Learning Dictionary From Signals under Global Sparsity Constraint

Deyu Meng[a], Qian Zhao[a], Yee Leung[b], Zongben Xu[a]

[a] *Institute for Information and System Sciences and Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, P.R. China*

[b] *Department of Geography & Resource Management, The Chinese University of Hong Kong, Hong Kong, P.R. China.*

Abstract

A new method is proposed in this paper to learn overcomplete dictionary from signals. Differing from the current methods that enforce uniform sparsity constraint on the coefficients of each input signal, the proposed method attempts to impose global sparsity constraint on the coefficient matrix of the entire signal set. This enables the proposed method to fittingly assign the atoms of the dictionary to represent various signals and optimally adapt to the complicated structures underlying the entire signal set. By virtue of the sparse coding and sparse PCA techniques, a simple algorithm is designed for the implementation of the method. The efficiency and the convergence of the proposed algorithm are also theoretically analyzed. Based on the experimental results implemented on a series of signal and image data sets, the capability of the proposed method is substantiated in original dictionary recovering, signal reconstructing and salient signal structure revealing.

*Corresponding author. Tel: 86 130 3290 4180; Fax: 86 29 8266 8559.
Email address: dymeng@mail.xjtu.edu.cn. (Deyu Meng)

Keywords: Dictionary learning, signal reconstruction, sparse principle component analysis, sparse representation, structure learning.

1. Introduction

In recent years, there has been a significant interest in using sparse representation over a redundant dictionary as a driving force for various signal processing tasks. All of these applications capitalize on the fact that salient features in signals can always be captured by their sparse representations over an appropriate dictionary. As such, the pre-specified dictionary is crucial to the success of the sparse representation model in practical applications. Most conventional studies use the "off-the-shelf" dictionaries, such as the wavelet [1] and DCT bases [2], to build a sparsifying dictionary based on a mathematical model of the data. Current studies, however, have demonstrated the advantages of learning an often overcomplete dictionary matched to signals of interest [3]-[6].

The dictionary learning task is mathematically described as follows: For a collection of signals $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in R^{d\times n}$, it is expected to find the dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m] \in R^{d\times m}$, composed by a collection of atoms $\mathbf{d}_i$ (the atom number m is set larger than d, implying that the dictionary is redundant), through the following optimization model [7][8]:

$$\min_{\mathbf{D},\mathbf{A}} \ \frac{1}{n}\sum\nolimits_{i=1}^{n} \left(\frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\mathcal{P}(\mathbf{a}_i)\right), \qquad (P_\lambda)$$

where the vector $\mathbf{a}_i$ contains the representation coefficients of $\mathbf{x}_i$. We denote the coefficient matrix as $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n] \in R^{m\times n}$. The objective function of (P_λ) involves two elements in the dictionary learning task: the expression

error term, $\frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2$, and the sparsity controlling term, $\mathcal{P}(\mathbf{a}_i)$, with respect to each input signal $\mathbf{x}_i$. The most widely utilized functions of $\mathcal{P}(\mathbf{a}_i)$ include the l_0 penalty $\|\mathbf{a}_i\|_0$ and the l_1 penalty $\|\mathbf{a}_i\|_1$. The other two generally utilized models for the dictionary task are [3][4]:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 \quad \text{s.t.} \quad \mathcal{P}(\mathbf{a}_i) \leq k, \ \ \forall \ \ 1 \leq i \leq n, \tag{P_k}$$

and

$$\min_{\mathbf{D},\mathbf{A}} \sum\nolimits_{i=1}^{n} \mathcal{P}(\mathbf{a}_i) \quad \text{s.t.} \quad \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2 \leq \epsilon, \ \ \forall \ \ 1 \leq i \leq n, \tag{P_ϵ}$$

where the notion $\|\cdot\|_F$ stands for the Frobenius norm. The tunable parameters λ, k and ε in the models (P_λ), (P_k) and (P_ϵ) play an important role in the model performance. They intrinsically control the compromise between the expression error and the sparsity of the representation coefficients.

It should be noted that a uniform parameter λ, k or ε formulated for the entire signal set is specified in the current model (P_λ), (P_k) or (P_ϵ), respectively. Such formulation facilitates the parameter selection and algorithm construction of the model. The signals in applications, however, are always of varying interior structures. On one hand, some signals may be composed of complicated features and need to be very densely represented under the dictionary; while some might be of very simple structure and can be precisely represented with very sparse coefficient vectors. On the other hand, some signals may seriously deviate from the original due to the analog-to-digital conversion errors or transmission through noisy channels, while some may be totally clean samples. This means that it is better to vary the parameter λ, k or ε with respect to different signals to make the dictionary learning model adaptive to the intrinsic structures underlying the entire signal set. The

uniform specification of the sparsity penalty λ, the maximal representation sparsity k, or the minimal representational error ε of the conventional model (P_λ), (P_k) or (P_ϵ), respectively, thus possibly conducts unstable performance of the model in applications.

The purpose of this paper is to formulate a new dictionary learning model, which is with simple form while will not impose uniform penalty or constraint on each signal like the conventional methods. Instead, the model will specify a global sparsity constraint on the coefficients of the entire signal set, so that it will adaptively tune the representation sparsity of diverse signals and properly reveal the intrinsic structures underlying the entire signal set. An efficient algorithm is specifically designed for the proposed model. It is efficient, convergent and easy to be implemented. By a series of experiments, it is verified that the proposed algorithm, in comparison with the current dictionary learning methods, can not only deliver more faithful dictionary underlying the input signals, but also can more precisely recover the original signals. Besides, the intrinsic structure underneath the entire signal set can be impressively depicted via the different representation sparsities of the signals under the learned dictionary.

In what follows, related work in the literature is first reviewed in Section II. Details of our algorithm and its basic model are then presented in Section III. The experimental results are given in Section IV for substantiation and verification. The paper is then concluded with a summary and outlook for future research. Throughout the paper, we denote matrices, vectors and scalars by the upper-case bold-faced letters, lower case bold-faced letters and lower-case letters, respectively.

2. Related work

Using sparse representations of signals under an appropriately specified redundant dictionary has advanced multiple signal processing tasks, and has drawn much attention recently. In the conventional studies, the "off-the-shelf" dictionaries have always been employed in various applications. The typical ones include the Fourier [9], the wavelet [10]-[11] and the DCT bases [2]. These bases have been successfully applied to many practical problems.

Recent research, however, has demonstrated the significance of learning an overcomplete dictionary, instead of a fixed one, matched to the signals of interest. Various algorithms along this line have been formulated in recent years. For example, the algorithm proposed by Olshausen and Field [12] can find sparse linear codes for natural scenes. The dictionary composed by these codes complies with the intrinsic features of the localized, oriented, bandpass receptive fields of the neurons of the primary visual cortex. The method of optimal directions (MOD), proposed by Engan et al. [13], is also an appealing dictionary training algorithm. It improves the efficiency of the work by Olshausen and Field [12] both in the sparse coding and dictionary updating stages. Through generalizing the K-means clustering process to alternate between sparse coding and dictionary updating, Aharon et al. [3, 4] designed the K-SVD method. There are two versions of the method: one achieves sparse signal representations under strict sparsity constraint (corresponding to the model (P_k), and is thus denoted as K-SVD$_{P_k}$) [3], and the other calculates the dictionary by allowing a bounded representation error for each signal (corresponding to the model (P_ε), and is thus denoted as K-SVD$_{P_\varepsilon}$) [4]. The method has empirically shown the state-of-the-art

performance in some image processing applications. Some methods have further been constructed to improve the efficiency of the dictionary learning problem: Mairal et al. [14] proposed an online model to efficiently solve the dictionary learning problem; Jenatton et al. [15] used a tree structured sparse representation to give a linear-time computation of the problem; and Lee et al. [7] designed an algorithm to speedup the sparse coding stage of the problem, allowing it to learn larger sparse codes than other algorithms. Recently, some algorithms have also been developed to extend the capability of dictionary learning based on some specific motivations. For instance, Shi et al. [8] developed an algorithm for dictionary learning with non-convex while continuous minimax concave penalty; Mairal et al. [16] established a discriminative approach, instead of the purely reconstructive methods, to build a dictionary. All of the aforementioned methods are addressed to the models (P_λ), (P_k) and (P_ϵ) introduced in Section I.

For many real signal processing applications, however, these models cannot fully tally with the practical signals possessing intrinsic complicated structures, such as those with different content of features, or with spatial and/or spectral non-uniform noises. Along this line of research, Mairal et al. [17] addressed the case of removing nonhomogeneous white Gaussian noise of images, while apriori knowledge of the noise deviation at each pixel of the objective images has to be pre-assumed. Such elaborate information about noise, however, generally cannot be attained in practice. Spielman et al. [18] proposed a method called ERSuD, which is also based on the global l_0 sparsity of a matrix. However, this method is applicable only when the dictionary matrix is square and invertible, which generally does not hold in real

dictionary learning applications. Very recently, Zhou et al. [5, 6] designed a nonparametric Bayesian method, called the beta process factor analysis (BPFA) method, for dictionary learning. The method can learn a sparse dictionary in situ for signals with spatially non-uniform noises, without having to know the apriori noise information. The method is thus used as one of the methods for comparison in our experiments.

In this paper, we propose a new dictionary learning method to enhance the capability of dictionary learning by imposing a global sparsity constraint on the coefficients of all training signals to enable adaptive atom assignment to individual signals based on their intrinsic structures. It should be noted that the concept of global-sparsity-constraint has also been involved in some other problems, such as image decomposition [19] and magnetic resonance imaging [20], where this idea was verified to be beneficial to achieve a global consistency of data structures. In machine learning area, there are also some literatures modeling the matrix with l_1-norm global sparsity [21]-[23]. Here we first introduce the l_0-norm global sparsity in dictionary learning problem.

3. Dictionary learning under global sparsity constraint: model and algorithm

3.1. Model: From local to global constraint on sparse representation

The current dictionary learning models, namely (P_λ), (P_k) and (P_ϵ), enforce uniform sparsity control parameter, including sparsity penalty λ, sparsity constraint k or representation error bound ϵ, for each involved signal. However, there are often counterexamples to such formulation in real applications, especially for signals embedded with intrinsic heterogeneous sparsity

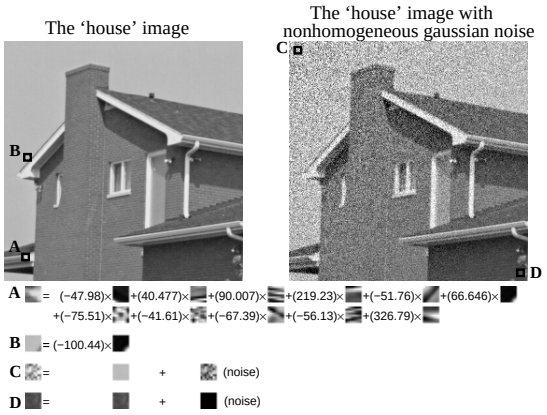

Figure 1: The left figure is the “house” image, and the right figure is the same image mixed with nonhomogeneous Gaussian noise. A,B,C,D are four patches cut from the two images. The upper two expressions at the bottom of the figure show the atoms and the coefficients utilized to sparsely represent the patches A and B over the dictionary learned from the algorithm proposed in Section 3.2, respectively. The lower two ones demonstrate the groundtruth of the noise separated from the patches C and D, respectively.

structures. We take the image case as an instance, in which the input signal set corresponds to the small local patches of the image in consideration. On one hand, the local parts of a real image may contain very different capacities of meaningful features, e.g., the region full of patterns with abundant textures, as compared to the area located at the background with small grayscale variations. In such case, smaller sparsity penalty λ of the optimization model (P_λ) or larger representation sparsity k of the optimization model (P_k), should be preset for the local patches located at the former region so that more atoms of the dictionary can be assigned to them. This phenomenon can be easily understood via the representations of the patches A and B located at the house eave and the background parts of the “house” image, respectively, as shown in Figure 1. On the other hand, the real noise mixed in the image is often of significant statistical heteroscedasticity. This means that the extents of noise corruption in various parts of the image,

such as the patches C and D in Figure 1, may be significantly different. It is easy to see that the patch C is highly corrupted by noise while D is almost clean and contains essentially no noise. Larger representation error bound ϵ of the optimization model (P_ϵ) should then be set for the former patches so that the model can properly capture the variation of noise corruption across the image. It is thus foreseeable that the performance of the current dictionary learning models can be substantially enhanced by adaptively regulating the sparsity control parameter(s) with respect to the underlying structural characteristics of the entire signals.

Based on the above rationale, we reformulate the model for dictionary learning into the following global-sparsity-constraint form:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 \quad \text{s.t.} \quad \|\mathbf{A}\|_0 \le K, \tag{P_K}$$

where K is the maximal size of the non-zero entries of the coefficient matrix $\mathbf{A}$ (i.e., the combination of the l_0 sparsities of all signals), and $\|\mathbf{A}\|_0$ counts the nonzero entries of the matrix $\mathbf{A}$. Differing from the current models in which uniform sparsity constraint is imposed on the coefficient vector for each input signal, our proposed model capitalizes on the global sparsity constraint superimposed upon the coefficient matrix for the entire signal set. This formulation enables the model to adaptively assign different number of atoms, k_i, of the dictionary to represent each signal $\mathbf{x}_i$ according to its intrinsic structure. This can be easily understood through the following equivalent reformulation of (P_K):

$$\min_{\mathbf{D},\mathbf{A},\{k_i\}_{i=1}^n} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 \quad \text{s.t.} \quad \|\mathbf{a}_i\|_0 \le k_i,\ 1\le i\le n,\ \ \sum\nolimits_{i=1}^n k_i = K. \tag{1}$$

In specific, for signals containing different capacities of features, more non-

zero atoms will be assigned to represent more complex signals by the proposed model; and for signals corrupted by the heterogeneous noises, the distribution of the non-zero entries of the coefficient matrix tends to be optimally balanced among the entire signal set, and the sparse representations of the signals over the dictionary attained by the proposed model will possibly reveal the major information (the original signals) while eliminate the minor (the mixed noises) at the global scale. The dictionary learning model (P_K) is thus expected to outperform the current models.

We now construct an efficient algorithm for solving (P_K).

3.2. Algorithm: Iteratively updating columns and rows of coefficient matrix

The main idea of our algorithm is to iteratively update the column and row vectors of the coefficient matrix $\mathbf{A}$ to approach the solution to (P_K). Denote the column and row vectors of the coefficient matrix $\mathbf{A}$ ($\in R^{m\times n}$) as $[\mathbf{a}_1^c, \mathbf{a}_2^c, \cdots, \mathbf{a}_n^c]$ and $[\mathbf{a}_1^r, \mathbf{a}_2^r, \cdots, \mathbf{a}_m^r]$, respectively. The column updating step is to update each column vector $\mathbf{a}_i^c$ ($i = 1, 2, \cdots, n$) of $\mathbf{A}$, with the number of its non-zero entries, k_i^c, fixed, while optimally relocate the column positions of these non-zero elements in an adaptive way (i.e., k_i^rs will be varied after this step), as graphically depicted in the upper of Figure 2. By decoupling the model (P_K), the corresponding task is to solve the following optimization model for each $\mathbf{a}_i^c$ ($i = 1, \cdots, n$):

$$\min_{\mathbf{a}_i^c} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i^c\|_2^2 \quad \text{s.t.} \quad \|\mathbf{a}_i^c\|_0 \leq k_i^c. \tag{2}$$

The row updating step is to update each row vector $\mathbf{a}_i^r$ ($i = 1, 2, \cdots, m$) of $\mathbf{A}$, with the sparsity k_i^r of $\mathbf{a}_i^r$ fixed, while optimally adapt its k_i^r non-zero elements to the proper row positions (i.e., k_i^cs will be changed after this step),

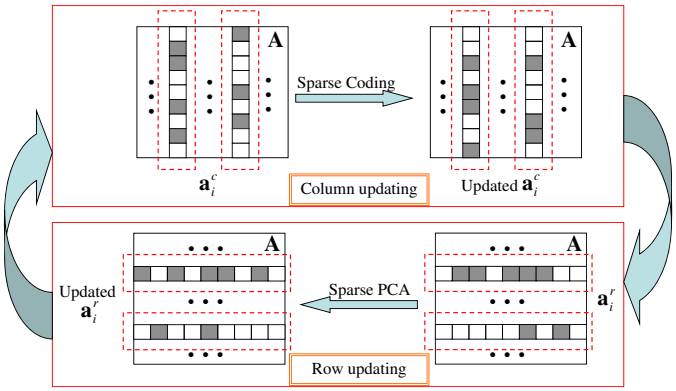

Figure 2: Graphical presentation of the iteration process between the column updating step and the row updating step in the proposed algorithm. The upper panel shows that in the column updating step, each column vector $\mathbf{a}_i^c$ of $\mathbf{A}$ is updated, with the number of its non-zero entries fixed while the column positions of these non-zero elements are optimally relocated in an adaptive way. The lower panel demonstrates that in the row updating step, each row vector $\mathbf{a}_i^r$ of $\mathbf{A}$ is updated, with its sparsity fixed while its non-zero elements are optimally adapted to the proper row positions.

as shown in the lower of Figure 2. Since the atom $\mathbf{d}_i$ of the dictionary $\mathbf{D}$ one-to-one corresponds to $\mathbf{a}_i^r$ in the sense that $\mathbf{DA} = \sum_{i=1}^m \mathbf{d}_i \left(\mathbf{a}_i^r\right)^T$, it is also simultaneously updated in this step together with $\mathbf{a}_i^r$. The corresponding optimization model is of the following form for each $\mathbf{a}_i^r$ $(i = 1, \cdots, m)$:

$$\min_{\mathbf{a}_i^r, \mathbf{d}_i} \left\| \mathbf{E}_i - \mathbf{d}_i (\mathbf{a}_i^r)^T \right\|_F^2 \quad \text{s.t.} \quad \|\mathbf{a}_i^r\|_0 \leq k_i^r, \quad \mathbf{d}_i^T \mathbf{d}_i = 1, \tag{3}$$

where $\mathbf{E}_i = \mathbf{X} - \sum_{j \neq i} \mathbf{d}_j \left(\mathbf{a}_j^r\right)^T$ stands for the representation error of all considered signals with the effect of the i-th atom $\mathbf{d}_i$ removed. It should be noted that the sparsity k_i^c of each $\mathbf{a}_i^c$ and k_i^r of each $\mathbf{a}_i^r$ are dynamically and adaptively adjusted during the iterations between the column updating and the row updating steps of the proposed algorithm.

The algorithm can then be summarized as Algorithm 1.

Now the question is how to efficiently solve the optimization models (2) and (3). For (2), it is actually the well known l_0-norm model of sparse coding

Algorithm 1 Algorithm for dictionary learning under global sparsity constraint (GDL)

Given: The input data $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in R^{d\times n}$, the global sparsity K

Execute:

1. Initialize the dictionary $\mathbf{D} \in R^{d\times m}$ and the coefficient matrix $\mathbf{A} \in R^{m\times n}$ with sparsity K, respectively.
2. Repeat
   2.1 (Column updating). Update the column vector $\mathbf{a}_i^c$ of $\mathbf{A}$ by solving (2) for each $i = 1, \cdots, n$.
   2.2 (Row updating). Update the row vector $\mathbf{a}_i^r$ of $\mathbf{A}$ and the atom $\mathbf{d}_i$ of $\mathbf{D}$ by solving (3) for each $i = 1, \cdots, m$.

   Until the termination condition is satisfied

Return: the solution $\mathbf{D}$, $\mathbf{A}$ of (P_K).

and multiple effective algorithms have been investigated to solve this model. The typical ones include the thresholding methods, e.g., the hard algorithm [24], and the greedy methods, e.g., the OMP algorithm [25].

For (3), we give the following theorem:

Theorem 1. *The optimum of the optimization model (3) can be attained by solving the optimization model:*

$$\max_{\boldsymbol{w}} \ \boldsymbol{w}^T \boldsymbol{E}_i^T \boldsymbol{E}_i \boldsymbol{w} \qquad s.t. \ \ \boldsymbol{w}^T \boldsymbol{w} = 1, \|\boldsymbol{w}\|_0 \le k_i^r, \tag{4}$$

in the sense of

$$\widehat{\boldsymbol{d}}_i = \frac{\boldsymbol{E}_i \widehat{\boldsymbol{w}}}{\|\boldsymbol{E}_i \widehat{\boldsymbol{w}}\|_2}, \widehat{\boldsymbol{a}}_i^r = \|\boldsymbol{E}_i \widehat{\boldsymbol{w}}\|_2 \, \widehat{\boldsymbol{w}}, \tag{5}$$

where $\widehat{\boldsymbol{d}}_i$, $\widehat{\boldsymbol{a}}_i^r$ are the optimum of (3), and $\widehat{\boldsymbol{w}}$ is the optimum of (4).

The proof of Theorem 1 is given in the appendix.

It is very interesting that (4) is just the sparse principal component analysis (sparse PCA) model [26], which has been thoroughly investigated in the last decade [26]-[30]. Many efficient algorithms have been constructed for solving the model, including SPCA [26], GPower [27], sPCA-rSVD [28] etc..

Thus, both models (2) and (3) can be efficiently solved, i.e., both of the column and row updating steps of the proposed algorithm can be effectively implemented, by employing the existing methods in sparse coding and sparse PCA, respectively.

The remaining issues are then how to appropriately specify the initial dictionary $\mathbf{D}$ and the coefficient matrix $\mathbf{A}$ in step 1, and when to terminate the iterative process in step 2 of the proposed algorithm. In our experiments, $\mathbf{D}$ and $\mathbf{A}$ were simply initialized with data signals and sparse matrix with K non-zero elements, whose positions are randomly generated in the matrix, respectively. By counting the nonzero element numbers of each column vector $\mathbf{a}_i^c$ and each row vector $\mathbf{a}_i^r$ of the initialized $\mathbf{A}$, the sparsity k_i^c and k_i^r are simultaneously specified. As for the termination of the algorithm, since the entire representation error of signals, i.e., the objective function of (P_K), decreases monotonically under the fixed global sparsity constraint throughout the iterative process, the algorithm can be reasonably terminated when the decrease in value of the objective function is smaller than some preset small threshold, or the process has reached the pre-specified number of iterations.

As for the convergence of the proposed algorithm, under the assumption that the models (3) and (4) can be precisely solved, each of the updating iterations in step 2 monotonically decreases the objective function of (2),

i.e., the total representation error $\|\mathbf{DA}-\mathbf{X}\|_F^2$, under the guarantee that the constraint $\|\mathbf{A}\|_0 \leq K$ is consistently held. Since this objective function is also lower bounded (≥ 0), the algorithm is guaranteed to be convergent. Although the above claim depends on the success of the sparse coding and sparse PCA techniques used to approximate the solutions to (3) and (4), respectively, the algorithms employed on the tasks always performed well in our experiments and can empirically generate a rational solution of (5) after multiple iterations, as demonstrated in the next section.

4. Experimental results

To test the effectiveness of the proposed algorithm on dictionary learning, it was applied to a series of synthetic signals and real images for substantiation. The results are summarized in the following discussion. All programs were implemented on the Matlab 7.0 platform.

4.1. Synthetic signal experiments with homogeneous noise

We first apply the proposed algorithm to synthetic signal data to test whether the algorithm can recover the generating dictionary and reconstruct the original signals. Two series of experiments were implemented, each involving 11 sets of signals. Each signal set contained 1500 20-dimensional signals, denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{1500}] \in R^{20\times 1500}$, which were created by a linear combination of a pre-specified dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_{50}] \in R^{20\times 50}$ and representation coefficients $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{1500}] \in R^{50\times 1500}$, and mixed with different extents of homogeneous Gaussian white noise. The entries of each dictionary $\mathbf{D}$ were first generated by random sampling, and each column (atom) $\mathbf{d}_i$ of $\mathbf{D}$ was then divided by its l_2-norm for normalization.

For the first series of experiments, each column $\mathbf{a}_i$ of $\mathbf{A}$ contained 3 non-zero elements with randomly chosen values and locations. For the second series of experiments, each coefficient matrix $\mathbf{A}$ consisted of 4500 randomly valued and located non-zero entries. Added to the 11 signal sets in both series of experiments were homogeneous Gaussian noises with standard deviations σ varying from 0 to 0.1 with interval 0.01. The corresponding SNR values of these signal sets ranged from infinity to around 10[1]. It should be indicated that for the first experimental series, the signals in each experiment were of similar representation sparsity over the preset dictionary, which complies with the preassumption of the model (P_k); and for both series of experiments, the signals were corrupted with homogeneous noises, which tallies with the preassumption underlying the model (P_ε).

Five of the current dictionary learning methods, including the MOD [13], K-SVD$_{P_k}$ [3], K-SVD$_{P_\varepsilon}$ [4], Efficient [7] and BPFA [5] methods were applied to these signal sets for comparison. The dictionaries of the first four methods, as well as the proposed GDL method, were initialized as the randomly selected signals from the input set, and the initialization of the BPFA method was based on the singular value decomposition technique [5]. Since both MOD and K-SVD$_{P_\varepsilon}$ need the apriori deviation of the noise mixed in the signals to preset the representation error parameter, we directly used the groundtruth information to optimally specify the parameter value [4]. For the K-SVD$_{P_k}$ and Efficient methods, the sparsity constraint parameters k and λ were specified by running the method 5 times on each signal set under

[1]The signal set mixed with Gaussian noise with deviation 0 means that the set is clean and contains no noise. The corresponding value of SNR is infinite.

different parameter values, and selecting the best one as the final output. All parameters involved in the BPFA were automatically inferred by using a full posterior on the model [5]. For the proposed GDL method, the global sparsity K was set as 4500 in all experiments. The results of K-SVD$_{P_k}$, K-SVD$_{P_\varepsilon}$, MOD, Efficient and GDL were attained after 100 iterations, and those of BPFA were obtained after 1000 iterations of Gibbs sampling.

Two criteria are utilized to assess the performance of the employed methods for dictionary learning. The first is computed by sweeping through each atom of the generating dictionary and checking whether it is recovered by the dictionary attained by a utilized method via the following formula [3]:

$$1 - |\mathbf{d}_i^T \widehat{\mathbf{d}}_i|, \tag{6}$$

where d_i is the atom of the original dictionary and $\widehat{\mathbf{d}}_i$ is its corresponding closest atom in the recovered dictionary. If the value of (6) was less than 0.01, then it was considered as a success. The rate of the successfully recovered atoms in the generating dictionary (called the dictionary recovery rate, or DR in brief) is then taken as the first criterion, which evaluates the capability of the method in delivering the original dictionary beneath the input signals. The second criterion is the mean of the standard deviations of the reconstructed signals from the original signals (called the representation error, or RE briefly). This value assesses the performance of the method in recovering the input signals.

In all of the implemented experiments, the DR and RE values in the iterative processes of the five current methods and the proposed GDL method were recorded. The upper and lower panels of Figure 3 depict the RE and DR curves of the six methods in the iterative processes of three of the first

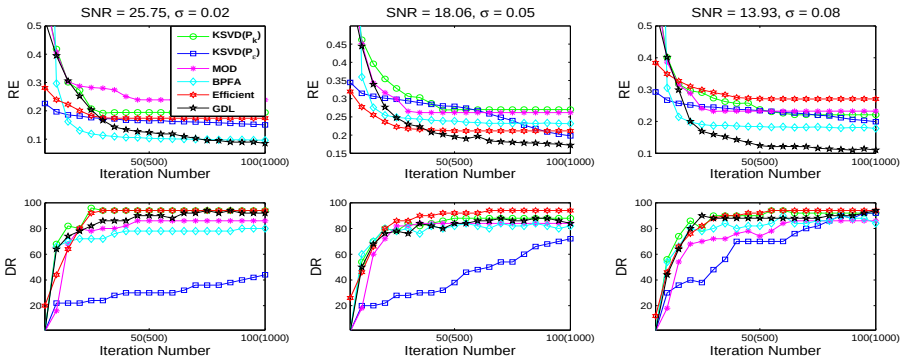

Figure 3: The upper panels: The RE curves of the K-SVD$_{P_k}$, K-SVD$_{P_\varepsilon}$, MOD, BPFA, Efficient and the proposed GDL methods in the iterative processes of three of the first series of experiments. The lower panels: the corresponding DR curves. It should be noted that the BPFA method implements 1000 iterations of Gibbs sampling, while other methods run 100 iterations.

series of experiments, respectively. Figure 4 shows the corresponding results in three cases of the second series of experiments. Panels (a) and (c) of Figure 5 show the final RE and DR values of the six methods in 11 experiments of the first series, respectively. For easy comparison, panels (b) and (d) of the figure display the mean values of RE and DR of the six methods as vertical bars. Figure 6 depicts the cases of the second series of experiments.

It can be easily observed from the upper panels of Figures 3 and 4 that the RE values obtained by the proposed method tend to decrease monotonically throughout the iterative process. Besides, after around 20 iterations, the GDL method attains the smallest or the second smallest RE values among the six methods. In the final output, the GDL method also achieves the comparatively small RE values in all of the experiments, as shown in Figure 5(a) and Figure 6(a). On the average, the proposed algorithm outperforms the other five methods in both series of experiments, as shown in 5(b) and Figure 6(b). This demonstrates the excellent capability of the proposed algorithm in reconstructing the input signals.

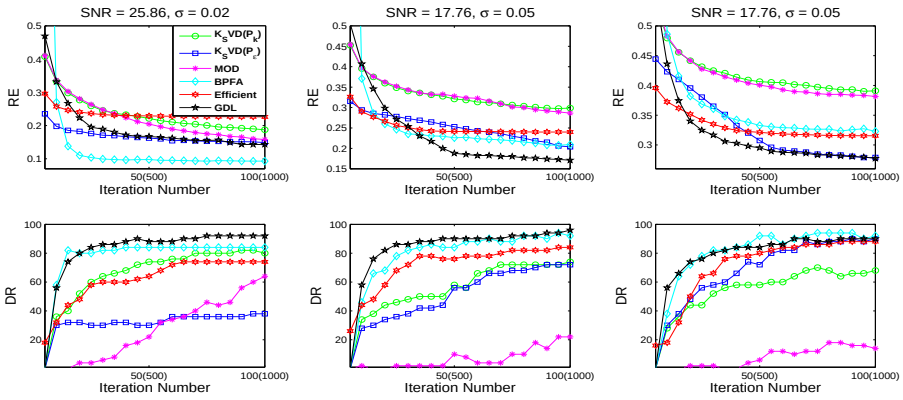

Figure 4: The upper panels: The RE curves of the six utilized methods in the iterative processes of three of the second series of experiments. The lower panels: the corresponding DR curves.

Furthermore, from the lower panels of Figures 3 and 4, it can be observed that the DR curves of the proposed method tend to increase monotonically in all experiments, and the method obtains the largest or the second largest values among the six methods after 60 iterations. Moreover, by observing Figure 5(c) and Figure 6(c), the proposed algorithm apparently yields the most stable DR values among the six methods, especially the second series of experiments, where the input signals are with heterogeneous sparsity structures. The proposed method successfully detects more than 85% atoms of the original dictionary in each of the experiments, and achieves the second largest average DR value in the first experimental series (only unsubstantially smaller than the Efficient method), and the largest in the second, as shown in Figure 5(d) and Figure 6(d), respectively. This substantiates the good capability of the GDL algorithm in recovering the original dictionary.

In the next section, we further verify the effectiveness of the proposed method on image reconstruction from data with more complex intrinsic structures and more complicated nonhomogeneous noises.

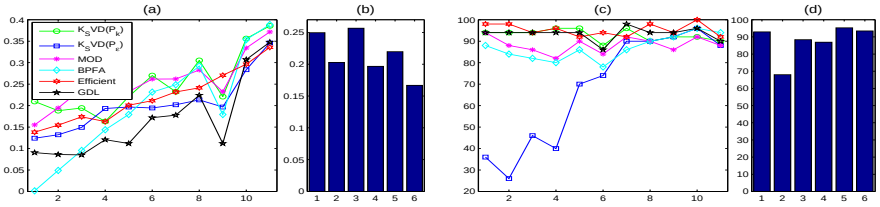

Figure 5: (a)(c): the final RE and DR values obtained by the six utilized methods in 11 experiments of the first series. (b)(d): The mean RE and DR values obtained by the six methods in the first series of experiments. The numbers 1-6 in the horizonal axis stand for the K-SVD$_{P_k}$, K-SVD$_{P_\varepsilon}$, MOD, BPFA, Efficient and GDL methods, respectively.

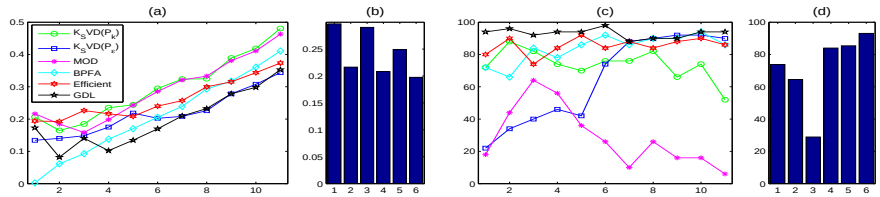

Figure 6: (a)(c): the final RE and DR values obtained by the six utilized methods in 11 experiments of the second series. (b)(d): The mean RE and DR values obtained by the six methods in the second experimental series.

4.2. Real image experiments with nonhomogeneous noise

A series of test images of 512×512 or 256×256 pixels were utilized for the image reconstruction problems. These images were generated by combining 6 gray-scale images, all of which are widely used in the image processing literature [31], with different levels of nonhomogeneous noise. In our experiments, four types of nonhomogeneous noise were employed for each image, constituting four series of experiments listed as follows.

Experiment 1 (E1): Nonhomogeneous Gaussian noise with extent δ. The standard deviation of Gaussian noise increasing uniformly from 0 for lower-right pixels to δ for upper-left pixels across the image. For each of the original 6 images, 8 noisy images of this type were generated, with noise extents $\delta = 10.22, 30.66, 51.10, 61.32, 81.76, 102.20, 127.75, 153.30$, respectively.

Experiment 2 (E2): Salt-pepper noise with extent p. Corrupting the image

with p percentage of dead pixels with either maximum or minimum intensity values. For each image, 6 noisy images of this type were utilized, with noise extents $p = 2, 6, 10, 12, 16, 20$, respectively.

Experiment 3 (E3): Mixture of homogeneous Gaussian and salt-pepper noise with extent (σ, p). Mixing the image with the combination of homogeneous Gaussian noise with deviation σ and salt-pepper noise with extent p. For each image, 4 corrupted images of this type were used, with noise extents $(\sigma, p) = (20, 5), (20, 10), (40, 5), (40, 10)$, respectively.

Experiment 4 (E4): Mixture of nonhomogeneous Gaussian and salt-pepper noise with extent (δ, p). Corrupting the image with the mixture of nonhomogeneous Gaussian noise with extent δ and salt-pepper noise with extent p. For each image, 5 noisy images of this type were used, with noise extents $(\delta, p) = (20.44, 4), (20.44, 10), (51.10, 4), (51.10, 10), (76.65, 10)$, respectively.

In each experiment, the dictionary was trained on the overlapping patches, of 8×8 pixels (i.e., the input signals are with dimension $d = 64$), of input images, and thus each experiment includes $n = (256 - 7)^2 = 62,001$ patches (all available patches from the 256×256 images, and every second patch from every second row in the 512×512 images). In each of the experiments, the dictionary D contains $m = 256$ atoms. For each utilized dictionary learning method, the images were rebuilt by averaging the overlapping reconstructed patches over the dictionary attained by the method.

Three of the current methods, DCT [2], K-SVD$_{P_\varepsilon}$ [4] and BPFA [6], were also applied to these images for comparison. The dictionary utilized by the DCT method is the overcomplete DCT bases, while dictionaries of K-SVD$_{P_\varepsilon}$, BPFA and GDL, were trained from the images. The randomly selected image

Table 1: Summary of the PSNR results of the image experiments. The best result in each experiment is highlighted. The last row is the average results of the four utilized methods over all noise cases for each image.

	Lena				Barbara				Boat			
	DCT	K-SVD	BPFA	GDL	DCT	K-SVD	BPFA	GDL	DCT	K-SVD	BPFA	GDL
$E1$	26.77	26.99	22.96	**27.55**	24.44	**24.69**	23.73	24.02	25.10	25.28	20.07	**25.54**
$E2$	25.49	25.33	15.95	**25.73**	23.06	22.56	15.85	**23.12**	24.15	23.78	16.06	**24.58**
$E3$	25.59	25.69	18.87	**26.33**	23.26	23.15	18.02	**23.52**	24.18	24.10	19.72	**24.94**
$E4$	25.70	25.62	17.34	**26.35**	22.98	22.90	17.17	**23.49**	24.19	24.03	18.41	**24.85**
Average	25.89	25.91	18.78	**26.49**	23.44	23.32	18.69	**23.54**	24.41	24.30	18.57	**24.98**

	Figureprint				House				Peppers			
	DCT	K-SVD	BPFA	GDL	DCT	K-SVD	BPFA	GDL	DCT	K-SVD	BPFA	GDL
$E1$	21.79	22.14	23.12	**22.67**	26.45	27.03	23.83	**28.79**	24.62	25.10	23.48	**25.93**
$E2$	21.09	20.70	15.91	**22.10**	25.51	25.22	16.02	**27.06**	23.61	23.44	15.84	**24.65**
$E3$	21.18	21.12	18.71	**22.40**	25.73	25.67	19.60	**27.37**	23.72	23.84	18.39	**24.89**
$E4$	21.07	20.96	17.55	**22.54**	25.48	25.51	18.47	**27.64**	23.61	23.88	17.18	**24.81**
Average	21.28	21.23	18.82	**22.43**	25.80	25.86	19.48	**27.72**	23.89	24.06	18.72	**25.07**

patches were used as the initialization of the K-SVD$_{P_\varepsilon}$ and GDL methods, and the SVD-based initialization was used for BPFA. The K-SVD$_{P_\varepsilon}$ and GDL results were obtained by 10 iterations, and that of BPFA was achieved by 30 iterations of Gibbs sampling. Since both DCT and K-SVD$_{P_\varepsilon}$ need to preset the parameter that evaluates the mean noise deviation of the entire image pixels, we implemented both methods 10 times on each test image under different initializations of this parameter and only recorded the best one as the final result. The global sparsity parameter K of the proposed GDL method was set as 15000 for all experiments.

The performance of each utilized method is quantitatively measured by the average PSNR value of the reconstructed images in each series of E1, E2, E3, E4 experiments for each of the 6 images, respectively. The results are listed in Table 1. Besides, by looking through the numbers of atoms re-

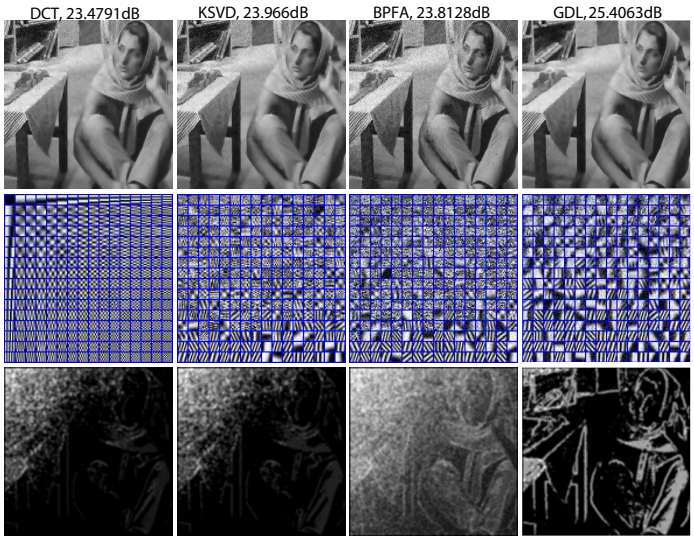

Figure 7: Results on the Barbara image mixed with nonhomogeneous Gaussian noise with extent 61.32. The panels from top to bottom: the reconstructed images, the dictionaries and the atom-using-frequency figures obtained by the DCT, K-SVD, BPFA and GDL methods, respectively.

quired for representing the image patches (i.e., the numbers of the non-zero elements in the corresponding representation coefficients) and averaging the results over the entire image, an atom-using-frequency figure, of the same resolution as the original image, can be achieved. For the atom-using-frequency figures so constructed, the brighter is a pixel, the more atoms are assigned to represent the image patches containing the pixel, and thus the more emphasis is placed on the region around the pixel by the corresponding method, and vice versa. Therefore, such an atom-using-frequency figure qualitatively reflects the intrinsic image structure explored by the utilized method.

For easy evaluation, Figures 7 and 8 depict the recovered images, along with their PSNR values obtained by applying the DCT, K-SVD, BPFA and GDL methods to two typical test images mixed with different types of noises, respectively. The corresponding dictionaries and atom-using-frequency figures attained by these methods are also displayed in these figures. The

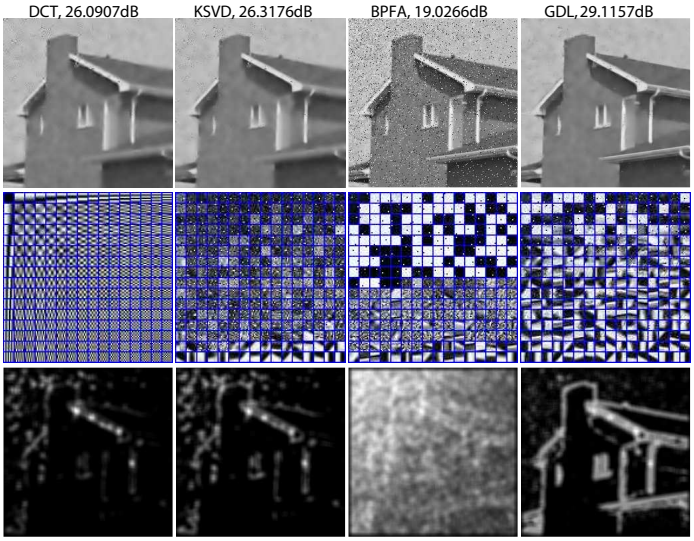

Figure 8: Results on the House image corrupted by mixture of nonhomogeneous Gaussian noise and salt-pepper noise with extents $\delta = 51.10$ and $p = 4$. The panels from top to bottom: the reconstructed images, the dictionaries and the atom-using-frequency figures obtained by the DCT, K-SVD, BPFA and GDL methods, respectively.

advantage of the proposed method can be easily observed from these results mainly in the following three-fold aspects. First, our algorithm best rebuilds the original images among all employed methods. Specifically, as compared with the DCT, K-SVD and BPFA methods, our method achieves the largest PSNR values among all competing methods for each experiment. This advantage can also be visualized in the first rows of Figures 7 and 8. It can be seen that by our method, the noise is most prominently removed from the noisy images, e.g. the bookshelf of the Barbara image in Figure 7, and the details of the original image are mostly recovered, e.g. the windows of the House image in Figure 8 (the details can be better seen by zooming in onto the images in the computer). These results show the excellent capability of our algorithm in reconstructing the original images.

Second, the proposed GDL method more robustly attains the proper dictionary underlying the images as compared with the other dictionary learning

methods. This can be easily observed in the second rows of Figures 7 and 8. The dictionaries attained by our method evidently capture more meaningful features underlying the images and are least affected by the noise in all cases. These results validate the capability of the proposed algorithm in properly generating the dictionary for images corrupted by nonhomogeneous noises.

Third, the atom-using-frequency figures obtained by our method faithfully reflect the intrinsic structures underlying the images. From the third rows of Figures 7 and 8, it can be observed that the atom-using-frequency figures obtained by our method clearly depict the basic edge information underlying the images. This is due to the fact that the patches around the image edges are of relatively complicated structures, and our method thus adaptively assigns more atoms to represent these image patches. In comparison, such meaningful structures are not so noticeably detected by the atom-using-frequency figures of the other utilized methods in the experiments. These results demonstrate the capability of the GDL method in detecting meaningful structure information underlying the images at the global scale.

4.3. Stability testing experiments

In this section, we want to further evaluate the stability of the proposed algorithm on different settings of global sparsity K and the initial coefficient matrix A. Also, we want to demonstrate the details of how the sparsity in the coefficient matrix changes in the iterative process of the column-updating and row-updating of our algorithm, to further clarify its intrinsic mechanism.

Like Section 4.1, we also constructed a series of signals, each having 1500 20-dimensional signals, denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{1500}] \in R^{20\times 1500}$, respectively. The signals were generated by a linear combination of a dictio-

nary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_{50}] \in R^{20\times 50}$ and representation coefficients $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{1500}] \in R^{50\times 1500}$, and mixed with Gaussian white noise with standard deviation 0.02. Different from Section 4.1, however, the coefficient matrix $\mathbf{A}$ has a more complicated sparsity structure: the number of nonzero elements in each column $\mathbf{a}_i$ of $\mathbf{A}$ is rounding from the Gaussian distribution $N(3, 3)$, and their positions are just randomly located. This means that the groundtruth sparsity of the coefficient matrix is not known in prior.

We employed the following two series of initializations of our algorithm for experiments: (1) The first 60 columns $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{60}$ of A are randomly valued, and the rest 1440 columns are simply set as zero vectors (the corresponding global sparsity parameter K is thus $60 \times 50 = 3000$). (2) A series of coefficient matrixes, with sparsities varying from 3000 to 6000 with interval 100, are specified, respectively. The nonzero entries of each coefficient matrix is randomly located. By (1) initialization, we want to depict the capability of our algorithm on adaptively and dynamically adjusting the sparsity (i.e., k_i in Eqn. (1)) of coefficients to appropriately represent signals, even on such singular specification; and by (2) initializations, we aim to show the stability of our algorithm with respect to different settings of the global sparsity K and the initial coefficient matrix A.

The left upper row of Figure 9 compares the sparsity diversity of the representation coefficients for all 1500 input signals in the 1, 3, 8, 50 iterations of our algorithm under (1) initialization, respectively; and the left lower row of the figure shows the standard deviations of the reconstructed signals from the original ones in these steps, correspondingly. It can be easily observed from this figure that although only the first 60 columns of A

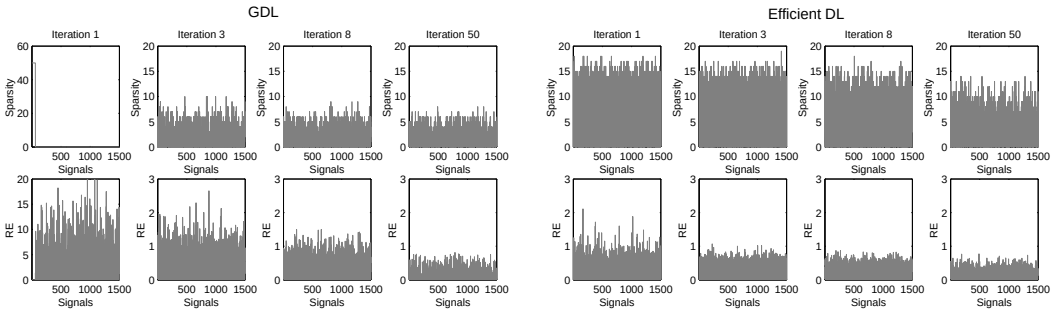

Figure 9: The left upper row: the sparsity diversity of the representation coefficients for all 1500 input signals in the 1, 3, 8, 50 iterations of the proposed algorithm under (1) initialization. The left lower row: the standard deviations of the reconstructed signals from the original ones in the 1, 3, 8, 50 iterations, respectively. The right upper and lower rows: the corresponding performance of the Efficient method.

are pre-specified as nonzeros, the nonzero elements is to be automatically scattered to all columns of A only after several iterations of the proposed algorithm. The representational errors for the signals are evidently decreasing during the implementation process of our algorithm, implying the nonzero elements of A tend to be gradually rearranged to the proper positions based on the various structures of the entire signal set. As comparison, we also implemented the Efficient method [7], which is constructed on the l_1-norm model P_λ, on this signal set (we have tried 10 different λs and selected the best one as final result). The right upper and lower rows of Figure 9 depict the diversity of the coefficient sparsity and standard deviations of signals in the 1, 3, 8, 50 iterations of this method. Since this l_1 minimization method pre-specifies the penalty λ on P_λ while not the sparsity k on P_k, the sparsities of signals can also be tuned in the iterations to a certain extent. It, however, always needs more nonzero elements (3.56 versus 2 in average) to achieve the comparable deviation (0.343 versus 0.338 in average) with the proposed method, as clearly depicted in the figure. This substantiates that

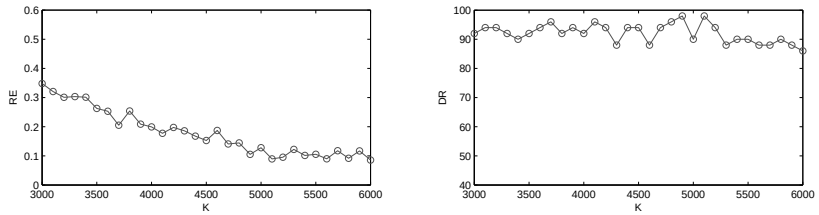

Figure 10: The RE and DR curves of the proposed algorithm with global sparsity parameter K varying from 3000 to 6000 with interval 100.

the introduced global sparsity constraint in Eqn. (1) does bring a flexible sparsity control mechanism to dictionary learning, and the GDL algorithm tends to adaptively represent different signals with proper sparsities and fittingly recover the original signals.

Figure 10 depicts the performance of the proposed GDL algorithm, in terms of RE and DR values, respectively, under (2) initializations. It can be observed that the DR values of all experiments are stabilized at the interval between 88 and 98. Besides, the RE values tend to be decreasing with K increasing since more nonzero elements are involved in the coefficient matrix, while after around $K = 5000$, the performance also becomes not sensitive to the preset values of K. This verifies that the proposed algorithm can perform stably well under different settings of the global sparsity K.

5. Conclusion and discussion

In this paper we have proposed a novel dictionary learning method for signals. Instead of enforcing uniform sparsity constraint on the coefficient vector of each input signal like the previous methods, the new method imposes global sparsity constraint on the coefficient matrix of all training signals, which makes the new method capable of adaptively assigning atoms for representing the various signals and fitting to the intrinsic signal structures

at the global scale. An efficient algorithm has also been correspondingly developed, which is easy to be implemented based on the sparse coding and sparse PCA techniques, and is guaranteed to be convergent. Based on the experimental results on a series of signal and image data sets, it has been substantiated that as compared with the current dictionary learning methods, the proposed method can more faithfully deliver the ordinary dictionary and properly reconstruct the input signals. Besides, it has been theoretically analyzed and empirically verified that by utilizing the proposed method, the atoms of the dictionary can be appropriately adapted to represent signals with various intrinsic complexities, and the frequency of atom-using can facilitate revealing the intrinsic structure underlying the input signals.

5.1. On computational complexity of GDL

Here we want to briefly discuss the complexity of the proposed method. The computational complexity of the proposed algorithm is essentially determined by the iterative process between the column and row updating steps. By employing the recent sparse coding and sparse PCA technologies, e.g., OMP [25] and sPCA-rSVD algorithms [28], respectively, both steps can be efficiently performed, requiring around $O(dnm\widehat{k}) \times T$ computational cost, where $\widehat{k}$ is the maximal value of k_i^cs, and T is the iteration number of the algorithm[2]. That is, the computational time of the proposed algorithm in-

[2]In each iteration of the proposed algorithm, the optimization model (2) needs to be solved for $i = 1, \cdots, n$, each requiring $O(dmk_i^c)$ cost from utilizing the OMP algorithm [4, 25], and the model (3) needs to be solved for $i = 1, \cdots, m$, each costing $O(dn)$ computation from employing the sPCA-rSVD algorithm [28]. Thus, the total computational complexity of the proposed algorithm is around $O(dnm\widehat{k}) \times T$.

creases linearly with the dimensionality and the size of the input signals, as well as the number of atoms in the dictionary. The computational complexity of the proposed algorithm is comparable to that of the current dictionary learning algorithms [4, 5, 12].

It should be noted that both the OMP and sPCA-rSVD methods employed in steps 2.1 and 2.2 of the proposed algorithm contain only simple computations. No complicated operations like matrix inverse calculation, eigenvalue decomposition and equation set solving are involved. The proposed method thus can always be efficiently implemented. For example, as compared with the Efficient method [7] constructed on the l_1 minimization problem, which costed 200.25s and 283.97s in average in two series of signal experiments, the proposed method only spent 82.87s and 85.48s.

5.2. On potential applications of GDL

In the paper, we demonstrate the applications of the proposed method to image denoising and edge detection. There are actually many other practical tasks the dictionary learning techniques can handle, including image deblurring, image impainting, image super-resolution, image classification, and etc. [32, 33]. Here we want to list some of the potential applications of the proposed method based on its specific adaptive-sparsity-arranging capability: (1) Image content assessment: Through adjusting the global sparsity K of the proposed method on a certain image such that the reconstruction error is smaller than some pre-specified small threshold, the magnitude of K so attained can then be used to measure the complexity of the content contained in the image. For example, the cartoon image generally contains only simple strokes and is thus expected to be perfectly reconstructed under

a small global sparsity K by our method, while a real image is always contains more complicated contexts, and has to be finely reconstructed under a comparatively large sparsity K by our method. Thus by comparing the value of the sparsity K so attained, we can then make a quantitative image content assessment, which is potentially useful for image categorization and taxonomy. (2) Object location: First learn a dictionary from images containing specific objective, e.g., faces, and then represent a new image under a small sparsity K by the proposed method under this dictionary. Since our method can adaptively arrange the K nonzero elements into the right positions of the coefficient matrix to make the representation error of the entire signals possibly small, these nonzero entries are expected to be adapted to the face area of the image since this area is more hopeful to be exactly reconstructed by the dictionary learned from faces. By detecting the atom-using-frequency image, the face can then be located in the image. (3) Virtual attention simulation: It should be noted that when the global sparsity K is set small, only small area of the image can be emphasized by virtue of the atom-using-frequency image obtained by our method. Such area reflects the most noticeable part in the image by humans, e.g., edges and peaks of the objects. When K is gradually specified larger, the atom-using-frequency image attained by our method tends to highlight more and more parts of the input image. This process complies with the real virtual phenomenon of human being. That is to say, it is hopeful to employ the proposed method to simulate the virtual attention mechanism of human by performing the proposed method under varying sparsities K. We thus expect to extract the physiological explanation of our method in our future research.

5.3. On future investigations of GDL

Other problems required to be further investigated include: (1) the effectiveness of the proposed algorithm requires to be further testified in real signals with complicated noise types, e.g., the poisson noise; (2) Qualitatively speaking, the more complex is the entire structure or the less noise is contained in an image, the larger the global sparsity parameter K should be properly preset. Investigation, however, still needs to be made to design an automatic quantitative parameter selection strategy to further improve the quality of the proposed method; (3) Research is needed to further improve the efficiency of the proposed algorithm by virtue of the online [14] or convexification [7] techniques.

Acknowledgement

This research was supported by the Geographical Modeling and Geocomputation Program under the Focused Investment Scheme at The Chinese University of Hong Kong, the National Grand Fundamental Research 973 Program of China under Grant No. 2013CB329404, the China NSFC project under contract 11131006 and Ph.D. Programs Foundation of Ministry of Education of China 20090201120056.

References

[1] S. Mallat, A wavelet tour of signal processing, Academic Press, 2009.

[2] O.G. Guleryuz, Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part I-Theory,

Part II-Adaptive algorithms, IEEE Trans. Image Process., vol. 15, no. 3, pp. 539-571, Mar. 2005.

[3] M. Aharon, M. Elad, A.M. Bruckstein, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process., vol. 54, no. 11, pp. 4311-4322, Nov. 2006.

[4] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process., vol. 15, no. 12, pp. 3736-3745, Dec. 2006.

[5] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images, IEEE Trans. Image Process., vol. 21, no. 1, pp. 130-144, Jan. 2011.

[6] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, L. Carin, Nonparametric bayesian dictionary learning for sparse image representations, in NIPS, 2009.

[7] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, in NIPS, 2007.

[8] J.P. Shi, X. Ren, G. Dai, J.D. Wang, Z.H. Zhang, A non-convex relaxation approach to sparse dictionary learning, in CVPR, 2010.

[9] G.P. Tolstov, Fourier series, Courier-Dover, 1976.

[10] E.J. Candes, D.L. Donoho, Recovering edges in ill-posed inverse prob-

lems: Optimality of curvelet frames, Ann. Statist., vol. 30, no. 3, pp. 784-842, Jun. 2002.

[11] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 9, pp. 891-906, Sep. 1991.

[12] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res., vol. 37, no. 23, pp. 3311-3325, Dec. 1998.

[13] K. Engan, S.O. Aase, J.H. Husϕy, Multi-frame compression: Theory and design, EURASIP Signal Process., vol. 80, no. 10, pp. 2121-2140, Oct. 2000.

[14] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in ICML, 2009.

[15] R. Jenatton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for sparse hierarchical dictionary learning, in ICML, 2010.

[16] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman. Supervised dictionary learning, in NIPS, 2009.

[17] J. Mairal, M. Elad, and G. Sapiro, Sparse representation for color image restoration, IEEE Trans. Image Process., vol. 17, no. 1, pp. 53-69, Jan. 2008.

[18] D.A. Spielman, H. Wang, J. Wright, Exact recovery of sparsely-used dictionaries, in COLT, 2012.

[19] L. Shen, C.H. Yeo, Intrinsic images decomposition using a local and global sparse representation of reflectance, in CVPR, 2011.

[20] F.A. Razzaq, S. Mohamed, A. Bhatti, S. Nahavandi, Non-uniform sparsity in rapid compressive sensing MRI, in SMC, 2012.

[21] P. Favaro, R. Vidal, A. Ravichandran. A closed form solution to robust subspace estimation and clustering. In CVPR, 2011.

[22] R. Gribonval, K. Schnass, Some recovery conditions for basis learning by L1-minimization, in ISCCSP, 2008.

[23] R Gribonval, K Schnass, Dictionary identifiability from few training samples, in EUSIPCO 2008.

[24] K. Bredies, D.A. Lorenz, Iterated hard shrinkage for minimization problems with sparsity constraints, SIAM J. Sci. Comput., vol. 30, np. 2, pp. 657-683, 2008.

[25] Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in Proc. Ann. Asilomar Conf. Signals Systoms and Computers, pp. 40-44, 1993.

[26] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat., vol. 15, no. 2, pp. 265-286, Jun. 2006.

[27] M. Journée, Y. Nesterov, P. Richtárik, R. Sepulchre, Generalized power method for sparse principal component analysis, J. Mach. Learn. Res., vol. 11, pp. 517-553, Feb. 2010.

[28] H.P. Shen, J.Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivar. Anal., vol. 99, no. 6, pp. 1015-1034, Jul. 2008.

[29] D.Y. Meng, Q. Zhao, Z.B. Xu, Improving robustness of sparse PCA by l_1-norm minimization, Pattern Recognit., vol. 45, no. 1, pp. 487-497, Jan. 2012.

[30] C.D. Sigg, J.M. Buhmann, Expectation maximization for sparse and non-negative PCA, in ICML, 2008.

[31] J. Portilla, V. Strela, M.J. Wainwright, E.P. Simoncelli, Image denoising using scale mixtures of gaussians in the wavelet domain, IEEE Trans. Image Process., vol. 12, no. 11, pp. 1338-1351, Nov. 2003.

[32] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proceedings of the IEEE, vol. 98, no. 6, pp. 1031-1044, Jun. 2010.

[33] M. Elad, M.A.T. Figueiredo, and Y. Ma, On the role of sparse and redundant representation in image processing, Proceedings of the IEEE, vol. 98, no. 6, pp. 972-982, Jun. 2010.