the term simultaneous codeword optimization (SimCO).

III. PROBLEM FORMULATION AND OPTIMIZATION FRAMEWORK

Given a set of training signals $Y \in \mathbb{R}^{m \times n}$, the ADL problem can be written as [7]

$$\{ \Omega^*, X^* \} = \arg \min_{(\Omega, X)} \| X - \Omega Y \|_F^2,$$

\[ \text{s.t. } \| X_{:,i} \|_0 = p - l, \ \forall i. \quad (2) \]

This is a general formulation without any additional constraint on $\Omega$ apart from the co-sparsity constraints $\| X_{:,i} \|_0 = p - l, \ \forall i.$
However, this formulation has ambiguities caused by scaling. In one case, when the training data \( Y \) admits exact sparse representations, there exists a dictionary \( \Omega \) with which the analysis representations of \( Y \), i.e. \( X = \Omega Y \), satisfy the co-sparsity constraints. If the dictionary \( \Omega \) is scaled by multiplying a scalar \( c \in \mathbb{R} \), the corresponding representations \( c \cdot X = c \cdot \Omega Y \) will also satisfy the constraints. Thus, the problem (2) has infinite optimal solutions \( c \cdot \Omega \) and \( c \cdot X \). This may introduce difficulty in optimization. In the other case, if the data \( Y \) admits approximation representations and \( \| X - \Omega Y \|_F^2 = \delta \), the value of the cost function with scaled \( X \) and \( \Omega \), i.e. \( \| X - \Omega Y \|_F^2 = c^2 \cdot \delta \), can be arbitrarily small. In other words, the cost function is unbounded from below, which makes it impossible to find an optimal solution. In addition, (2) has trivial solutions where \( \Omega \) contains all-zero rows.

In order to avoid these problems, we apply the unit \( \ell_2 \)-norm constraints on the rows of \( \Omega \), leading to the following formulation of the ADL problem

\[
\{ \Omega^*, X^* \} = \arg \min_{\{ \Omega, X \}} \| X - \Omega Y \|_F^2, \\
\text{s.t. } \| X_{i,:}\|_0 = p - l, \quad \forall i \\
\| \Omega_j,:\|_2 = 1, \quad \forall j.
\]

(3)

The unit \( \ell_2 \)-norm constraints on the rows of \( \Omega \) are able to eliminate the scaling ambiguity mentioned above. Besides, the trivial solutions where \( \Omega \) has zero rows can be excluded. The formulation (3) is different from that of Analysis K-SVD [17] which minimizes the error in the signal domain. It also differs from the objective function of LOST [20] where the penalty terms as described earlier in Section I are included.

The problem (3) can be addressed by an optimization framework alternating between two stages: analysis sparse coding and dictionary update. Given a dictionary \( \Omega \), the first stage finds \( X \) satisfying the co-sparsity constraints \( \| X_{i,:}\|_0 = p - l, \quad \forall i \). In the dictionary update stage, \( \Omega \) is updated assuming known and fixed \( X \) obtained in the first stage.

Here we attempt to update the dictionary using a similar method as in SimCO and refer to our proposed algorithm as Analysis SimCO. The optimization framework of Analysis SimCO is presented in Algorithm 1. In our original algorithm SimCO, the use of the term “simultaneous” comes from the following two facts: (1) multiple dictionary atoms are updated simultaneously, and (2) their corresponding coefficients are also updated simultaneously with these atoms. In the analysis case, we borrow the term “SimCO” mainly because in the proposed algorithm the dictionary atoms are updated simultaneously.

A common problem with the popular analysis dictionary learning algorithms, such as Analysis K-SVD [17], is that the learned dictionary \( \Omega \) may contain similar atoms. Such a dictionary is regarded as a degenerate solution [7], [21]. This issue is also observed in the dictionary learned from (3) with the Analysis SimCO algorithm, as will be shown in Section VI. Thus, we develop an extended version of Analysis SimCO to avoid this kind of degenerate dictionary, which will be presented in Section V in detail.

### Algorithm 1 Optimization Framework of Analysis SimCO

**Input:** \( Y, p, l \)

**Output:** \( \Omega^* \)

**Initialization:**

Initialize the iteration counter \( k = 1 \) and the analysis dictionary \( \Omega^{(k)} \). Perform the following steps.

**Main Iterations:**

1. Analysis sparse coding: Compute the representations \( X^{(k)} \) with the fixed dictionary \( \Omega^{(k)} \) and the training signals in \( Y \).
2. Dictionary update: Update the dictionary \( \Omega^{(k+1)} \leftarrow \Omega^{(k)} \).
3. If the stopping criterion is satisfied, \( \Omega^* = \Omega^{(k+1)} \) and quit the iteration. Otherwise, increase the iteration counter \( k = k + 1 \) and go back to step 1).

IV. **ANALYSIS SIMCO ALGORITHM**

As the dictionary update stage in our algorithm is based on optimization on matrix manifolds, we begin this section with a brief introduction to the optimization on matrix manifolds to make this paper self-contained. The details of the analysis sparse coding and dictionary update are then presented respectively, followed by the convergence and computational complexity analysis of our proposed algorithm.

A. **Optimization on Matrix Manifolds**

The Stiefel manifold \( St(p, m) \) is defined as \( St(p, m) := \{ U \in \mathbb{R}^{m \times p} : U^T U = I \} \) [24, pp. 26]. For \( p = 1 \), the Stiefel manifold \( St(p, m) \) reduces to the unit sphere, i.e. \( S = \{ u \in \mathbb{R}^m : u^T u = 1 \} \). At each point \( u \in S \), there exists a tangent space \( T_u S \) which consists of all vectors orthogonal to \( u \) in \( \mathbb{R}^m \), i.e. \( T_u S = \{ v \in \mathbb{R}^m : u^T v = 0 \} \). The vectors in \( T_u S \) are tangent vectors to \( S \) at the point \( u \). The tangent space \( T_u S \) can be regarded as a vector space approximation of the manifold \( S \) at the point \( u \) [24, pp. 34].

Before dealing with the optimization problem on manifolds, we consider a more general class of problems, i.e., the unconstrained optimization problem, from which the optimization methods on matrix manifolds can be adapted,

\[
\min_u f(u),
\]

where \( u \in \mathbb{R}^m \) and \( f : \mathbb{R}^m \to \mathbb{R} \) is a differentiable function. This problem can be addressed by the standard line search method. In the \( k \)th iteration, the standard line search method selects a descent direction \( p \) along which the current point \( u_k \) is moved to a new point \( u_{k+1} \) leading to a smaller or equal objective function value, i.e.

\[
u_{k+1} = u_k + \alpha \cdot p
\]

with \( f(u_{k+1}) \leq f(u_k) \). Here \( \alpha \) is the scalar step size which can be selected carefully to guarantee the reduction of the cost function [25]. In order to determine the search direction \( p \), the value and the derivatives of the objective function can be used. The most obvious choice is the steepest descent direction \( p_k = -\nabla f(u_k) \) along which the objective function value decreases most rapidly among all the directions [25, pp. 20].
Now we consider the optimization problem where the variable $u$ is restricted on the manifold $S$, i.e.

$$\min_{u \in S} f(u). \quad (6)$$

Analogous line search methods on manifolds have been developed by generalizing the standard line search methods for the unconstrained optimization problem (4). Specifically, in the $k$th iteration, the search direction $q_k$ should be chosen as a tangent vector to $S$ at $u_k$, i.e., $q_k \in T_{u_k}S$. Thus the search direction $q_k$ is the projection of the search direction $p_k$ of the unconstrained optimization methods to the tangent space $T_{u_k}S$ [24, pp. 49], that is

$$q_k = (I - u_ku_k^T)p_k. \quad (7)$$

The new point $u_{k+1}$ obtained by moving $u_k$ in the direction of $q_k$ should stay on $S$. As a result, the line search path (5) is replaced by a curve on $S$ [24, pp. 103], i.e.

$$u_{k+1} = u_k \cos(\|q_k\|_2) + \frac{q_k}{\|q_k\|_2} \sin(\|q_k\|_2). \quad (8)$$

B. Analysis Sparse Coding Stage

The purpose of the analysis sparse coding stage is to get the sparse representations $X$ of the training signals in $Y$ based on a given dictionary $\Omega$. Unlike the corresponding problem of the synthesis model, here the exact representations $X$ can be calculated directly by simply multiplying the signals in $Y$ by the dictionary $\Omega$, that is

$$X = \Omega Y. \quad (9)$$

Since the initial dictionary is an arbitrary one, the representations obtained in this way may not satisfy the co-sparisty constraints on $X$ in (3). A hard thresholding operation is therefore applied to enforce the co-sparisty

$$\hat{X} = HT_l(X), \quad (10)$$

where $HT_l(X)$ is the non-linear operator that sets the smallest $l$ elements (in magnitude) of each column of $X$ to zero. The representations $\hat{X}$ obtained via equation (10) are the best approximation of the exact representations $X$ in terms of the error in Frobenius norm among all the matrices satisfying the co-sparisty constraints.

C. Dictionary Update Stage

The dictionary update stage aims at optimizing the following problem (by fixing $\hat{X}$ in (3))

$$\arg \min_{\Omega} f(\Omega) = \|X - \Omega Y\|_F^2 \quad \text{s.t.} \quad \|\Omega_{j,:}\|_2 = 1, \forall j. \quad (11)$$

The cost function can be rewritten as a function of the rows of $\Omega$. Besides, the constraint that $\Omega$ only contains unit $\ell_2$-norm rows restricts the transposes of the rows of $\Omega$ to lie on the unit sphere $S$, i.e. $\Omega_{j,:}^T \in S, \forall j$. Thus the problem (11) can be rewritten as

$$\arg \min_{\Omega} f(\Omega) = \sum_{j=1}^p \|X_{j,:} - \Omega_{j,:}Y\|_2^2 \quad \text{s.t.} \quad \Omega_{j,:}^T \in S, \forall j. \quad (12)$$

As a result, the “line” search methods on manifolds can be utilized in this stage. Here we use the first order optimization procedures as in SimCO [10], i.e. the gradient descent line search method. We explain below the key points of this method including search direction, line search path, and step size respectively. The dictionary update stage is summarized in Algorithm 2.

Algorithm 2 Dictionary Update Stage

Input: $\Omega^{(k)}$, $X^{(k)}$, $Y$
Output: $\Omega^{(k+1)}$

Main Steps:

1) Calculate the search direction, based on equations (13) and (14).
2) Find a proper step size $\alpha$ using golden section search.
3) Update the dictionary $\Omega^{(k+1)} \leftarrow \Omega^{(k)}$, based on equation (15).

1) Search direction: We use the steepest descent direction as the search direction, i.e. the negative gradient of the objective function with respect to $\Omega$ as follows

$$H = -\nabla f(\Omega)$$

$$= -\frac{\partial\|X - \Omega Y\|_F^2}{\partial \Omega}$$

$$= 2XY^T - 2\Omega YY^T. \quad (13)$$

2) Line search path: The search direction of the $j$th row of $\Omega$, i.e. the projection of each row of $H$ onto the tangent space of $S$, is [24, pp. 49]

$$h_j = H_j.(I - \Omega_{j,:}^T \Omega_{j,:}). \quad (14)$$

According to equation (8), the line search path for the $j$th row of $\Omega$ can be written as

$$\Omega_{j,:}(\alpha) = \begin{cases} 
\Omega_{j,:} & \text{if } \|h_j\|_2 = 0,
\Omega_{j,:} \cos(\|h_j\|_2) + (\frac{h_j}{\|h_j\|_2}) \sin(\|h_j\|_2) & \text{otherwise}
\end{cases} \quad (15)$$

where $\alpha$ is the step size.

3) Step size: In order to find a proper step size $\alpha$, we apply the golden section search method [10]. This method consists of two stages. In the first stage, it finds a range which contains a local minimum and within which the objective function is unimodal. In the second stage, the golden section ratio is used to successively narrow the range until the minimizer is located and thus $\alpha$ is determined.

D. Convergence

Our proposed algorithm alternates between the analysis sparse coding stage and the dictionary update stage. For a fixed dictionary $\Omega$, $X$ obtained in the analysis sparse coding stage is the optimal solution under the constraint of co-sparisty. Thus, the cost function can only decrease in this stage. In the dictionary update stage, since the update of $\Omega$ is along a descent direction and the step size is chosen to guarantee that the updated $\Omega$ will not increase the cost function. Thus, the cost function is decreasing monotonically in our proposed
algorithm. In addition, the cost function of our formulation (3) is lower bounded by zero, i.e. $\|X - \Omega Y\|_F^2 \geq 0$. According to the monotone convergence theorem [26], given the cost function decreases monotonically and is lower bounded, the algorithm must converge. The convergence will also be demonstrated experimentally in Section VI-A.

### E. Computational Complexity

The time complexity of the AnalysisSimCO algorithm can be analyzed as follows. The time complexity of the sparse coding stage is dominated by the calculation of $\Omega Y$, at $O(pmn)$, in terms of the analysis in [20]. In the dictionary update stage, the calculation of $H$ is the dominant part. Computing the product $XY^T$ requires $O(pmn)$ operations. The time complexity of $\Omega YY^T$ is $O(pm^2)$ with pre-computed $YY^T$. As a result, the dictionary update stage requires $O(pmn)$ operations with the usual case $n > m$. The total time complexity of each iteration of the Analysis SimCO algorithm thus scales as $O(pmn)$.

The computational complexity of Analysis SimCO, similar to those of LOST [20], (NL)AOL [18], and Transform K-SVD [21], shows a reduction compared with those of Analysis K-SVD and (NA)AOL. The complexity of Analysis K-SVD is $O(pm^2n)$ using BG or $O(pm^2n)$ using OBG, and (NA)AOL requires $O(pmnk)$ operations with $k$ being the number of dictionary update per iteration. The running time of these algorithms in practice will be given in Section VI.

### V. INCOHERENT ANALYSIS SIMCO

As mentioned in Section III, dictionaries learned by the existing ADL algorithms may contain similar atoms, which can degrade the representation performance for signal recovery. To address this problem, several methods have been proposed. For example, in Analysis K-SVD and Transform K-SVD, the similar atoms are replaced by randomly generated atoms, as mentioned in Section I. In LOST [20] and GOAL [19], a penalty term is used in the objective function to restrict the correlations between atoms. As will be observed in the experiments of Section VI, Analysis SimCo has the same issue, where some of the atoms in the learned dictionary may appear similar. Here, we present an alternative solution to this problem based on [27]. The method in [27] was developed to mitigate the correlations between atoms learned by a synthesis model. Here we adapt this method to our model and optimization problem.

In the context of the sparse synthesis model, the coherence of the dictionary has been defined as a measure of the similarities between the atoms [28]. We extend this definition for an analysis dictionary $\Omega$ and define the coherence $\mu(\Omega)$ in a row-wise way as

$$\mu(\Omega) = \max_{\forall i,j,i \neq j} \left( \frac{|\Omega_{i,:}, \Omega_{j,:}|}{\|\Omega_{i,:}\|_2 \|\Omega_{j,:}\|_2} \right),$$

From this definition, we have $0 \leq \mu(\Omega) \leq 1$. With the unit $\ell_2$-norm constraints on the rows of $\Omega$, the coherence $\mu(\Omega)$ can be simplified as

$$\mu(\Omega) = \max_{\forall i,j,i \neq j} |\Omega_{i,:}, \Omega_{j,:}|.$$  

The coherence $\mu(\Omega)$ reflects the maximum correlation of two distinct atoms in $\Omega$. If $\mu(\Omega)$ is close to 1, it means that there are very similar rows in $\Omega$, which is the case we attempt to avoid. Thus, we add a coherence constraint $\mu(\Omega) \leq \mu_0$ to the formulation (3), i.e.

$$\{\Omega^*, X^*\} = \arg \min_{\Omega,X} \|X - \Omega Y\|_F^2$$

$$\text{s.t. } \|X_{i,:}\|_0 = p - l, \forall i$$

$$\|\Omega_{j,:}\|_2 = 1, \forall j$$

$$\mu(\Omega) \leq \mu_0,$$

where $\mu_0$ is the coherence limit for the learned dictionary $\Omega$.

To enforce the incoherence constraint, we add an extra step in the dictionary update stage, aiming to find the closest dictionary $\Omega_0$ to $\Omega$ in Frobenius norm, with the coherence of the dictionary $\hat{\Omega}$ bounded by a threshold $\mu_0$, that is

$$\arg \min_{\Omega} \|\hat{\Omega} - \Omega\|_F^2$$

$$\text{s.t. } \|\hat{\Omega}_{i,:}\|_2 = 1, \forall i$$

$$\mu(\hat{\Omega}) \leq \mu_0.$$  

Here the unit $\ell_2$-norm constraints for the atoms in the dictionary are also applied to ensure that the transposes of the atoms in the output dictionary are still on the manifold. This problem is addressed by applying the decorrelation method [27] in a row-wise fashion, as presented in Algorithm 3. The general idea is to determine the atom pairs whose correlations are greater than $\mu_0$, via a labeling process (from line 5 to line 9 of Algorithm 3), and decorrelate these atom pairs, via a decorrelation process (from line 10 to line 20). This method keeps alternating between the two processes until the coherence of the estimated dictionary $\hat{\Omega}$ reaches the threshold $\mu_0$. Although this is a heuristic algorithm, it typically involves only a few loops to output an incoherent dictionary. The convergence and the effectiveness of this algorithm will be numerically demonstrated in Section VI.

In the labeling process, the atoms of $\Omega$ are labeled as either the atom pairs to be decorrelated or atoms that do not need to be modified. An index-pair set $F$ is used to store the index pairs of atom pairs labeled to be decorrelated and an index set $E$ is employed to save the indices of the remaining atoms. $\Omega_E$ represents the submatrix of $\Omega$ only containing the rows indexed by the set $E$. In each iteration, the correlations of any two distinct rows belonging to $\Omega_E$ are calculated to determine the most correlated pair. The indices of these two atoms will be saved, as an index-pair, into the set $F$, i.e., $F \leftarrow F \cup \{(i, j)\}$, indicating that these two atoms are labeled as an atom pair to be decorrelated in the following decorrelation process. Their indices will be removed from $E$ to avoid being detected again, i.e. $E \leftarrow E \setminus \{i, j\}$.

In the decorrelation process, the atom pairs indexed by the members of $F$ are decorrelated successively. The decorrelation of each atom pair is achieved by rotating the two atoms symmetrically with respect to their mean so that their correlation reaches $\mu_0$ [27]. The rotated atoms are determined based on the orthonormal basis $\{b_1, b_2\}$ developed using the atoms to be decorrelated (lines 11 and 12) and the angle $\theta$ determined
Algorithm 3 Atom Decorrelation Step

1: **Input:** $\Omega$, $\mu_0$
2: **Output:** $\Omega$
3: **Initialization:**
   $$\Omega = \Omega_0, \theta = \frac{1}{2}\arccos \mu_0, c_1 = \cos \theta, c_2 = \sin \theta$$
4: **while** $\mu(\Omega) > \mu_0$ **do**
5:   **if** $E = \{1, 2, ..., p\}$ **then**
6:     $F_0 = \emptyset$
7:     **for** $(i, j) \in F$ **do**
8:       $b_1 = (\Omega_{i,:} + \Omega_{j,:})/(\|\Omega_{i,:} + \Omega_{j,:}\|_2)$
9:       $b_2 = (\Omega_{i,:} - \Omega_{j,:})/(\|\Omega_{i,:} - \Omega_{j,:}\|_2)$
10:      **if** $(\Omega_{i,:}, \Omega_{j,:}) > 0$ **then**
11:        $\Omega_{i,:} = c_1 b_1 + c_2 b_2$
12:        $\Omega_{j,:} = c_1 b_1 - c_2 b_2$
13:      **else**
14:        $\Omega_{i,:} = c_1 b_2 + c_2 b_1$
15:        $\Omega_{j,:} = -c_1 b_2 + c_2 b_1$
16: **end if**
17: **end for**
18: **end if**
19: **end while**
20: **return** $\Omega$

Algorithm 4 Incoherent Analysis SimCO

**Input:** $Y$, $p$, $l$, $\mu_0$
**Output:** $\Omega^*$

**Initialization:**
- Initialize the iteration counter $k = 1$ and the analysis dictionary $\Omega^{(k)}$. Perform the following steps.

**Main Iterations:**
1. **Analysis sparse coding:** Compute the representations $X^{(k)}$ with the fixed dictionary $\Omega^{(k)}$ and the training signals in $Y$, based on equations (9) and (10).
2. **Dictionary update:** Update the dictionary $\Omega^{(k+1)} \leftarrow \Omega^{(k)}$, using Algorithm 2.
3. **Atom decorrelation:** Decorrelate the atoms $\Omega^{(k+1)} \leftarrow \Omega^{(k+1)}$, using Algorithm 3.
4. If the stopping criterion is satisfied, $\Omega^* = \hat{\Omega}^{(k+1)}$, quit the iteration. Otherwise, increase the iteration counter $k = k + 1$ and go back to step 1).

SimCO-Random (ASimCO-Random) and Analysis SimCO-IKSVD (ASimCO-IKSVD) respectively.

VI. SIMULATION RESULTS

In this section we present two categories of experiments to demonstrate the performance of our proposed algorithms. The first category contains experiments with synthetic data, and the second one provides image denoising results using the dictionaries learned with different ADL algorithms.

A. Experiments with Synthetic Data

Now we test the ADL algorithms with synthetic data. First of all, the approach to generating the synthetic data sets and the performance metrics employed are introduced. Second, we test our proposed algorithms with different initial dictionaries, showing their convergence and robustness to initializations. Third, the effect of the atom decorrelation step of the Incoherent Analysis SimCO algorithm is demonstrated. Fourth, experiments with different parameters are conducted to provide a more comprehensive comparison between our proposed algorithms and other ADL algorithms.

1) Synthetic data generation and performance metrics: A set of synthetic data consists of a reference analysis dictionary $\Omega \in \mathbb{R}^{p \times m}$ and a set of signals in $Y \in \mathbb{R}^{m \times n}$ that is sparse with respect to $\Omega$ with co-sparsity $l$. The reference dictionary $\Omega$ is generated as detailed in the settings of the experiments. The generation of the signals in $Y$ is based on the fact that the sparse analysis model can be used as a generative model with a given dictionary [6]. For generating each signal, $l$ rows of $\Omega$ are selected randomly and a basis for the null space of these $l$ rows is determined. Multiplying this basis by a random vector gives a vector which can be one member of the signal set, i.e. one column of $Y$. In the following experiments, $Y$ and its noisy version will both be used as the training samples. The noisy training signals are obtained by adding Gaussian noise with zero mean and standard deviation 0.04 to $Y$, set as in [17]. Learning with the original signals $Y$ is referred to as the noiseless case and learning with the noisy signals as the noisy case.
An advantage of using synthetic data is that the reference dictionary which can sparsify the signals exactly is available, and therefore the quality of a learned dictionary can be evaluated by comparing it with the reference dictionary. We use the recovery rate of the atoms to measure the performance of the algorithms for recovering the reference dictionary, following the experiments in [17]. An atom $\Omega_{j,:}$ of the reference dictionary is regarded as recovered if
\[ \min_i (1 - |\Omega_{j,:}^T \Omega_{j,:}|) < \tau, \] (20)
where $\hat{\Omega}_{j,:}$ are the atoms of the learned dictionary and $\tau$ is the threshold value to determine whether the atoms are recovered. The value of $\tau$ is typically set as 0.01 [17].

Another way for evaluating a learned dictionary is to consider the average co-sparsity of the original signals in $Y$ with respect to this dictionary since the final goal of ADL is to acquire a dictionary with which the analysis representations of the signals are sparse. We introduce an operator $\|x\|_0$ counting the number of the elements of $x \in \mathbb{R}^p$, which are below the threshold $\epsilon$, i.e.
\[ \|x\|_0 = \text{card}\{i : |x_i| < \epsilon, i = 1, 2, ..., p\}, \] (21)
where $x_i$ denotes the $i$th element of $x$ and $\epsilon > 0$. The threshold value $\epsilon$ should be close to zero and it is set as $\epsilon = 0.001$ throughout our experiments. The co-sparsity of a signal can be obtained by applying this operator to the product of the learned dictionary and this signal. The average co-sparsity of all signals are used as the second metric to evaluate the learned dictionaries.

2) Convergence of the proposed algorithms: Different initial dictionaries are used to demonstrate the convergence of our proposed algorithms. The reference dictionaries were generated with the random variables satisfying the i.i.d. Gaussian distribution with zero mean and unit variance and then the rows of the dictionaries were normalized. The size of the reference dictionaries was $50 \times 25$ (i.e. $p = 50$, $m = 25$). The number of training signals was 50000 (i.e. $n = 500000$) and their co-sparsity was 21 (i.e. $l = 21$) set as in [17]. Analysis SimCO and Incoherent Analysis SimCO were applied to learn analysis dictionaries respectively. The co-sparsity parameters of these two algorithms were both set as the reference co-sparsity. The coherence limit of Incoherent Analysis SimCO was set as $\mu_0 = 0.6$, based on our empirical tests.

Three types of matrices were used as initial dictionaries, following the experiments of [7]. The first type is the random matrix consisting of i.i.d. zero mean and unit variance Gaussian elements. The other two types are vertical concatenations of two matrices. One type is the vertical concatenations of two $25 \times 25$ 2D DCT matrices (defined as the Kronecker product of two $5 \times 5$ 1D DCT matrix), and the other is composed of two $25 \times 25$ identity matrices. We have used 100 independent runs to test the proposed algorithms, and the change of the objective function in (3) shows similar patterns in different runs. Fig. 1 shows the objective function value averaged from ten independent tests of the proposed algorithms over the iterations in noiseless case and noisy case. The objective function decreases monotonically for all the initializations in both the noiseless case and noisy case. Though the dictionaries were initialized in different ways, the algorithms converge to a similar final value. This indicates that our proposed algorithms can converge robustly with different initializations.

3) Effect of the atom decorrelation step: Now we compare the correlations of the atoms in the dictionaries learned by Analysis SimCO and Incoherent Analysis SimCO to show the effect of the atom decorrelation step. The initial dictionaries were set as random Gaussian matrices with normalized rows. Other settings were the same as those in the experiments of Fig. 1.

In order to observe the correlations of the atoms of a learned dictionary $\Omega$, we define its mutual correlation matrix $M$ as follows
\[ M(\Omega) = \text{abs}(I - \Omega \Omega^T), \] (22)
where operator $\text{abs}(\cdot)$ takes the element-wise absolute value of a matrix. The non-diagonal elements of $M(\Omega)$ represent the correlations between atoms of $\Omega$ and thus the coherence of $\Omega$ is the maximum value of all the elements of $M$, i.e. $\mu(\Omega) = \max(M(\Omega))$. The histograms of the mutual correlation matrices of the dictionaries learned by Analysis SimCO and Incoherent Analysis SimCO in one test are presented in Fig. 2, in both the noiseless case and noisy case. In the mutual correlation matrix obtained by Analysis SimCO, there are some elements close to 1, which means that highly-correlated atoms exist in the learned dictionary. These highly-correlated atoms disappear in the dictionary learned by Incoherent Analysis SimCO, as shown in the right plot of Fig. 2. This demonstrates that the atom decorrelation step can effectively avoid the highly-correlated atoms in the learned dictionary.

The recovery rate and average co-sparsity averaged from ten independent tests are shown in Fig. 3. It can be seen that the recovery rate is higher in both the noiseless case and noisy case, when the Incoherent Analysis SimCO algorithm is applied. The average co-sparsity in the noisy case also increases due to the atom decorrelation step. In the noiseless case, the average co-sparsity obtained by Incoherent Analysis SimCO is
are selected empirically to show the trends of the learning results of the algorithms in terms of recovery rate and average co-sparsity. The reference dictionaries were generated with random variables satisfying i.i.d. Gaussian distribution with zero mean and unit variance and their rows are normalized. The initial dictionaries used in all the algorithms were also generated in the same way.

<p>| TABLE I |</p>
<table>
<thead>
<tr>
<th>PARAMETERS USED IN THE COMPARISON OF DIFFERENT ADL ALGORITHMS WITH SYNTHETIC DATA.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>

Analysis SimCO and Incoherent Analysis SimCO were applied for 2500 iterations. For Incoherent Analysis SimCO, the coherence limit was $\mu_0 = 0.6$. The parameters of Analysis K-SVD were set as the experiments with synthetic data in [17]. We found that the LOST algorithm fails to recover any atom of the reference dictionary if the parameters as in the original paper [20] are used. This may be because the experiments with synthetic data scale differently from the experiments with image patches in [20]. Extensive experiments were conducted to find good parameters of LOST for the experiments with synthetic data. The coefficients of the penalty terms in the objective function were chosen as 50 and the index parameter in the correlation penalty term was 20. The step size and the iteration number of the inner gradient conjugate algorithm were $10^{-4}$ and 30 respectively. The number of iterations for LOST was fixed to 1000. For the AOL algorithms, its noiseless version (NL)AOL and noise-aware version (NA)AOL were applied to the noiseless case and the noisy case respectively. The iteration numbers of (NL)AOL and (NA)AOL were 50000 and 10 respectively, according to the settings in [18]. The coefficient of the objective function of (NA)AOL was $\lambda = 0.3$. Other parameters of these two algorithms were the same as suggested in [18]. The parameters of Transform K-SVD were set at their default values as in [21]. The parameters of GOAL were set as in the original code. The threshold used to replace similar rows in ASimCO-Random is also set as $\mu_0$ to be consistent with the coherence limit of Incoherent Analysis SimCO. The parameters for the decorrelation method in ASimCO-IKSVD are set as recommended in [29].

The recovery rate and average co-sparsity averaged from five independent tests with different $l$, $n$ and $p$ are presented in Figs. 4, 5 and 6 respectively. Abbreviations are used in the legends because of space limitation (IN-ASimCO, ASimCO, AKSVD and TKSVD are short for Incoherent Analysis SimCO, Analysis SimCO, Analysis K-SVD and Transform K-SVD respectively). In general, our proposed algorithms, 

4) Simulations with different parameters: Our proposed algorithms are compared with seven baseline algorithms: ASimCO-Random, ASimCO-IKSVD, Analysis K-SVD [17], LOST [20], AOL [18], Transform K-SVD [21] and GOAL [19]1. The algorithms were tested with different parameters, i.e. co-sparsity $l$, the number of training signals $n$ and the number of atoms $p$. In each test, one parameter was changed while the others were fixed, as shown in Table I. These parameters

1The code of GOAL was downloaded from http://www.gol.ei.tum.de/index.php?id=25&type=98.
Analysis K-SVD, LOST and Transform K-SVD show similar trends over the varying parameters. This may result from the same measurements used for co-sparsity, i.e. \( \ell_0 \)-norm, and their similar optimization procedure which alternates between the update of the analysis representation and the update of the dictionary. For these five algorithms, better dictionaries can be learned with larger co-sparsities (cf. Fig. 4) and more training samples (cf. Fig. 5). With the increase of the number of atoms, the recovery rates obtained by these algorithms decrease (cf. Fig. 6). The results of Incoherent Analysis SimCO are similar to the results of Analysis K-SVD and Transform K-SVD, which are better than the results of LOST. The recovery rates of the dictionaries obtained by Incoherent Analysis SimCO are higher than Analysis SimCO in all cases due to the restriction of the coherence of the learned dictionary. The average co-sparsities obtained by Analysis SimCO are closer to the reference co-sparsity than those obtained by Incoherent Analysis SimCO in the noiseless case, but the Incoherent Analysis SimCO algorithm shows advantage for the average co-sparsity in the noisy case. The results of (NL)AOL, (NA)AOL and GOAL appear to be quite different from the other methods compared. This might be due to the "\( \ell_1 \)-norm" or "\( \ell_p \)-norm" (\( 0 \leq p \leq 1 \)) used to estimate the co-sparsity of the coefficients, as opposed to the "\( \ell_0 \)-norm" used in the other algorithms. The relatively limited performances of (NL)AOL and (NA)AOL may result from the application of the UNTF constraint to the learned dictionaries, that the reference dictionaries do not satisfy. The results of Incoherent Analysis SimCO and ASimCO-Random are very similar to each other, and they both outperform the ASimCO-IKSVD algorithm.

The time (in seconds) of one test with different parameters is presented in Table II\(^3\). From Table II, we can see that our proposed algorithms are faster than Analysis K-SVD, LOST and (NL)AOL, but slower than Transform K-SVD and GOAL. It seems that Transform K-SVD is the best choice to learn dictionaries with synthetic data considering its good performance and efficient computation. However, for the application to image denoising, our proposed algorithm outperforms Transform K-SVD, which will be presented in the next two subsections. The running time of (NA)AOL changes substantially in different cases since sometimes the subgradient algorithm applied to update the dictionary requires longer time to converge. (NA)AOL seems to be faster than our proposed algorithms as shown in Table II. It should be noted that the running time presented here is the time for one test, however, the conditions for terminating the iterations in the algorithms compared are different. For each iteration, our proposed algorithms are faster than (NA)AOL which is consistent with the analysis of the computational complexities.
in Section IV-E.

**TABLE II**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of co-sparsities (l)</th>
<th>Number of training signals (n) (10^4)</th>
<th>Number of atoms (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASimCO</td>
<td>4</td>
<td>0.5</td>
<td>30</td>
</tr>
<tr>
<td>ASimCO</td>
<td>8</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>ASimCO</td>
<td>12</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>ASimCO</td>
<td>16</td>
<td>4</td>
<td>60</td>
</tr>
<tr>
<td>ASimCO</td>
<td>20</td>
<td>6</td>
<td>70</td>
</tr>
<tr>
<td>ASimCO</td>
<td>24</td>
<td>8</td>
<td>80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of co-sparsities (l)</th>
<th>Number of training signals (n) (10^4)</th>
<th>Number of atoms (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASimCO</td>
<td>1273</td>
<td>1263</td>
<td>1264</td>
</tr>
<tr>
<td>ASimCO</td>
<td>1323</td>
<td>1327</td>
<td>1327</td>
</tr>
<tr>
<td>ASimCO</td>
<td>1290</td>
<td>1291</td>
<td>1291</td>
</tr>
<tr>
<td>ASimCO</td>
<td>1329</td>
<td>1306</td>
<td>1306</td>
</tr>
<tr>
<td>ASimCO</td>
<td>1488</td>
<td>1488</td>
<td>1488</td>
</tr>
</tbody>
</table>

\(\sqrt{m} \times \sqrt{m}\) patches extracted from the noisy image are reshaped as column vectors which are concatenated as a matrix \(Z \in \mathbb{R}^{m \times n}\), where \(n\) is the number of patches. The recovering operation is directly applied to \(Z\) using the learned dictionary \(\Omega\), resulting in a noiseless estimation \(\hat{Y} \in \mathbb{R}^{m \times n}\). Overlapping patches are used to mitigate the blockiness artifacts caused by this patch-based framework. The denoised image can be obtained by reshaping the columns of \(\hat{Y}\) as image patches and averaging these overlapping patches.

The key idea of estimating \(\hat{Y}\) is to solve an optimization problem where the learned analysis dictionary \(\Omega\) serves in the regularization term reflecting the co-sparsity prior of \(\hat{Y}\), that is

\[
\arg \min_{Y} ||\Omega Y||_1 + \frac{\lambda}{2} ||Z - Y||_F^2,
\]

where \(\lambda\) is the Lagrangian multiplier to balance the data fidelity term \(||Z - \hat{Y}||_F^2\) and the regularization term \(||\Omega Y||_1\).

The alternating direction method of multipliers (ADMM) [30], [18] is applied to tackle this problem. It should be noted that the methods used for image recovery in the experiments of LOST [20], Analysis K-SVD [18], Transform K-SVD [21] and GOAL [19] are different, which makes it difficult to evaluate the dictionaries learned by different algorithms consistently. To make a fair comparison, the same image recovering method, formulated as (23), is used in our experiments. This method is selected because of its high computational efficiency.

2) Denoising performance evaluation and parameter selection: The images to be denoised were artificially corrupted by additive white Gaussian noise with the standard deviation being either \(\sigma = 12.8\) or \(\sigma = 45\), choosing empirically to represent the case of a relatively low or high level of noise respectively. Peak signal to noise ratio (PSNR) was used to measure the denoising performance. For an \(N\)-pixel noise-free image \(y \in \mathbb{R}^N\), the PSNR in decibels (dB) of its denoised version \(\hat{y} \in \mathbb{R}^N\) is defined as

\[
\text{PSNR} = 10 \log_{10} \frac{255^2 N}{||y - \hat{y}||_2^2},
\]

where \(||y - \hat{y}||_2^2\) is the mean squared error between the original image and its denoised version.

Throughout our experiments, we followed the same set up as in [18], [21] and fixed the size of the image patches to \(8 \times 8\), i.e. \(m = 64\). The overlap of the patches was set as 7. The size of the learned dictionaries was \(128 \times 64\), i.e. \(p = 128\). In the image recovering process, the proper selection of the Lagrangian multiplier \(\lambda\) is related to the noise level. In general, \(\lambda\) needs to be smaller when the noise level is higher. The method to choose optimal \(\lambda\) is out of the scope of our work. Herein a set of different \(\lambda\)'s was tested, and only the results of \(\lambda \in \{0.002, 0.01, 0.05, 0.1, 0.3, 0.5\}\) are presented to show the trends of the denoising results.

We still compare our proposed algorithms with the baseline algorithms as employed in the experiments for the synthetic data. The parameters about co-sparsity were set as follows. For Analysis SimCO and Incoherent Analysis SimCO, the co-sparsities were set as either \(l = 40\) or \(l = 80\). The corresponding parameters of the baseline algorithms were set.
based on the value of $l$, in order to ensure the equal co-sparisty. For Analysis K-SVD, only the $l = 40$ case was tested since $l$ cannot be greater than the signal dimension $m$ in its parameter settings [17]. There is no parameter related to the co-sparsity $l$ in (NA)AOL and GOAL. Other parameters were set as those employed in their original papers except for GOAL whose parameters are the same as in the experiments for the synthetic data. The coherence limit of Incoherent Analysis SimCO and the correlation threshold of ASimCO-IKDVD were both set as 0.2, which was lower than the value used in the experiments with synthetic data, since we found that, in general, the image dictionaries learned have a relatively lower coherence, as compared with that in the synthetic case. The same initial dictionaries, generated with i.i.d. Gaussian distribution with zero mean and unit variance, were used for different algorithms.

3) Face image denoising: Now we denoise face images using the learned analysis dictionaries, following the experiments in [18]. The face images are centred and cropped [31] and can be modelled as piecewise smooth signals approximately. The original face and the noisy face images are shown in Fig. 7. Two types of training data were tested: Type I consists of the patches extracted from 13 other clean face images [18]; Type II includes the patches extracted from the face image to be denoised. 16384 patches were randomly selected as training data in both cases [18].

The PSNR (in dB) values of the denoised face images averaged from five independent tests with varying $\lambda$ are presented in Fig. 8 ($\sigma = 12.8$) and Fig. 9 ($\sigma = 45$). In each of these two figures, the top and bottom sub-figures show the results using the Type I and Type II training data, respectively. The left and right sub-figures present the denoising results with $l = 40$ and $l = 80$, obtained by our proposed algorithms, LOST and Transform K-SVD. The results of Analysis K-SVD are only shown in the left sub-figures. The results of (NA)AOL and GOAL, which are not related to the co-sparsity setting, are plotted without modifications in both the left and right sub-figures. The best denoising results obtained via different algorithms with varying $\lambda$, i.e. the peak PSNR values of the lines in Fig. 8 and Fig. 9, are summarized in Table III.

Fig. 8 and Fig. 9 reveal that the denoising results change considerably with various $\lambda$. The best $\lambda$ of the set tested in the case $\sigma = 45$ is smaller than that of the case $\sigma = 12.8$ due to the increase of the noise level. Some common features of the ADL algorithms can also be observed from Table III. The dictionaries learned with higher co-sparsity ($l = 80$) perform better in general. In terms of the types of training data, the image patches from the noisy face image itself (Type II training data) seem to be more suitable to be the training data. As shown in Table III, the performance of GOAL is competitive, compared with other baseline algorithms. For the lower noise level with Type I training data, it obtains the best result. However, our proposed algorithms can obtain better results in other cases. Incoherent Analysis SimCO is able to get higher PSNR than Analysis SimCO in some cases. The Incoherent Analysis SimCO is outperforms ASimCO-Random and ASimCO-IKSV in all cases, except for the cases when Type I training data is applied with co-sparsity being 40.

4) Natural image denoising: Now we examine the denoising of the natural images shown in Fig. 10. The size of the images is $256 \times 256$. Similar to the denoising of face images,
two types of training data are tested: Type I contains image patches extracted from 5 other clean images shown in Fig. 11; Type II includes the patches of the image to be denoised. The number of training patches is 20000 [17]. The results (PSNR in dB) averaged from the denoised versions of the four test images are plotted in Fig. 12 (σ = 12.8) and Fig. 13 (σ = 45), and the peak results of each curve are listed in Table IV.

![Test images for natural image denoising.](image)

![Training images used for learning analysis dictionaries.](image)

According to Fig. 12 and Fig. 13, the best λ is bigger for the lower noise level, which is consistent with the objective function (23). Table IV indicates that the patches extracted from other clean natural images are preferred to the patches of the image to be denoised. In the lower noise level case, for Type I training data (NA)AOL and GOAL gives slightly better results than our proposed algorithms and for Type II training data ASimCO-Random performs better. Our proposed algorithms outperform the baseline algorithms in other cases.

The results obtained by Analysis SimCO and Incoherent Analysis SimCO are similar to each other.

![Natural image denoising (σ = 12.8).](image)

![Natural image denoising (σ = 45).](image)
<table>
<thead>
<tr>
<th>Training data type</th>
<th>Type I</th>
<th>Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>co-sparsity ( l )</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>ASimCO</td>
<td>31.45</td>
<td>31.17</td>
</tr>
<tr>
<td>IN-ASimCO</td>
<td>31.40</td>
<td>30.91</td>
</tr>
<tr>
<td>ASimCO-Random</td>
<td>31.29</td>
<td>31.25</td>
</tr>
<tr>
<td>ASimCO-IKSVD</td>
<td>30.22</td>
<td>32.26</td>
</tr>
<tr>
<td>AKSVD</td>
<td>28.12</td>
<td>—</td>
</tr>
<tr>
<td>LOST</td>
<td>29.47</td>
<td>31.05</td>
</tr>
<tr>
<td>TKSVD</td>
<td>29.59</td>
<td>30.17</td>
</tr>
<tr>
<td>(NA)AOL</td>
<td>31.47</td>
<td>29.57</td>
</tr>
<tr>
<td>GOAL</td>
<td>31.53</td>
<td>29.56</td>
</tr>
</tbody>
</table>

\( \sigma = 12.8 \)

<table>
<thead>
<tr>
<th>Training data type</th>
<th>Type I</th>
<th>Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>co-sparsity ( l )</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>ASimCO</td>
<td>25.73</td>
<td>24.24</td>
</tr>
<tr>
<td>IN-ASimCO</td>
<td>25.74</td>
<td>25.37</td>
</tr>
<tr>
<td>ASimCO-Random</td>
<td>25.54</td>
<td>25.71</td>
</tr>
<tr>
<td>ASimCO-IKSVD</td>
<td>22.22</td>
<td>22.53</td>
</tr>
<tr>
<td>AKSVD</td>
<td>22.17</td>
<td>—</td>
</tr>
<tr>
<td>LOST</td>
<td>22.17</td>
<td>22.39</td>
</tr>
<tr>
<td>TKSVD</td>
<td>22.17</td>
<td>22.19</td>
</tr>
<tr>
<td>(NA)AOL</td>
<td>23.54</td>
<td>22.18</td>
</tr>
<tr>
<td>GOAL</td>
<td>23.85</td>
<td>22.19</td>
</tr>
</tbody>
</table>

\( \sigma = 45 \)

have confirmed the competitive performance of our proposed algorithms. The applications of learned analysis dictionaries in other signal processing tasks merit more study, which we leave for future work.

**ACKNOWLEDGEMENT**

The authors thank the Associate Editor and the anonymous reviewers for their contributions to improving the quality of the paper.

**REFERENCES**


**Jing Dong** received the B.Eng. degree in automation in 2010 and the M.Eng. degree in navigation, guidance and control in 2013, both from Harbin Engineering University, Harbin, China. She is now pursuing her Ph.D. degree in the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, U.K. Her current research interests include sparse representation and image processing.

**Wei Dai** is currently a Lecturer in Electrical and Electronic Engineering at Imperial College London. He received his Ph.D. degree in Electrical and Computer Engineering from the University of Colorado at Boulder in 2007. From 2007 to 2010, he was a Postdoctoral Research Associate in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign.

**Wenwu Wang** (M02-SM11) was born in Anhui, China. He received the B.Sc. degree in automatic control in 1997, the M.E. degree in control science and control engineering in 2000, and the Ph.D. degree in navigation guidance and control in 2002, all from Harbin Engineering University, Harbin, China. He then joined Kings College, London, U.K., in May 2002, as a postdoctoral research associate and transferred to Cardiff University, Cardiff, U.K., in January 2004, where he worked in the area of blind signal processing. In May 2005, he joined the Tao Group Ltd. (now Antix Labs Ltd.), Reading, U.K., as a DSP engineer working on algorithm design and implementation for real-time and embedded audio and visual systems. In September 2006, he joined Creative Labs, Ltd., Egham, U.K., as an R&D engineer, working on 3D spatial audio for mobile devices. Since May 2007, he has been with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K., where he is currently a Reader in Signal Processing, and a Co-Director of the Machine Audition Lab. He is a member of the Ministry of Defence (MoD) University Defence Research Collaboration (UDRC) in Signal Processing (since 2009), a member of the BBC Audio Research Partnership (since 2011), and an associate member of Surrey Centre for Cyber Security (since 2014). During spring 2008, he has been a visiting scholar at the Perception and Neurodynamics Lab and the Center for Cognitive Science, The Ohio State University. His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection. He is currently an Associate Editor for IEEE Transactions on Signal Processing.

**Mark D. Plumbley** received the B.A. (honours) degree in electrical sciences and the Ph.D. degree in neural networks from the University of Cambridge, United Kingdom, in 1984 and 1991, respectively. From 1991 to 2001, he was a lecturer at Kings College London, before moving to Queen Mary University of London in 2002, later becoming Director of the Centre for Digital Music. In 2015 he joined the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey, as Professor of Signal Processing. His research focuses on the automatic analysis of audio and music, including automatic music transcription, audio event detection and acoustic scene analysis, using methods such as source separation and sparse representations. He is a member of the IEEE Signal Processing Society Technical Committees on Audio & Acoustic Signal Processing and Signal Processing Theory and Methods. He is a Fellow of the IEEE.

**Zifa Han** received the B.Sc. degree in information and computing science from Northeast Petroleum University, China in 2010 and the M.Eng. degree in navigation, guidance and control from Harbin Engineering University, China in 2013. He is currently a Ph.D. student in the Department of Electronic Engineering, City University of Hong Kong. His research interests include machine learning and image processing.

**Jonathon Chambers** (S83-M90-SM98-F11) received the Ph.D. and D.Sc. degrees in signal processing from the Imperial College of Science, Technology and Medicine (Imperial College London), London, U.K., in 1990 and 1994, respectively. From 1991 to 1994, he was a Research Scientist with the Schlumberger Cambridge Research Center, Cambridge, U.K. In 1994, he returned to Imperial College London as a Lecturer in signal processing and was promoted to Reader (Associate Professor) in 1998. From 2001 to 2004, he was the Director of the Center for Digital Signal Processing and a Professor of signal processing with the Division of Engineering, Kings College London. From 2004 to 2007, he was a Cardiff Professorial Research Fellow with the School of Engineering, Cardiff University, Cardiff, U.K. Between 2007-2014, he led the Advanced Signal Processing Group, within the School of Electronic, Electrical and Systems Engineering and is now a Visiting Professor. In 2015, he joined the School of Electrical and Electronic Engineering, Newcastle University, where he is a Professor of signal and information processing and heads the ComS2IP group. He is also a Guest Professor at Harbin Engineering University, China. He is a coauthor of the books Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability (New York, NY, USA:Wiley, 2001) and EEG Signal Processing (New York, NY, USA: Wiley, 2007). He has advised more than 60 researchers through to Ph.D. graduation and published more than 400 conference proceedings and journal articles, many of which are in IEEE journals. His research interests include adaptive and blind signal processing and their applications.

Dr. Chambers is a Fellow of the Royal Academy of Engineering, U.K., and the Institution of Electrical Engineers. He was the Technical Program Chair of the 15th International Conference on Digital Signal Processing and the 2009 IEEE Workshop on Statistical Signal Processing, both held in Cardiff, U.K., and a Technical Program Cochair for the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic. He received the first QinetiQ Visiting Fellowship in 2007 for his outstanding contributions to adaptive signal processing and his contributions to QinetiQ, as a result of his successful industrial collaboration with the international defense systems company QinetiQ. He has served on the IEEE Signal Processing Theory and Methods Technical Committee for six years and the IEEE Signal Processing Society Awards Board for three years. He is currently a member of the IEEE Signal Processing Conference Board and the European Signal Processing Society Best Paper Awards Selection Panel. He has also served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING for three terms over the periods 1997-1999, 2004-2007, and as a Senior Area Editor 2011-15.