A Dictionary Learning Method for Sparse Representation Using a Homotopy Approach

Milad Niknejad[1(✉)], Mostafa Sadeghi[2], Massoud Babaie-Zadeh[2], Hossein Rabbani[3], and Christian Jutten[4]

[1] Islamic Azad University, Majlesi Branch, Isfahan, Iran
milad3n@gmail.com
[2] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
[3] Biomedical Engineering Department, Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
[4] GIPSA-lab, Institut Universitaire de France, Grenoble, France

Abstract. In this paper, we address the problem of dictionary learning for sparse representation. Considering the regularized form of the dictionary learning problem, we propose a method based on a homotopy approach, in which the regularization parameter is overall decreased along iterations. We estimate the value of the regularization parameter adaptively at each iteration based on the current value of the dictionary and the sparse coefficients, such that it preserves both sparse coefficients and dictionary optimality conditions. This value is, then, gradually decreased for the next iteration to follow a homotopy method. The results show that our method has faster implementation compared to recent dictionary learning methods, while overall it outperforms the other methods in recovering the dictionaries.

Keywords: Dictionary learning · Sparse representation · Homotopy · Adaptive · Warm-start method

1 Introduction

In recent years, it has been shown that sparse representation leads to promising results in many applications of signal processing [3]. Sparse representation deals with approximating a signal as a linear combination of a few known signals, called atoms, chosen from a signal collection, called dictionary. The performance of sparse coding for a particular class of signals is highly related to a dictionary having the ability to represent all signals in the class by linear combinations of a few atoms. Learning sparsifying dictionaries has also been shown to outperform known and predetermined dictionaries in some applications for classes of signals such as images [4] and audio [6].

A common approach to obtain the dictionary is to use alternating minimization in an iterative procedure [5,13]. In the sparse coding stage, sparse coefficients

This work was partially funded by European project 2012-ERC-AdG-320684 CHESS.

E. Vincent et al. (Eds.): LVA/ICA 2015, LNCS 9237, pp. 271–278, 2015.
DOI: 10.1007/978-3-319-22482-4_31

are obtained while the previously found dictionary is fixed, and in the dictionary update stage, the dictionary is found based on the obtained coefficients. In the sparse coding stage, Orthogonal Matching pursuit (OMP) [10] and Iterative Shrinkage Thresholding (IST) algorithm [12] have been used in Method of Optimal Directions (MOD) [5] and Majorization Method (MM) [13] dictionary learning, respectively. Among some examples of dictionary update stages, the MOD used the observation matrix multiplied by pseudo inverse of representation matrix, and a Maximum A Posteriori (MAP)-based dictionary learning in [7] used a gradient descent method, both followed by normalization of dictionary columns. However, all methods proposed so far have not considered the adaptivity of uncertain parameters of the cost function, such as the regularization parameter, to the data.

In this paper, we propose a dictionary learning method for sparse representations, which benefits from a homotopy (continuation) method. Generally, the homotopy is a heuristic that, first, computes the solution of an initial simpler problem, in which the global minimum can be easily found, and then, gradually deforms the initial problem to the desired one. The homotopy has also been used in solving nonlinear equations [8], and in the optimization relating to sparse representation with fixed dictionary [2,12]. Inspired by a homotopy approach, we propose a method which starts solving the dictionary learning cost function from a higher value of the regularization parameter, and adaptively decreases this parameter along iterations. Although our method uses an alternating minimization approach, as we explain through this paper, being capable of changing the value of the regularization parameter enables us to choose a regularization parameter such that it keeps the sparse representation solutions near the optimal after updating the dictionary. Our method can also be seen as a method that uses a homotopy approach with an adaptive regularization parameter selection.

In the following sections, the dictionary learning problem is first discussed in Sect. 2. Then, Sect. 3 is devoted to the description of our proposed method. In Sect. 4, we evaluate the performance and speed of our method in comparison to other dictionary learning algorithms.

2 The Dictionary Learning Problem

Let $\{\mathbf{y}_l \in \mathbb{R}^p\}_{l=1}^L$ be the set of training signals, and $\{\mathbf{x}_l \in \mathbb{R}^q\}_{l=1}^L$ be the set of corresponding representation coefficients over the dictionary $\mathbf{D} \in \mathbb{R}^{p\times q}$. Forming a training data matrix by $\mathbf{Y} \triangleq [\mathbf{y}_1 \dots \mathbf{y}_L]$, and the representation matrix by $\mathbf{X} \triangleq [\mathbf{x}_1 \dots \mathbf{x}_L]$, the dictionary learning problem for sparse representations, as used in [13], can be mathematically modeled by the joint optimization problem of the form

$$\underset{\mathbf{D}\in\mathscr{D},\mathbf{X}}{\operatorname{argmin}}\{\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\lambda_d\|\mathbf{X}\|_{1,1}\} \tag{1}$$

where $\|.\|_F$ indicates the Frobenius norm, and $\|\mathbf{X}\|_{r,s} \triangleq \sum_i (\sum_j |x_{i,j}|^r)^{s/r}$. Although other matrix norms (generally $0 < r < 1$ and $0 < s < 1$), promotes sparsity in the representations in (1). $\|\mathbf{X}\|_{1,1}$ is used for this purpose due

to its convexity and also its separability into the absolute sum of the individual entries of the matrix i.e. $\|\mathbf{X}\|_{1,1} = \sum_i \sum_j |x_{i,j}|$. In (1), λ_d is the desired value for the regularization parameter, and is set to achieve a suitable tradeoff between the accuracy of the representations and the sparsity level in $\mathbf{X}$. The desired value of the regularization parameter depends on the application in which the dictionary learning is employed. As an example in [4], this value is set proportional to the variance of Gaussian noise for an image denoising application. Since solving the optimization problem in (1) tends to increase the norms of the atoms, which unfavorably affects some sparse representation algorithms, it is desirable to constrain the norms of the dictionary atoms by defining the admissible set of

$$\mathscr{D} = \{\mathbf{D} \in \mathbb{R}^{p \times q} \quad \text{s.t} \quad \forall j \ ||\mathbf{d}_j||_2 \leq 1\}. \tag{2}$$

3 Our Proposed Method

Using a homotopy approach, we start to solve the optimization problem in (1) with a high value of the regularization parameter, and then decrease it along the iterations adaptively until reaching the desired value of λ_d. The starting value for the regularization parameter and the procedure of choosing its values along iterations is discussed in Sect. 3.3. So our proposed method at the n^{th} iteration, instead of a fixed value of λ_d, solves the optimization problem of the form

$$\underset{\mathbf{D} \in \mathscr{D}, \mathbf{X}}{\operatorname{argmin}} \{\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda^{(n)} \|\mathbf{X}\|_{1,1}\} \tag{3}$$

where as n grows, $\lambda^{(n)}$ decreases adaptively.

In order to solve the minimization problem in (3), our algorithm alternates among the sparse coding stage, the dictionary update stage and the update of $\lambda^{(n)}$. Our method also uses $\mathbf{X}^{(n)}$ and $\mathbf{D}^{(n)}$ found in the optimization problem with $\lambda^{(n)}$ as a warm-start for solving the optimization problem with the nearby value of $\lambda^{(n+1)}$. Using a warm-start strategy has been previously shown to be effective in improving the speed of dictionary learning algorithms [11,13].

3.1 Sparse Coding Stage

In our method, a sparse coding algorithm which belongs to the class of IST methods is used. These methods benefit from a proper initialization enabling us to use the warm-start strategy. At the k^{th} iteration of the sparse coding algorithm in the n^{th} dictionary learning iteration, the sparse coding solves

$$\mathbf{X}^{(k+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \{\|\mathbf{X} - \mathbf{U}^{(k)}\|_F^2 + \frac{\lambda^{(n)}}{c} \|\mathbf{X}\|_{1,1}\} \tag{4}$$

in which c should satisfy $c > \|\mathbf{D}^T\mathbf{D}\|$ where $\|.\|$ stands for the spectral norm, and $\mathbf{U}^{(k)} \triangleq \mathbf{X}^{(k)} + \frac{1}{c}(\mathbf{D}^T(\mathbf{Y} - \mathbf{D}\mathbf{X}^{(k)}))$ [12]. The global optimum of the convex and non-smooth problem in (4), is the point with zero subgradient i.e.

$$2c(\mathbf{X} - \mathbf{U}^{(k)}) - \lambda^{(n)} \mathscr{P} \in 0 \tag{5}$$

where $\mathscr{P} = \nabla\|\mathbf{X}\|_{1,1}$ is a set of matrices whose entries satisfy

$$\begin{cases} p_{i,j} = 1 & \text{if } \mathbf{x}_{i,j} > 0 \\ p_{i,j} \in [-1\ 1] & \text{if } \mathbf{x}_{i,j} = 0 \\ p_{i,j} = -1 & \text{if } \mathbf{x}_{i,j} < 0. \end{cases}$$

The point that satisfies the optimality condition (5), is obtained by $\mathbf{X}^{(k)} = \mathcal{S}_{\frac{\lambda^{(n)}}{2c}}(u_{i,j})$ which is a soft thresholding operator on entries of $\mathbf{U}$ with the threshold value of $\frac{\lambda^{(n)}}{2c}$. Using the soft thresholding operator there is one single matrix $\mathbf{P}^{(k)}$ from the set $\mathscr{P}$ which makes the condition in (5) turn into

$$2c(\mathbf{X}^{(k)} - \mathbf{U}^{(k)}) - \lambda^{(n)}\mathbf{P}^{(k)} = 0. \tag{6}$$

3.2 Dictionary Update Stage

The dictionary update stage is to find the minimization problem in (1), while $\mathbf{X}$ is fixed with the value found in the previous sparse coding stage. Similar to [7], we use the gradient descent algorithm. So in the k^{th} iteration of the gradient descent of the n^{th} iteration of dictionary learning algorithm, our method updates the dictionary by

$$\mathbf{D}^{(k+1)} = \mathbf{D}^{(k)} + \rho(\mathbf{Y} - \mathbf{D}^{(k)}\mathbf{X}^{(n)})\mathbf{X}^{(n)^T} \tag{7}$$

where ρ is an appropriate constant and is set to .001 in our implementations. Using the gradient descent has a fast implementation, and due to using a proper initialization, enables us to employ a warm-start strategy, similar to sparse coding algorithm in our method. Then, our algorithm normalizes the atoms whose norms are more than one to the unit norm and keep the other atoms intact.

3.3 Determining the Regularization Parameter

In many homotopy methods, decreasing the regularization parameter is done heuristically by a linear or exponential decay [9]. However in this section, we propose a more sophisticated choice for this value.

One of the disadvantages of the alternating minimization between the two stages of dictionary learning algorithms is that each stage may not preserve the optimality of the other one. So the solutions in an alternating minimization approach might oscillate around an optimal point. To understand this, assume without loss of generality that dictionary update is performed after sparse coding stage at each iteration. Updating the dictionary may not preserve the optimality condition derived for the sparse coding in (5) or equivalently in (6), since this condition is not considered in the dictionary update stage. So updating $\mathbf{D}$ might lead to a deviation from the optimality condition of sparse coding at end of each iteration of dictionary learning algorithm. Being capable of changing the regularization parameter in our method, in order to alleviate this, after the

dictionary update stage we choose the regularization parameter in such a way that it best preserves the optimality condition in (6) for sparse coding. The criterion for optimality of sparse coding stage could be the Frobenius norm of the term that is set to zero in (6). After the alteration of $\mathbf{D}$, the value of $\mathbf{U}$ changes, and the mentioned term might not be equal to zero. So we find the value of the regularization parameter which minimizes the Frobenius norm of the term set to zero in (6), after updating the dictionary, based on the current estimate of $\mathbf{D}$ and $\mathbf{X}$ i.e.

$$\begin{aligned}\lambda_{opt} &= \underset{\lambda}{\operatorname{argmin}} \|\mathbf{R}^{(n)} - \lambda \mathbf{P}^{(n)}\|_F^2 \\ &= \underset{\lambda}{\operatorname{argmin}} \{Tr({\mathbf{R}^{(n)}}^T \mathbf{R}^{(n)}) + \lambda^2 Tr({\mathbf{P}^{(n)}}^T \mathbf{P}^{(n)}) - 2\lambda Tr({\mathbf{P}^{(n)}}^T \mathbf{R}^{(n)})\} \quad (8)\end{aligned}$$

where $\mathbf{R}^{(n)} = 2c(\mathbf{X}^{(n)} - \mathbf{U}^{(n)})$. The global optimum of the above least square minimization problem can be found by setting its derivative to zero which leads

$$\lambda_{opt} = \frac{Tr({\mathbf{P}^{(n)}}^T \mathbf{R}^{(n)})}{Tr({\mathbf{P}^{(n)}}^T \mathbf{P}^{(n)})}. \quad (9)$$

Having found the optimal value for $\lambda^{(n)}$ based on the current estimation of $\mathbf{X}$ and $\mathbf{D}$, in order to follow a homotopy, we gradually decrease this value by a constant factor which leads to $\lambda^{(n+1)} = (1-\epsilon)\lambda_{opt}$ where ϵ is a small constant. However, implementing some iterations of our algorithm without applying $\lambda^{(n+1)} = (1-\epsilon)\lambda_{opt}$ is desirable, since it leads to equilibrium for a joint point of $(\mathbf{X}, \mathbf{D}, \lambda)$ (Note that the value of λ here is higher than the desired value). The value of regularization parameter is also forced to be bounded to the desired value of λ_d which makes final iterations be implemented with this value of regularization parameter. It is worth mentioning that the procedure of finding the optimal value for regularization parameter and decreasing it by a constant factor has been used in a homotopy based sparse coding (with fixed dictionary) in [12]. However the procedure of obtaining the optimal value is completely different and novel in our method, and is adapted to the dictionary learning application.

The initial optimal value of the regularization parameter is set to $\|D^T Y\|_\infty$, where $\|.\|_\infty$ returns the maximum absolute value of the matrix entries, since for $\lambda^{(1)} > \|D^T Y\|_\infty$, the solution of zero is optimal in (5) [12], and consequently, no update of initial dictionary is occurred in the dictionary update stage (Fig. 1).

It is worth mentioning that the performance of homotopy methods depends on the tracing the optimal solutions while the value of the regularization parameter changes. As we discussed in this subsection, by the proposed optimal choice for the regularization parameter, our algorithm tends to keep the optimal solutions along iterations for both dictionary and sparse coefficients.

4 Simulations Results

In this section, we compare our method with other methods using synthetic signals to evaluate the performance of algorithms in recovering the dictionary that produces the data.

– Initialization: Choose an initial dictionary $\mathbf{D} \in \mathbf{R}^{p \times q}$
– For $n = 1, \ldots, N$ (main loop)
Sparse coding stage:
1. Initialize with $\mathbf{D} = \mathbf{D}^{(n-1)}$, $\mathbf{X}^{(k=0)} = \mathbf{X}^{(n-1)}$
2. For $k = 1, \ldots, K_s$
$$\mathbf{U}^{(k)} = \mathbf{X}^{(k-1)} + \frac{1}{c}(\mathbf{D}^T(\mathbf{Y} - \mathbf{D}\mathbf{X}^{(k-1)})), \mathbf{X}^{(k)} = \mathcal{S}_{\frac{\lambda^{(n)}}{2c}}(\mathbf{U}^{(k)})$$
End For
3. Set $\mathbf{X}^{(n)} = \mathbf{X}^{(K_s)}$
Dictionary update stage:
1. Initialize with $\mathbf{X} = \mathbf{X}^{(n)}$, $\mathbf{D}^{(k=0)} = \mathbf{D}^{(n-1)}$
2. For $k = 1, \ldots, K_d$
$$\mathbf{D}^{(k)} = \mathbf{D}^{(k-1)} + \rho(\mathbf{Y} - \mathbf{D}^{(k-1)}\mathbf{X})\mathbf{X}^T$$
Normalize columns of dictionary whose norms are more than 1.
3. End For
4. Set $\mathbf{D}^{(n)} = \mathbf{D}^{(K_d)}$
Regularization parameter selection:
1. Obtain the optimum regularization parameter λ_{opt} by (9)
2. decrease the regularization parameter by $\lambda^{(n+1)} = \max((1-\epsilon)\lambda_{opt}, \lambda_d)$
– End For (main loop)
– Final answer is $\mathbf{D} = \mathbf{D}^{(N)}$

Fig. 1. Our proposed dictionary learning algorithm

A dictionary of size 30×60 is randomly generated with independent identically distributed (i.i.d.) Gaussian entries, and its columns are normalized to have unit norms. 4000 sample signals $\{\boldsymbol{y}_l\}_{l=1}^{4000}$ are produced by linear combination of a few (precisely determined by Q in each experiment) number of atoms with the coefficients which are i.i.d. Gaussian in random and independent locations. We compare our method with MOD [5] and K-SVD [1] as two well-known methods, and also with the Majorization dictionary learning algorithm [13] which has improved those methods and its sparse coding algorithm is similar to our method. For other methods, the MATLAB codes published online by the authors were used. All the experiments were done with core i5 CPU with 4 GB of memory using Matlab 2011a under Microsoft Windows 7 operating system.

The percentages of recovered atoms are compared for different methods during the execution time with data generated by $Q = 4$ number of atoms in Fig. 2. The CPU time is considered in this experiment to roughly compare the computational complexity of the algorithms. The value of ϵ for homotopy decreasing factor is set to 0.05 and applied every 4 iterations (to obtain an equilibrium point for a higher value of the regularization parameter, as described in the previous section). We found that this implementation leads to an appropriate tradeoff between the speed and preserving the performance in our algorithm. The desired value of λ_d in our algorithm and the value of the regularization parameter in MM method are both set to 0.18. Figure 2 shows that our method converges faster

and more accurate in this case. In order to better compare the speed of the algorithms, the CPU times are reported for different algorithms while the sparsity level Q varies from 3 to 6. Also, the percentages of the recovered atoms for this experiment are shown in Fig. 3(a) to compare the performance of algorithms in recovering dictionary atoms. The corresponding implementation times are shown in Fig. 3(b). The values are averaged over three independent implementations of algorithms. Based on this figure, our algorithm is more successful in recovering the dictionary except for $Q = 3$, and is faster in all the cases, compared to other methods.

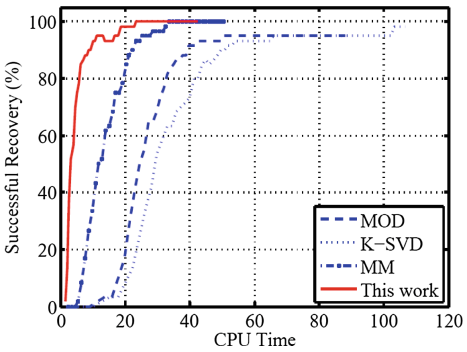
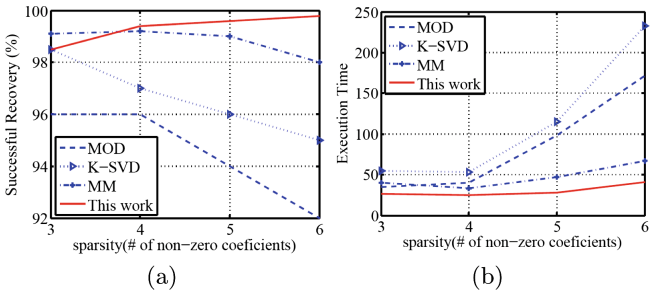


Fig. 2. Comparison of the percentage of recovered atoms vs. the computational time for different methods in $Q = 4$.



(a) (b)

Fig. 3. Comparison of performance of dictionary learning algorithms for datasets with different values of sparsity level Q: (a) Percentage of recovered atoms, (b) implementation time.

5 Conclusion

In this paper, we proposed a homotopy-based method for dictionary learning for sparse representation in which the value of the regularization parameter

decreases along iterations. We proposed an adaptive selection for the regularization parameter which best preserves the optimality of sparse coefficients after the dictionary update at each iteration. The results showed that our method is more successful in recovering the dictionaries compared to other methods, and it has faster implementation time.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Trans. Sig. Process. **54**, 4311–4322 (2006)
2. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
3. Elad, M.: Sparse and redundant representations: from theory to applications in signal and image processing. Springer, New York (2010)
4. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. Image Process. **15**, 3736–3745 (2006)
5. Engan, K., Aase, S.O., Hakon-Husoy, J.H.: Method of optimal directions for frame design. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, pp. 2443–2446 (1999)
6. Jafari, M., Plumbley, M.: Fast dictionary learning for sparse representations of speech signals. IEEE Sel. Top. Sign. Process. **5**, 1025–1031 (2011)
7. Kreutz-Delgado, K., Murray, J.F., Rao, B.D., Engan, K., Lee, T., Sejnowski, T.: Dictionary learning algorithms for sparse representation. Neural Comput. **15**(2), 349–396 (2003)
8. Liao, S.: Homotopy analysis method in nonlinear differential equations. Springer, Heidelberg (2012)
9. Mancera, L., Portilla, J.: Non-convex sparse optimization through deterministic annealing and applications. In: 15th IEEE International Conference on Image Processing, pp. 917–920 (2008)
10. Pati, Y., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: 27th Annual Asilomar Conference Signals, Systems and Computers vol. 1, pp. 40–44 (1993)
11. Smith, L.N., Elad, M.: Improving dictionary learning: multiple dictionary updates and coefficient reuse. IEEE Signal Process. Lett. **20**(1), 79–82 (2013)
12. Wright, S.J., Nowak, R.D., Figueiredo, M.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**(7), 2479–2493 (2009)
13. Yaghoobi, M., Blumensath, T., Davies, M.E.: Dictionary learning for sparse approximations with the majorization method. IEEE Trans. Signal Process. **57**(6), 2178–2191 (2009)