Transform Learning Algorithm Based on the Probability of Representation of Signals

Gayatri Parthasarathy and G. Abhilash
Department of Electronics and Communication Engineering
National Institute of Technology Calicut, Kerala, India 673601
Email: gayatri_p120094ec@nitc.ac.in, abhilash@nitc.ac.in

Abstract—**Compressed sensing is a signal acquisition scheme that measures signals at sub-Nyquist rate amenable to sparse recovery, with high probability, from a reduced set of measurements. One of the main requirements of compressive sensing is the sparsity of the class of signals of interest in some basis. A method to construct a sparsifying basis for a class of signals using information theoretic measures is proposed in this paper. The algorithm constructs the sparsifying basis from a known non-sparsifying basis by concentrating the probability distribution of the basis in the representation of a class of signals. Simulation studies using speech and image signals confirm that the basis constructed using the proposed method results in an improved sparsity of the signals with thresholded coefficients but without degrading the signal quality.**

Index Terms—**Transform learning, Sparse representation, Compressed sensing, Representation entropy, Sparse modeling.**

I. Introduction

Compressive sensing is a method of capturing signals into significantly less number of samples than stipulated by the classical sampling theorem. The key to successful recovery from these reduced set of random measurements lies with the sparsity of the signal. Most natural signals belonging to a high dimensional space have an underlying low-dimensional structure in an appropriate model, which enables their sparse representation [1].

An N-dimensional signal x is said to be K-sparse when the representation of the signal in some known basis has l_0-norm[1] equal to K, with $K \ll N$. For this signal, improving the sparsity means reducing K. Increased sparsity allows proportionately large reduction in the number of measurements without affecting stable recovery [2]. In general, the sparsifying basis for a signal is not known. Hence, the construction of the sparsifying dictionaries or transforms for a class of signals through learning becomes important.

Recently, there is large interest on the construction of overcomplete sparsifying dictionaries by learning from the data itself [3] - [7]. There are two well known sparsity models [8] to which the dictionary learning approach is applied. One is the synthesis sparsity model which states that a linear combination of a few of the dictionary atoms is sufficient to represent the signal [8] [9]. The other is analysis sparsity model which suggests that the representation of a signal in the dictionary is sparse [8] [9]. Ravishankar and Bresler explored the transform sparsity model and proposed a parameter dependent transform learning (TL) method for square transforms in [9], for orthogonal transforms (TLortho) in [11] and for overcomplete transforms in [10]. In [12], Eksioglu *et al* proposed a parameter independent transform learning method called the Transform K-SVD (T-KSVD).

The condition number of the transform and the compressibility of the signals relative to the transform generated by TL algorithm [9] depend strongly on the parameters chosen. The orthogonal transform learning of [11] is parameter independent and gives good performance in attaining the specified sparsity. Tipping and Bishop [13] proposed probabilistic Principal Component Analysis (PPCA), where the PCA is effected using an iterative expectation maximization algorithm.

The algorithm proposed in this paper is inspired by the methods proposed in [9], [11], [12] and [13]. The proposed algorithm (section III) is to learn a square sparsifying transform for a class of signals from a known non-sparsifying transform. It identifies an orthogonal sparsifying basis by iteratively maximizing the concentration of the probability distribution of the representation basis for a class of signals. The probability of representation used in this paper is defined in section II, and is not the same as the statistical probability of the data set used in [13]. The experiments with speech and image signals (section IV) show that the algorithm is capable of constructing transforms that capture the underlying low-dimensional structure of the class of signals. This method can find applications in the dimension reduction of a dense data set for sparse modeling. The algorithm depends on only one parameter and the number of iterations is fixed. Hence the algorithm provides freedom in its application to any class of signals, without changing its framework.

II. Probability of representation Based transform learning (PTL)

To derive the algorithm for transform learning using information theoretic approach, we need to define the probability of selecting each basis vector that represents the signal, and the associated Shannon entropy of representation. In the sequel, Shannon entropy and entropy are used interchangeably.

Definition 1: [14] Let $\Phi = \{\phi_i\}_{i=1}^{N}$ be an orthonormal basis of an N-dimensional space. Let x be a normalized signal

[1] l_0, the number of non-zero elements in a vector, does not satisfy the homogeneity property required for a norm. We use the term l_0-norm for readability.

ISBN 978-0-9928626-7-1 © EURASIP 2017

belonging to a class of signals $\boldsymbol{X}$ such that $x = \sum_{i=1}^{N} b_i \phi_i$, where $b = [b_1, b_2, ...b_N]^T$ is the vector of representation coefficients of x relative to Φ and $b_i = \langle x, \phi_i \rangle$. The probability of choosing ϕ_i in the representation of x relative to Φ is $p_i = |b_i|^2$.

The basis Φ is not unique. Hence the basis relative to which a class of signals has sparse representation may exist such that the maximum information of the signals is concentrated on a small set of vectors of the representing basis. Hence to identify the maximum information bearing vectors of the basis, the entropy of representation relative to such vectors should be minimum. This entropy of representation is conditioned to Φ, which means that the amount of information left in x is $H(x \mid \Phi)$. The conditional entropy $H(x \mid \Phi)$ depends only on the probability of representation of x relative to Φ, which is $p = \{p_i\}_{i=1}^{N}$ as stated in definition 1. Since $H(x \mid \Phi)$ depends on p which in turn depends on b, $H(x \mid \Phi) = H(b)$, where $H(b)$ is the Shannon entropy of representation of x relative to Φ. Thus,

$$H(x \mid \Phi) = -\sum_{i=1}^{N} p_i ln(p_i), \tag{1}$$

where, by definition $0 \times ln(0) = 0$ [17].

Let $\Psi = \{\psi_i\}_{i=1}^{N}$ be another orthonormal basis for the class of signals $\boldsymbol{X}$. The representation of x in Ψ is $x = \sum_{i=1}^{N} c_i \psi_i$ with $c = [c_1, c_2, \ldots, c_N]^T$ as the vector of representation coefficients of x with respect to the basis Ψ and $c_i = \langle x, \psi_i \rangle$. Hence the probability of selecting the i-th basis function of Ψ is $q_i = |c_i|^2$. The entropy of representation of the signal in the basis Ψ is $H(c) = H(x \mid \Psi) = -\sum_{i=1}^{N} q_i ln(q_i)$.

The entropy $H(c)$ is low when maximum information is captured by a small number of basis vectors, and the coefficients of representation, arranged in the descending order of magnitude, follow the power law decay as $|c_i| \leq Ri^{-r}$, where $R > 0$ and $r > 1$. In general, natural signals are not strictly sparse in any basis, rather they are compressible in some basis.

The algorithm proposed in this paper constructs a sparsifying basis Ψ from a known non-sparsifying basis Φ using theoretical dimension as the measure of sparsity. The relation between the theoretical dimension n_{th}^{Ψ} in a basis Ψ and the entropy of representation $H(c)$ is given by [15] [16]

$$n_{th}^{\Psi} = \lceil exp(H(c)) \rceil. \tag{2}$$

The advantage of theoretical dimension lies in the fact that it gives the number of basis vectors required to capture atleast 90% of the signal energy. Hence it gives a measure for quantifying the sparsity of a compressible signal.

If the two representation bases Φ and Ψ are known, the mutual coherence (μ) between the two bases is given by [18]

$$\mu = \max_{1 \leq i,j \leq N} \langle \phi_i, \psi_j \rangle, \tag{3}$$

where $\frac{1}{\sqrt{N}} \leq \mu \leq 1$. The relation between the entropies of representation of a signal in these bases is given by the entropy uncertainty relation [19] - [21].

$$H(b) + H(c) \geq -2ln(\mu). \tag{4}$$

If the value of μ is low, and the entropy of representation $H(b) \geq -2ln(\mu)$, then the representation entropy $H(c)$ can attain its minimum value, zero. In other words, if the theoretical dimension of the signal relative to Φ is at least μ^{-2} (non-sparse representation), then the theoretical dimension of the representation in Ψ can be unity (sparsest representation).

While learning the transform for a class of signals,we would need the representation entropy for all the signals in the class to be the lowest, ideally. But, $H(c) = 0$ means that the probability of selecting a particular basis vector is unity. Since we are learning the transform for a class of signals, this lower bound can hold good for a maximum of N signals, ideally. So, in general, the lower bound on the representation entropy would be non-zero; that is $H(c) > 0$.

Remarks: (1) In the algorithm proposed, μ is not used as a parameter to construct Ψ. (2) It is not necessary that Φ be a non-sparsifying basis for the algorithm to hold good.

III. Problem Formulation

Dimension reduction can be attained through the reduction of representation entropy. To maximally reduce the entropy of representation relative to a basis, it is sufficient to make the probability distribution of the basis, in the representation of the signal, maximally concentrated.

Consider a normalized basis vector ψ_i and a signal x. By Definition 1, the probability of choosing the basis vector ψ_i to represent the signal x can be given by $|\langle \psi_i, x/\|x\|_2 \rangle|^2$. The sparsest representation would be obtained if the probability of selecting a basis vector is unity. This would give us the lowest entropy $H(c) = 0$ and the theoretical dimension $n_{th}^{\Psi} = 1$. This knowledge is used for finding the basis relative to which the class of signals assumes low n_{th}^{Ψ}.

Considering a set of training signals as columns of the matrix X, we attempt to find a basis vector which would have a probability close to one in the representation of all the signals in the training set. Let P_d be the row vector of the desired probabilities (row of 1's as the initial vector) for all the signals in the training set. Let ψ_i be the basis vector that is to be learned such that its probability in the representation of the signals in the training set is as close to P_d as possible. The probability of selecting the basis vector ψ_i in the representation of the signal X_j is $|\psi_i^T X_j|^2$, where X_j is the j-th column of X. The vector of probabilities associated with the basis vector ψ_i in the representation of the signals in the training set is given by

$$\left(\psi_i^T X\right)^2 = \{|\psi_i^T X_j|^2\}_{j=1}^{L}, \tag{5}$$

where L is the number of signals in X. By Definition 1, (5) holds good as probability vector if and only if $\|\psi_i\|_2 = \|X_j\|_2 = 1$. Without loss of generality, we consider X to be a collection of normalized signals. The requirement $\|\psi_i\|_2 = 1$ constrains the ensuing optimization problem. Since we have taken the probability of representation to be in terms of the inner product, $|\langle \psi_i, x \rangle|^2$, we need to ensure that the basis vectors are orthogonal. Hence the optimization problem is

$$\min_{\psi_i} \|P_d - (\psi_i^T X)^2\|_2^2$$
$$\text{subject to } \langle \psi_i, \psi_j \rangle = \delta_{ij} \ ; \ j = 1, 2 \cdots i, \tag{6}$$

where δ_{ij} is the Kronecker delta. To simplify the optimization problem, we have eliminated the constraint by using an alternative approach to incorporate the constraint. The unit-norm constraint can be incorporated by normalizing the basis vector obtained as the solution to the optimization problem. The orthogonality constraint can be taken care of by including a penalty term in the objective function.

A. The Algorithm

The problem in (6) forms the first iteration in the algorithm giving the first basis vector ψ_1. As mentioned in section II, it is not possible to obtain the extreme entropy minimization for the class of signals because $H(c) > 0$. Hence, the basis vector obtained as a solution to (6) need not capture all the information present in the class of signals. Though we expect most of the information to be captured by ψ_1, practically that is not possible and its probability of representation in a few of the signals will be small. Hence, we need to find subspaces orthogonal to the subspace spanned by ψ_1 to get the complete representation of all the signals. The subsequent basis vectors can be obtained similarly, with a slight alteration in the formulation as described below.

After finding each basis vector, the transform matrix can be updated as

$$\Psi^{(i)} = \Psi^{(i-1)} \cup \{\widehat{\psi_i}\}, \tag{7}$$

where $\Psi^{(i)}$ is the matrix of basis vectors that results at the i-th iteration by maximally concentrating the probability of representation of the training signals such as to capture maximum information from the class of signals. The vector $\widehat{\psi_i}$ is orthonormal to the vectors in $\Psi^{(i-1)}$. Since the vectors in $\Psi^{(i)}$ are orthonormal, the signal estimate $\widehat{X}^{(i)}$ at the i-th iteration is given by

$$\widehat{X}^{(i)} = \Psi^{(i)} \left(\Psi^{(i)}\right)^T X. \tag{8}$$

The residual at the i-th iteration is updated as $R^{(i)} = X - \widehat{X}^{(i)}$. This ensures that $R^{(i)}$ is orthogonal to $\Psi^{(i)}$. Hence the basis vector ψ_{i+1} that has maximum probability in the representation of $R^{(i)}$ should be orthogonal to the vectors in $\Psi^{(i)}$. Thus, the orthogonality constraint of the problem (6) is ensured.

The desired probability vector, $P_d^{(i)}$ at the i-th iteration is

$$P_d^{(i)} = P_d^{(i-1)} - \left(\widehat{\psi_i}^T R^{(i-1)}\right)^2, \tag{9}$$

where $P_d^{(i-1)}$ is the desired probability vector that results at the $(i-1)$-th iteration, and $(\widehat{\psi_i}^T R^{(i-1)})^2$ is calculated as in (5), with X replaced with $R^{(i-1)}$. Hence at the i-th iteration, the optimization problem becomes

$$\psi_i^* = \arg\min_{\psi_i} \left\| P_d^{(i-1)} - \left(\frac{\psi_i^T}{\|\psi_i\|_2} R^{(i-1)}\right)^2 \right\|_2^2 + \lambda \|\psi_i^T \Psi^{(i-1)}\|_2^2. \tag{10}$$

Algorithm 1 Probability-based Transform Learning Algorithm

```
Input: Training set $X_{N\times L}$, Initial transform $\Phi_{N\times N}$
Output: Sparsifying transform $\Psi_{N\times N}$
Initialize:
    Desired probability vector $P_d^{(0)} = [1, 1, \cdots, 1]_{1\times L}$
    Residual $R^{(0)} = X$
    Sparsifying transform $\Psi^{(0)} = \{\emptyset\}$
    Penalty parameter $\lambda = 1$
for $i = 1$ to $N$ do
    Set initial value of $\psi_i = \phi_i$
    $\psi_i^* = \arg\min_{\psi_i} \left\| P_d^{(i-1)} - \left(\frac{\psi_i^T}{\|\psi_i\|_2} R^{(i-1)}\right)^2 \right\|_2^2 + \lambda\|\psi_i^T \Psi^{(i-1)}\|_2^2$
    Normalize the resulting $\psi_i^*$
        $\widehat{\psi_i} = \psi_i^* / \|\psi_i^*\|_2$
    $\Psi^{(i)} = \Psi^{(i-1)} \cup \{\widehat{\psi_i}\}$
    Update the desired probability
        $P_d^{(i)} = P_d^{(i-1)} - \left(\widehat{\psi_i}^T R^{(i-1)}\right)^2$
    Find the signal estimate
        $\widehat{X}^{(i)} = \Psi^{(i)} \left(\Psi^{(i)}\right)^T X$
    Update the residual
        $R^{(i)} = X - \widehat{X}^{(i)}$
end for
```

The term $\frac{\psi_i^T}{\|\psi_i\|_2}$ ensures the unit norm requirement of the basis vectors as stipulated by Definition 1, making $\left(\frac{\psi_i^T}{\|\psi_i\|_2} R^{(i-1)}\right)^2$ the probability vector associated with $\psi_i / \|\psi_i\|_2$. As mentioned above, since $R^{(i-1)}$ is orthogonal to $\Psi^{(i-1)}$, the basis vector ψ_i should be orthogonal to $\Psi^{(i-1)}$, ideally. The penalty term $\|\psi_i^T \Psi^{(i-1)}\|_2^2$ is introduced to ensure that ψ_i is strictly orthogonal to $\Psi^{(i-1)}$.

It is sufficient that the weight of the penalty λ is unity for all classes since orthogonality is primarily ensured by the residual. The performance of the algorithm is not altered by increasing λ, but bringing λ close to zero may make the transform non-orthogonal. The problem stated in (10) can be solved using any conventional non-linear optimization algorithm.

The algorithm continues for N iterations, where N is the dimension of the class of signals, ensuring the generation of a complete basis.

B. Convergence

We have chosen the quasi-Newton method [22] to carry out the minimization in (10), the convergence of which is well known. To establish the convergence of the proposed algorithm as a whole, we define the error terms of (10) at the end of the i-th iteration as

$$e_i = P_d^{(i-1)} - \left(\widehat{\psi_i}^T R^{(i-1)}\right)^2. \tag{11}$$

To ensure the convergence, we need $\{\|e_i\|_2\}_{i=1}^N$ to form a bounded, monotonically decreasing sequence. Since the underlying vector space is finite dimensional, and hence complete, $\{\|e_i\|_2\}_{i=1}^N$ is bounded by $\|e_N\|_2 = 0$. From (9) and (11), we have $P_d^{(i)} = e_i$. Substituting in the expression for e_{i+1},

$$\|e_{i+1}\|_2 = \left\| e_i - \left(\widehat{\psi}^{\,T}_{i+1} R^{(i)} \right)^2 \right\|_2 . \qquad (12)$$

Since $\left(\widehat{\psi}^{\,T}_{i+1} R^{(i)} \right)^2$ is a vector of non-negative numbers, $\|e_i\|_2 \geq \|e_{i+1}\|_2$ and $\{\|e_i\|_2\}_{i=1}^N$ forms a monotonically decreasing sequence. Hence the algorithm converges.

IV. Experimental results

In this section, we demonstrate the potential of the transform generated using the PTL algorithm proposed in this paper in comparison with the transforms generated using the TL algorithm [9], the T-KSVD algorithm [12], the TLortho algorithm [11] and the Principal component analysis (PCA). The TL and TLortho were studied experimentally using the softwares available in [23] and T-KSVD algorithms using the software available in [24], respectively. The values of the parameters used in the TL algorithm were chosen so as to generate a well conditioned transform, that is, the weight of log determinant penalty and Forbenius norm penalty were 10^5. The step-size of the optimization was taken to be 10^{-8} with 128 conjugate gradient iterations. The number of alternating minimization iterations considered was 20. The number of iterations for TLortho was 5. A square transform case was considered for the T-KSVD algorithm which was carried out for 70 iterations. The sparsity was fixed to 5 for TL, TLortho and T-KSVD. The initial basis used for all the algorithms was the standard ordered basis (identity matrix).

A comparison is made in terms of the theoretical dimension (n^{Ψ}_{th}). To study the efficiency of the algorithm in sparsifying the class of signals, the sparsity in representing the signals within the training set, and the signals belonging to the same class but outside the training set were calculated. The classes of training signals used were speech signals and image signals.

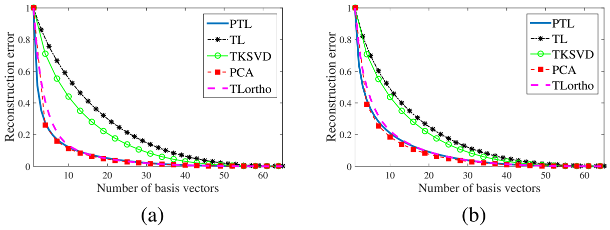
The minimization problem of (10) was solved using the quasi-Newton optimization technique [22] because it does not require the calculation of the Hessian, and hence fast. The initial vectors taken for the optimization were the columns of the identity matrix, that is $\Phi = I$.

A. Speech signals

A set of 2450 signals of dimension 64 at 8kHz sample rate were used in the training set. The performance comparison of the transforms generated using PTL, TL, TLortho, T-KSVD , and PCA for this class of signals is shown in Table I. The table gives the average n^{Ψ}_{th} of the representation of the signals, in the training set and signals belonging to the same class but outside the training set (test signals), with respect to the basis obtained by PTL, TL, TLortho, T-KSVD, and PCA. The reconstruction error energy normalized to the signal energy of a signal x is given by $\|x - x'\|_2^2 / \|x\|_2^2$, where x' is the signal reconstructed using the thresholded coefficients. The reconstruction error curves of Fig. 1 show that the compressibility of signal representation in the basis obtained from PTL is superior to the transforms obtained through other learning methods and at par with PCA, indicating that PTL generates a transform that sparsifies the representation of a class of signals.

TABLE I
Average theoretical dimension of speech signals

Signal	PTL	PCA	TL	TKSVD	TLortho	DCT
Training	9	9	32	30	11	8
Test	14	14	33	31	16	13



(a)

(b)

Fig. 1. Average normalized reconstruction error energies of speech signals (a) in the training set and (b) outside the training set

The mutual coherence μ of the generated transform with the standard ordered basis is approximately 0.3. Hence, if the entropy of representation of the signal relative to the standard ordered basis is greater than $-2ln(\mu) = 2.4$, the minimum entropy of the representation in the new basis should be zero (by (4)). But since we are learning for a class of signals, this minimum entropy situation is not attained.

B. Image signals

The algorithm was applied to a set of natural images for identifying a sparsifying basis for image signals. The training set was created by picking 8×8 non-overlapping patches from a set of images. The number of training patches used was 7725. The 8×8 image patches were converted to a set of 64-dimensional vectors that forms the set of training signals.

Fig. 2 shows an image reconstructed using five of the basis vectors that capture maximum energy. The amounts of energy captured in the five basis vectors of the basis generated by PTL, PCA, TL, TLortho, T-KSVD, and DCT are 99.98%, 99.98%, 80%, 98.5%, 95.5%, and 99.99%, respectively. The perceptual quality of the image, reconstructed using the sparse set of basis vectors identified by the PTL algorithm shows that the basis is capable of achieving 80% gain in sparsity without unduly degenerating the signal. The PSNRs (in dB) of the images in Fig. 2 (b)-(g) are 26.05, 26.04, 13.6, 21.2, 25.1, and 26.1, respectively.

C. Discussion

The performance of the algorithms were studied with different initial bases. For a given parameter setting of TL, and initial basis chosen to be DCT, the compressibility curves of the representation of the training signals in the basis obtained by TL, T-KSVD and PTL are shown in Fig. 3(a). For the same parameter setting of TL and initial basis chosen to be the standard ordered basis, the compressibility curves of the representation are shown in Fig. 3(b). It can be seen that the basis generated using the TL algorithm depends strongly on

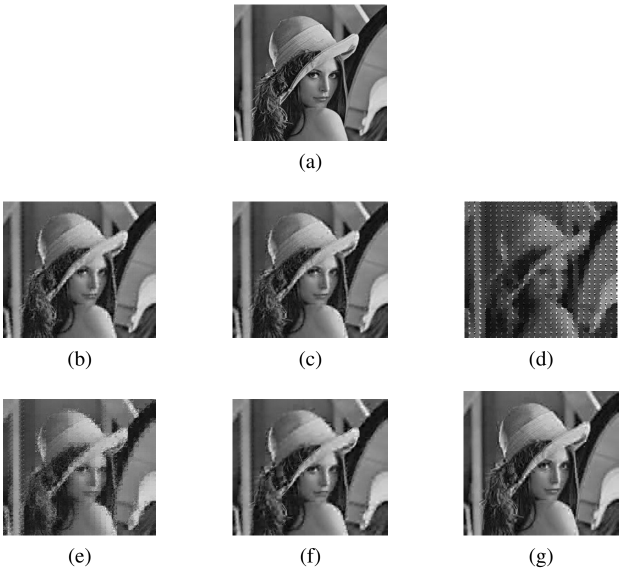


Fig. 2. (a) Original Image. Images reconstructed using five vectors in the representation bases generated by (b) PTL (c) PCA (d) TL (e) TKSVD (f) TLortho (g) DCT

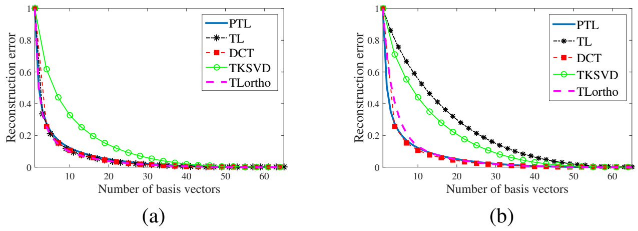


Fig. 3. Average normalized reconstruction error energies of speech training signals with initial basis set as (a) DCT and (b) standard ordered basis.

the initial basis used. The basis generated by T-KSVD and TLortho algorithms also depends on the initial basis but to a smaller extent.

When the initial basis is the standard ordered basis, the compressibility achieved with the representation basis obtained by TL algorithm can be improved by changing the parameter setting. But this shows the sensitivity of the algorithm to the parameters. The PTL algorithm presented in this paper has the least sensitivity to parameters and initial basis. The value of λ can be taken close to unity for all classes of signals. As depicted in Fig. 3, the compressibility curve of the representation of the signals in the basis generated by PTL is independent of the initial basis selected. The basis generated by selecting different initial basis may be different but irrespective of the initial basis, the PTL generates a basis that is capable of capturing the information content in the class of signals under consideration.

V. Conclusion

We have presented a novel algorithm for transform learning by concentrating the probability distribution of the basis, in the representation of a class of signals, to attain maximum sparsity. The concentration of the probability distribution leads to the reduction of dimension of the class of signals. This algorithm can be used to identify the low dimensional structure that underlies a dense collection of a given class of data. The experiments with speech and image signals confirm that the theoretical dimension of the signal relative to the new basis is reduced significantly.

References

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

[2] R. G. Baraniuk, "Compressive sensing," *IEEE signal processing magazine*, vol. 24, no. 4, 2007.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[4] B. Ophir, M. Elad, N. Bertin, and M. D. Plumbley, "Sequential minimal eigenvalues-an approach to analysis dictionary learning," in *19th European Signal Processing Conference*. IEEE, 2011, pp. 1465–1469.

[5] B. A. Olshausen *et al.*, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[6] R. Rubinstein, T. Peleg, and M. Elad, "Analysis k-svd: A dictionary-learning algorithm for the analysis sparse model," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 661–677, 2013.

[7] M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies, "Analysis operator learning for overcomplete cosparse representations," in *19th European Signal Processing Conference*. IEEE, 2011, pp. 1470–1474.

[8] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse problems*, vol. 23, no. 3, p. 947, 2007.

[9] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, 2013.

[10] S. Ravishankar and Y. Bresler, "Learning overcomplete sparsifying transforms for signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 3088–3092.

[11] S. Ravishankar and Y. Bresler, "Closed-form solutions within sparsifying transform learning," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 5378–5382.

[12] E. M. Eksioglu and O. Bayir, "K-svd meets transform learning: Transform k-svd," *Signal Processing Letters, IEEE*, vol. 21, no. 3, pp. 347–351, 2014.

[13] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[14] V. Meena and G. Abhilash, "Robust recovery algorithm for compressed sensing in the presence of noise," *IET Signal Processing*, vol. 10, no. 3, pp. 227–236, 2016.

[15] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[16] V. Meena and G. Abhilash, "Sparse representation and recovery of a class of signals using information theoretic measures," in *2013 Annual IEEE India Conference (INDICON)*. IEEE, 2013, pp. 1–6.

[17] Jan C. A Van der Lubbe, *Information Theory*, Cambridge University Press, 1997.

[18] M. Elad, *Sparse and redundant representation*. Springer, 2010.

[19] X. Guanlei, W. Xiaotong, and X. Xiaogang, "Entropic uncertainty inequalities on sparse representation," *IET Signal Processing*, 2016.

[20] H. Maassen and J. B. M Uffink, "Generalized entropic uncertainty relations," *Physical Review Letters*, vol. 60, no. 12, p. 1103, 1988.

[21] B. Ricaud and B. Torrésani, "A survey of uncertainty principles and some signal processing applications," *Advances in Computational Mathematics*, vol. 40, no. 3, pp. 629–650, 2014.

[22] D. G. Luenberger and Y. Ye, *Linear and nonlinear programming*, 3rd ed. Springer, 1984.

[23] Y. Bresler, "Transform Learning Software," http://transformlearning.csl.illinois.edu/software/index.html/, 2015, [accessed 09-April-2016].

[24] E. Eksioglu, "Transform K-SVD Software," http://web.itu.edu.tr/~eksioglue/pubs/transform_ksvd.htm/, 2014, [accessed 09-April-2016].