LEARNING SPARSE DICTIONARIES WITH A POPULARITY-BASED MODEL

Jianzhou Feng[1]*, Li Song*[1]*, Xiaoming Huo*[2]*, Xiaokang Yang*[1] *and Wenjun Zhang*[1]

[1]Institute of Image Comm. & Information Proc.,
Shanghai Jiaotong University, 200240, Shanghai, China
[2] School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

ABSTRACT

Sparse signal representation based on overcomplete dictionaries has recently been extensively investigated, rendering the state-of-the-art results in signal, image and video processing. We propose a novel dictionary learning algorithm—the PK-SVD algorithm—which assumes prior probabilities on the dictionary atoms and learns a sparse dictionary under a popularity-based model. The prior distribution brings the flexibility that is desirable in applications. We examine our algorithm in both synthetic tests and image denoising experiments.

Index Terms— Dictionary learning, sparse representation, K-SVD, PK-SVD, OMP.

1. INTRODUCTION

Sparse decomposition over a redundant dictionary has been proven as an efficient technique to convert complex signal sets into low entropy coefficient vectors, which are suitable for compression, regularization in inverse problems, feature extraction, and many more. A sparse coding model suggests that for any observed signal $y \in \mathbb{R}^n$, there exists a dictionary $D \in \mathbb{R}^{n\times K}$ that is composed of K atoms $\{d_j\}_{j=1}^K$ as its columns, and one can represents y as a linear combination of a few atoms from D. The sparsest representation is the solution of the following,

$$\min_x \|x\|_0 \quad \text{subject to } \|y - Dx\|_2 \leq \sigma, \tag{1}$$

where $\|\cdot\|_0$ is the l^0 norm (i.e., the number of nonzero entries of a vector) and σ is the noise level.

The overcomplete dictionary D that leads to sparse representations can either be chosen as a pre-defined set of functions (such as steerable wavelets, curvelets, contourlets) or be designed by adapting its content to fit a given set of signal examples. Using a pre-specified transform matrix is simple and fast, however experiments show that the learned overcomplete dictionaries are superior for signal representation. Among the developed algorithms, K-SVD [1] is an excellent one, which achieves comparable or better performance than the state-of-the-art algorithms in denoising, inpainting, compression, and so on.

We propose a novel dictionary learning algorithm (denoted by PK-SVD) that can further improve the performance of K-SVD. K-SVD is based on the sparse coding model, which assumes that every atom d_j shares the same popularity in the linear combinations. This is *not* the case for many real signals. For example, images are dominated by low frequency atoms, with some seldom used high frequency ones. The aforementioned fact motivates us to decompose images into the sub-sparsest linear combinations of low frequency atoms, rather than the sparsest representations that may contain high frequency ones. Armed by this, we modify the sparse coding model to a generative popularity-based model. We set a prior distribution on the models, which are made by subsets of atoms in D. Under the new probabilistic framework, the best dictionary and decomposition should not only minimize the representation error but also maximize the likelihood of the corresponding model.

In Section 2 we present the popularity-based model in details and propose a new objective function that is made by both reconstruction accuracy and atom group popularity for dictionary learning. In Section 3, we design an iterative algorithm to solve under the newly built formulation. Section 4 presents results of our algorithm, demonstrating its superiority. Section 5 concludes the paper.

2. THE POPULARITY-BASED MODEL

Let $y \in \mathbb{R}^n$ denote the observed signal, $D \in \mathbb{R}^{n\times K}$ denote the dictionary, and $x \in \mathbb{R}^K$ be the weight vector. We have the formulation

$$y = Dx + \varepsilon, \tag{2}$$

where $\varepsilon \in \mathbb{R}^n$ represents the Gaussian white noise with variance σ^2. An equalization form of (1) is

$$\hat{x} = \text{argmin}_{\ x} \|y - Dx\|_2^2 + \lambda\|x\|_0. \tag{3}$$

Actually, (3) can be interpreted as a maximum *a posteriori* (MAP) estimate under the following generative model, in

which T_0 is the sparsity. The generative model is:

1. **Select T_0 atoms**: Select the atoms out in turn. At the J^{th} turn, $J-1$ atoms has been selected and the rest atoms has the same possibility $\frac{1}{K-J+1}$ to be chosen. The final chosen out atoms are noted as $d_{sel_1}, \cdots, d_{sel_{T_0}}$.

2. **Generate T_0 coefficients for atoms**: Uniformly sample T_0 values from the interval $[-L, L]$ and sort them in an absolute value descending order $c_1, \cdots, c_{T_0}$ (i.e., $L \geq |c_1| \geq \cdots \geq |c_{T_0}| \geq 0$). So the probability density function is $f(c_1, \cdots, c_{T_0}) = \frac{T_0!}{(2L)^{T_0}}$.

3. **Generate noise ε**: Gaussian white noise.

4. **Generate signal y**:

$$y = \sum_{i=1}^{T_0} c_i d_{sel_i} + \varepsilon.$$

Under the generative model assumption, we have

$$\begin{aligned} p(x) &= p(\text{Support Set of x}) p(\text{Non-zero coefficients of x}) \\ &= \frac{1}{K(K-1)\cdots(K-\|x\|_0+1)} \cdot \frac{T_0!}{(2L)^{T_0}} \\ &\approx \frac{T_0!}{(2L)^{T_0}} (\frac{1}{K})^{\|x\|_0}. \end{aligned} \tag{4}$$

The last approximation comes from

$$\|x\|_0 \leq T_0 \ll K.$$

Hence the MAP estimate becomes

$$\begin{aligned} \hat{x} &= \text{argmax } p(x|y) = \text{argmax } p(x)p(y|x) \\ &= \text{argmax } (\frac{1}{K})^{\|x\|_0} e^{-\frac{\|y-Dx\|_2^2}{2\sigma^2}} \\ &= \text{argmin } \|y - Dx\|_2^2 + 2\sigma^2 \ln(K)\|x\|_0, \end{aligned} \tag{5}$$

which is the same as (3), where $\lambda = 2\sigma^2 \ln(K)$.

Furthermore, we modify the above generative model into a popularity-based one by changing the equal probability $\frac{1}{K-J+1}$ in step 1. to $p_{i,j}$, the element of a prior probability matrix $P \in \mathbb{R}^{T_0 \times K}$. If $p_{i,j}$ is large, then atom d_j is more likely to be chosen at the i^{th} turn. Using matrix P, the MAP estimate becomes

$$\begin{aligned} \hat{x} = \text{argmin } & \|y - Dx\|_2^2 - 2\sigma^2 \sum_{i=1}^{\|x\|_0} \ln(p_{i,sel_i}), \\ & \text{subject to } \|x\|_0 \leq T_0 \end{aligned} \tag{6}$$

and (6) can further lead to our new dictionary learning objective function.

Task: Find the best decomposition of y when D and P are known

$$\min_x \|y - Dx\|_2^2 - 2\sigma^2 \sum_{i=1}^{\|x\|_0} \ln(p_{i,sel_i}).$$

subject to $\|x\|_0 \leq T_0$

Initialization: $x^{(0)} = 0$, support set $\Gamma^{(0)} = \varnothing$.
For $J = 1, 2, \ldots, T_0$

- *Update Residual*: $r^{(J)} = y - Dx^{(J-1)}$.
- *Sweep*: For $k = 1, 2, \ldots, K$,

$$m(k) = \langle r^{(J)}, d_k \rangle^2 + 2\sigma^2 \ln(p_{J,k}).$$

- *Update Support Set*: Find a maximizer

$$k^* = \text{argmax } m(k).$$

If $m(k^*) > 0$, set $\Gamma^{(J)} = \Gamma^{(J-1)} \cup \{k^*\}$. Else, break out of the iteration.

- *Update weight vector*:

$$x^{(J)} = \text{argmin}_{\text{Support}(x)=\Gamma^{(J)}} \|y - Dx\|_2^2.$$

Fig. 1. The POMP algorithm.

Suppose $Y = \{y_i\}_{i=1}^N$, $y_i \in \mathbb{R}^n$ is the example set, the best dictionary D, the popularity matrix P and the decomposition $X = \{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^K$ are the solution of

$$\begin{aligned} \text{argmin}_{D,X,P} & \|Y - DX\|_F^2 - 2\sigma^2 \sum_{j=1}^{N} \sum_{i=1}^{\|x_j\|_0} \ln(p_{i,sel_i^{(j)}}), \\ & \text{subject to } \forall j, \|x_j\|_0 \leq T_0 \end{aligned} \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm (i.e., we have $\|\cdot\|_F = \sqrt{\sum_{ij} \cdot^2}$). The goal of (7) is not only to minimize the reconstruction error but also to maximize the atoms' popularity.

3. PK-SVD ALGORITHM

In the PK-SVD algorithm, as shown in Fig. 2, we solve (7) iteratively, using three stages. The first two stages are parallel to those in the K-SVD [1]. In the sparse coding stage, we solve (6) using a greedy algorithm extended from the Orthogonal Matching Pursuit (OMP) algorithm [2]. We name it Probabilistic OMP (POMP). As shown in Fig. 1, in each turn, the atom which can best reduce the residual and is most popular by itself is added into the atom group. The dictionary update stage is identical with K-SVD, as their approach has

Task: Find the best dictionary D with its matrix P and decomposition X:

$$\text{argmin}_{D,X,P}\|Y-DX\|_F^2-2\sigma^2\sum_{j=1}^{N}\sum_{i=1}^{\|x_j\|_0}\ln(p_{i,sel_i^{(j)}}),$$

subject to $\forall j,\ \|x_j\|_0\le T_0$.

Initialization: Set the random normalized dictionary matrix $D^{(0)}\in\mathbb{R}^{n\times K}$. Set the popularity matrix $P^{(0)}=(\frac{1}{K})_{T_0\times K}$. Set $\sigma_{(0)}^2=\max(\frac{\|Y\|_F^2}{6nN},\sigma^2)$. Set $J=1$.
For $J=1,2,\ldots,$ IterNum
Sparse Coding Stage:
Compute the initial $X^{(J)}=\{x_j^{(J)}\}_{j=1}^N$ using POMP algorithm with $D^{(J-1)}$, $P^{(J-1)}$ and $\sigma_{(J-1)}$ as parameter.
Dictionary Update Stage:
Update the initial $X^{(J)}$ to the updated $X^{(J)}$ and $D^{(J-1)}$ to $D^{(J)}$ using the same step as in K-SVD[1].
Popularity Matrix Update Stage:

- Define $N_{ij}=\sharp\{k|sel_i^{(k)}=j,k=1,2,\ldots,N\}$.
- For all the element of $P^{(J)}$, set $p_{i,j}^{(J)}=\frac{N_{ij}}{\sum_{k=1}^K N_{ik}}$.

Set $\sigma_{(J)}^2=\sigma_{(J-1)}^2-\frac{\sigma_{(0)}^2-\sigma^2}{\text{IterNum}-1}$.
Set $J=J+1$.

Fig. 2. The PK-SVD algorithm.

been verified to be powerful and efficient. Finally we update the prior probabilities. If one fixes dictionary D and decomposition X, matrix P becomes the solution to the following:

$$\begin{aligned}&\max_P\sum_{j=1}^{N}\sum_{i=1}^{\|x_j\|_0}\ln(p_{i,sel_i^{(j)}})=\max_P\sum_{i=1}^{T_0}\sum_{j=1}^{K}N_{ij}\ln(p_{i,j}),\\&\text{subject to }\forall i=1,2,\ldots,T_0\quad\sum_{j=1}^{K}p_{i,j}=1,\end{aligned}\tag{8}$$

where N_{ij} is the number of elements in set

$$\{k|sel_i^{(k)}=j,k=1,2,\ldots,N\}.$$

Using the Lagrange multiplier method, the solution is

$$p_{i,j}=\frac{N_{ij}}{\sum_{k=1}^K N_{ik}}.\tag{9}$$

During the iteration, the noise level used by the POMP algorithm is generally reduced to the real value in the case we don't know the exact sparsity. If the sparsity T_0 we set is too big and the real noise level is small, the randomly initialized dictionary shall not approximate Y sparsely. This means the sparse coding result $X^{(J)}$ in each iteration shall be too far from the real X, and the following update steps become wasteful. But if we use a larger noise level at beginning, the POMP algorithm can stop at the proper sparsity, benefit from the stopping rule $m(k^*)>0$, so that the PK-SVD shall run all right. The beginning noise level we set is $\sigma_{(0)}^2=\max(\frac{\|Y\|_F^2}{6nN},\sigma^2)$ and reduce it by $\frac{\sigma_{(0)}^2-\sigma^2}{\text{IterNum}-1}$ in every iteration.

SNR	10db	20db	30db	No Noise
K-SVD	50.8%	77.6%	78.4%	76.4%
PK-SVD	89.2%	89.2%	89.2%	95.2%

Table 1. The percentage of retrieved atoms (special popularity matrix).

In the implementation, there are two minor issues: 1) K-SVD has a clear dictionary step, which replaces a seldom used atom with a new one. Instead of choosing the observed signal y_i, whose corresponding residual norm $\|r_i\|_2$ is the biggest, we calculate the angle θ_{ij} between any two residuals r_i and r_j, find $i^*=\text{argmax}_i\sharp\{j|\theta_{ij}<\theta_{threshold}\}$, and select r_{i^*} as the new atom. 2) We calculate the value of (7) in each iteration and record $(D^{(J)},P^{(J)},X^{(J)})$ corresponding to the smallest value as the final result.

An important question that arises is: will the algorithm converge? Obviously, the algorithms used in the dictionary update stage and the popularity matrix update stage are powerful enough to ensure a monotonic reduction. As for the POMP, it's a direct extension of OMP and OMP is proved to succeed when T_0 is small enough. So when T_0 is really small, we can expect POMP to be reliable and PK-SVD is guaranteed to convergent to a local minimum. Further research can focus on how to jump from the current local minimum to a better one.

4. EXPERIMENT RESULTS

4.1. Synthetic experiments

In order to study PK-SVD, we conduct the synthetic tests in the same way as in [1]. We compare PK-SVD with K-SVD.

We randomly choose a normalized dictionary $D\in\mathbb{R}^{20\times 50}$, set the sparsity $T_0=3$ and assign the hidden popularity matrix P to be $\{p_{1,j}\}_{j=1}^{10}=\frac{1}{10}$, $\{p_{2,j}\}_{j=11}^{30}=\frac{1}{20}$ and $\{p_{3,j}\}_{j=31}^{50}=\frac{1}{20}$, and generate $Y=\{y_i\}_{i=1}^{4000}$. We then add white noises with various strengths to those signals. For each noise level, we execute 5 experiments. Both the K-SVD and PK-SVD was executed for a maximum number of 80 iterations to learn a new dictionary $\hat{D}$, which is compared with D, using the same method as in [1]. As shown in Table 1, the difference of PK-SVD and K-SVD is obvious, because PK-SVD tend to adapt to the special popularity matrix P.

σ	Lena		Barbara		Boats		House	
5	**38.61**	38.48	**38.08**	37.98	**37.24**	36.88	39.39	**39.48**
10	**35.50**	35.49	34.41	**34.50**	33.64	**33.67**	35.91	**36.18**
15	33.70	**33.84**	32.40	**32.52**	31.70	**31.89**	34.29	**34.42**
25	31.37	**31.58**	29.60	**29.88**	29.30	**29.64**	32.09	**32.10**
50	27.80	**27.94**	25.55	**26.07**	26.00	**26.38**	**28.01**	27.97

Table 2. Summary of the denoising PSNR results in decibels. In each cell, two denoising results are reported and the higher one is bolded. Left: Results of Elad etc. [3]. Right: Results of the method proposed in this paper ($\lambda = 2.5/\sigma^2$).

4.2. Image denoising experiments

Image denoising is an important research problem, which plays a fundamental role in many applications. Elad and coauthors studied the problem using K-SVD [3]. In this paper, we apply PK-SVD to image denoising. For a certain image, we first randomly extract thousands of image patches for dictionary learning. Assume Y is the observed image, X is the underlying "true" noise-free image, $\hat{D}$ is the learnt dictionary, $R_{ij}X$ is the image patch at position (i, j) and $\hat{\alpha}_{ij}$ is the corresponding linear representation using POMP. We need to solve

$$\hat{X} = \text{argmin}_{\,x} \lambda\|X-Y\|_F^2 + \sum_{ij} w_{ij}\|R_{ij}X - \hat{D}\hat{\alpha}_{ij}\|_2^2, \quad (10)$$

where w_{ij} is defined as

$$w_{ij} = \frac{\sigma^2}{\|R_{ij}Y - \hat{D}\hat{\alpha}_{ij}\|_2^2 - 2\sigma^2 \sum_{k=1}^{\|\hat{\alpha}_{ij}\|_0} \ln(p_{k,sel_k})}.$$

The weight represents the confidence of $R_{ij}X$ to be $\hat{D}\hat{\alpha}_{ij}$. Better decomposition yields larger w_{ij}.

This is a simple quadratic term that has a closed-form solution of the form

$$\hat{X} = \left(\lambda I + \sum_{ij} w_{ij} R_{ij}^T R_{ij}\right)^{-1} \left(\lambda Y + \sum_{ij} w_{ij} R_{ij}^T \hat{D}\hat{\alpha}_{ij}\right).$$

Table 2 summarizes the denoising results achieved by applying K-SVD and PK-SVD on several test images. In this set of experiments, the dictionary size is 64×256, designed to handle image patches of size 8×8. Each result is reported from one experiment. We note that the initial dictionary of K-SVD is a redundant DCT dictionary, while our proposed method use DC atom and AC part of randomly chosen image patches. This initialization may need more iteration for dictionary learning, however can better fit the corrupted image. We apply 10 iteration for both K-SVD and PK-SVD.

As can be seen from Table 2, our method outperforms the K-SVD in most cases. The average PSNR improvement for all the noise levels is over 0.1dB. The performance difference becomes larger as σ increases.

Learned Adaptive Dictionary

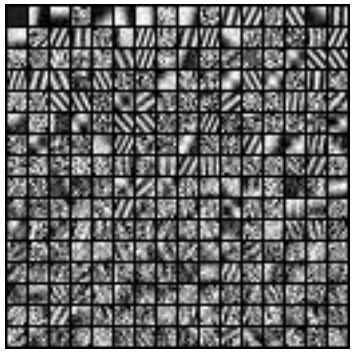

Fig. 3. Example of the denoising results for the image "Barbara" with $\sigma = 50$—the learned dictionary.

In the image "Barbara", which contains high frequency texture features, PK-SVD can better learn the specific characteristics of the image. The learnt dictionary is shown in Fig 3 for $\sigma = 50$. In the figure, we sort the atoms according to its popularity, the upper ones are used more often.

5. SUMMARY

We solve the problem of dictionary learning with a new popularity-based model. Our proposed model is more flexible and can better fit the given signal examples than another existing sparse coding method. Experiments on synthetic data and natural images show the strength of PK-SVD, in both dictionary learning and image denoising.

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (60702044, 60932006, 60625103) and 973 Program (2010CB731401).

7. REFERENCES

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

[2] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. on Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[3] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. on Image Processing*, vol. 15, no. 12, pp. 3736–3745, December 2006.