INK-SVD: Learning incoherent dictionaries for sparse representations

Mailhe, B; Barchiesi, D; Plumbley, MD; IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)

INK-SVD: LEARNING INCOHERENT DICTIONARIES FOR SPARSE REPRESENTATIONS

Boris Mailhé, Daniele Barchiesi and Mark D. Plumbley

Queen Mary University of London
School of Electronic Engineering and Computer Science
Centre For Digital Music
E14NS London, United Kingdom
(e-mail: firstname.name@eecs.qmul.ac.uk)

ABSTRACT

This work considers the problem of learning an incoherent dictionary that is both adapted to a set of training data and incoherent so that existing sparse approximation algorithms can recover the sparsest representation. A new decorrelation method is presented that computes a fixed coherence dictionary close to a given dictionary. That step iterates pairwise decorrelations of atoms in the dictionary. Dictionary learning is then performed by adding this decorrelation method as an intermediate step in the K-SVD learning algorithm. The proposed algorithm INK-SVD is tested on musical data and compared to another existing decorrelation method. INK-SVD can compute a dictionary that approximates the training data as well as K-SVD while decreasing the coherence from 0.6 to 0.2.

Index Terms— Sparse coding, Dictionary learning, Coherence, K-SVD

1. INTRODUCTION

In the method of sparse representations, a signal is expressed as a linear combination of a few vectors named *atoms* taken from a set called a *dictionary*. A good dictionary must obey several criteria. First it has to be adapted to the data being represented. Good pre-constructed dictionaries are known for common classes of signals, but sometimes it is not enough and the dictionary has to be learned from examples of the data to represent [1]. Second, even when the dictionary is known, finding the sparsest representation of the data is in general an NP-Hard problem. However several polynomial-time algorithms have been proven to be optimal if the dictionary is sufficiently close to orthogonal [2]. Coherence is one measure of this proximity. The coherence $\mu(\mathbf{\Phi})$ of a dictionary $\mathbf{\Phi}$ is the maximal correlation of any two different atoms:

$$\mu(\mathbf{\Phi}) = \max_{\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j \in \mathbf{\Phi}, i \neq j} \left| \left\langle \frac{\boldsymbol{\varphi}_i}{\|\boldsymbol{\varphi}_i\|_2}, \frac{\boldsymbol{\varphi}_j}{\|\boldsymbol{\varphi}_j\|_2} \right\rangle \right| \tag{1}$$

The μ function is valued between 0 and 1. The minimum is reached for an orthogonal dictionary and the maximum for a dictionary containing at least two collinear atoms. This work aims at merging those two criteria: learn an incoherent dictionary that is adapted to the training data.

There have already been a few early attempts at this task, but they either restrict the dictionary to a parametric form [3] or use a relaxed constraint penalization method that makes it harder to tune the exact value of the coherence [4] while the results on exact recovery provide hard bounds [2]. Therefore we propose a new algorithm called INcoherent K-SVD (INK-SVD) based on the addition of a decorrelation step to the K-SVD algorithm [5]. This step iteratively selects highly correlated pairs of atoms in the dictionary and decorrelates them until the desired coherence is reached.

Section 2 recalls the grounds of the K-SVD algorithm. Section 3 details the decorrelation method we present. Section 4 evaluates our method on musical audio data.

2. DICTIONARY LEARNING

2.1. Dictionary learning problem

Let $\mathbf{S}$ be a matrix of N training signals $\{\mathbf{s}_n\}_{n=1}^{N} \in \mathbb{R}^D$. Dictionary learning consists in finding a dictionary $\mathbf{\Phi}$ of size $D \times M$ with $M \geq D$ and sparse coefficients $\mathbf{X}$ such that $\mathbf{S} \approx \mathbf{\Phi X}$. For example, if the exact sparsity level K is known, the problem can be formalized as minimizing the error cost function $f(\mathbf{\Phi}, \mathbf{X})$ defined as

$$f(\mathbf{\Phi}, \mathbf{X}) = \|\mathbf{S} - \mathbf{\Phi X}\|_{\text{FRO}}^2 \tag{2}$$

under the constraints

$$\forall m \in [1, M],\ \|\boldsymbol{\varphi}_m\|_2 = 1 \tag{3}$$

$$\forall n \in [1, N],\ \|\mathbf{x}_n\|_0 \leq K \tag{4}$$

This work was supported by the EPSRC Project EP/G007144/1 Machine Listening using Sparse Representations, the EU FET-Open project FP7-ICT-225913-SMALL and the Queen Mary University of London School Studentships.

with $\boldsymbol{\varphi}$ an atom (or column) of $\boldsymbol{\Phi}$ and $\|\mathbf{x}_n\|_0$ the number of non-zero coefficients in the n^{th} column of $\mathbf{X}$.

2.2. K-SVD algorithm

Many dictionary algorithms follow an iterative scheme that alternates between updates of $\mathbf{X}$ and $\boldsymbol{\Phi}$ to minimize the cost function (2). Although this work only presents the combination with K-SVD, our decorrelation method can be combined with any alternate algorithm. K-SVD iterates two steps:

- the sparse approximation step: knowing $\mathbf{S}$ and $\boldsymbol{\Phi}$, we estimate $\mathbf{X}$, using a sparse approximation algorithm such as Orthogonal Matching Pursuit,
- the dictionary update step: jointly re-estimate each atom and its non-zero coefficients to minimize the cost function (2). The reader can refer to the original article for more details [5].

2.3. Incoherent dictionary learning problem

The low coherence of the dictionary can be enforced by adding another constraint to the dictionary learning problem. The problem is still one of minimizing the cost function f described in Equation (2) under the constraints (3) and (4), and another low-coherence constraint is added:

$$\mu\left(\boldsymbol{\Phi}\right) \leq \bar{\mu} \tag{5}$$

with $\bar{\mu}$ a fixed coherence threshold. The proposed INK-SVD algorithm solves that problem by inserting a decorrelation step in the K-SVD loop after the dictionary update.

3. DICTIONARY DECORRELATION

3.1. Previous work

Methods for constructing incoherent over-complete dictionaries include an alternate projection technique initially proposed by Tropp et al. in [6] and modified by Elad in [7], which we will use in Section 4 as a benchmark for INK-SVD.

In the cited work, the goal is to realize a Grassmannian tight frame, that is, an over-complete dictionary with minimal mutual coherence $\mu_{\min} = \sqrt{(M-D)/D(M-1)}$, by iteratively decorrelate an initial dictionary $\boldsymbol{\Phi}$.
The alternate projection consists in iteratively optimizing the Gram matrix $\mathbf{G} = \boldsymbol{\Phi}^{\mathcal{H}}\boldsymbol{\Phi}$, so that its off-diagonal values are shrunk towards $\mu_{\min}$ and that its spectrum has non-negative eigenvalues with rank smaller or equal than the ambient dimension D. This way, we obtain an updated gram $\mathbf{G}_{\text{new}}$, that can be in turn factorized into the product $\mathbf{G}_{\text{new}} = \boldsymbol{\Phi}^{\mathcal{H}}_{\text{new}}\boldsymbol{\Phi}_{\text{new}}$.

This factorization is not unique and does not take into account the sparse approximation objective of dictionary learning. While we are planning to fill this gap in future work, this paper presents an alternative method for dictionary decorrelation that follows a greedy strategy.

3.2. Decorrelation problem

This section proposes a general method that attempts to find the closest dictionary $\hat{\boldsymbol{\Phi}}$ to a given dictionary $\bar{\boldsymbol{\Phi}}$ with a coherence lower than a given $\bar{\mu}$. Formally, $\hat{\boldsymbol{\Phi}}$ is defined as:

$$\hat{\boldsymbol{\Phi}} = \underset{\boldsymbol{\Phi}\in\Gamma}{\operatorname{argmin}} \ \left\|\boldsymbol{\Phi} - \bar{\boldsymbol{\Phi}}\right\|^2_{\text{FRO}} \tag{6}$$
$$\text{with} \quad \Gamma = \left\{\boldsymbol{\Phi} \middle|\ \mu(\boldsymbol{\Phi}) \leq \bar{\mu} \ \wedge \ \|\boldsymbol{\varphi}_m\|_2 = 1, m \in [1, M]\right\}$$

The original dictionary $\bar{\boldsymbol{\Phi}}$ is the unconstrained minimum of the cost function (6). However, either $\mu(\bar{\boldsymbol{\Phi}}) \leq \bar{\mu}$ and it does not need to be decorrelated or $\bar{\boldsymbol{\Phi}}$ is not in the admissible set Γ of Problem (6). In that case $\bar{\boldsymbol{\Phi}}$ is a good candidate for a starting point for the dual problem that minimizes the coherence of the dictionary while staying close to $\bar{\boldsymbol{\Phi}}$:

$$\hat{\boldsymbol{\Phi}} = \underset{\boldsymbol{\Phi}\in\Gamma'}{\operatorname{argmin}} \ \mu(\boldsymbol{\Phi}) \qquad\qquad \text{with} \tag{7}$$

$$\Gamma' = \left\{\boldsymbol{\Phi} \middle|\ \left\|\boldsymbol{\Phi} - \bar{\boldsymbol{\Phi}}\right\|^2_{\text{FRO}} \leq \rho \ \wedge \ \|\boldsymbol{\varphi}_m\|_2 = 1, m \in [1, M]\right\}$$

with ρ the unknown minimum value reached by the criterion (6). We will rather address the dual problem (7).

3.3. Decorrelation of two atoms

Let the initial dictionary $\bar{\boldsymbol{\Phi}}$ be composed of only two unitary atoms $\bar{\boldsymbol{\varphi}}_1$ and $\bar{\boldsymbol{\varphi}}_2$ with a correlation higher than $\bar{\mu}$. In that simple case we can directly express the optimum of Problem (6). Let us assume without loss of generality that $\langle\bar{\boldsymbol{\varphi}}_1, \bar{\boldsymbol{\varphi}}_2\rangle > 0$ (the opposite case can be derived by considering the couple $(\bar{\boldsymbol{\varphi}}_1, -\bar{\boldsymbol{\varphi}}_2)$) and let $\bar{\theta}$ be the half-angle between $\bar{\boldsymbol{\varphi}}_1$ and $\bar{\boldsymbol{\varphi}}_2$. We are looking for the solution $\hat{\boldsymbol{\Phi}} = \begin{pmatrix}\hat{\boldsymbol{\varphi}}_1 & \hat{\boldsymbol{\varphi}}_2\end{pmatrix}$ of Problem (6). The problem only has two degrees of freedom because of the normalization constraint. We choose the half-angle $\hat{\theta}$ between $\hat{\boldsymbol{\varphi}}_1$ and $\hat{\boldsymbol{\varphi}}_2$ and the angle α between the sums $\bar{\boldsymbol{\varphi}}_1 + \bar{\boldsymbol{\varphi}}_2$ and $\hat{\boldsymbol{\varphi}}_1 + \hat{\boldsymbol{\varphi}}_2$ for parameters as shown on Figure 1. In the orthonormal basis

$$\begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{pmatrix} = \begin{pmatrix}\frac{\bar{\boldsymbol{\varphi}}_1 + \bar{\boldsymbol{\varphi}}_2}{\|\bar{\boldsymbol{\varphi}}_1 + \bar{\boldsymbol{\varphi}}_2\|_2} & \frac{\bar{\boldsymbol{\varphi}}_1 - \bar{\boldsymbol{\varphi}}_2}{\|\bar{\boldsymbol{\varphi}}_1 - \bar{\boldsymbol{\varphi}}_2\|_2}\end{pmatrix}$$

all the considered vectors have a simple expression:

$$\bar{\boldsymbol{\Phi}} = \begin{pmatrix}\bar{\boldsymbol{\varphi}}_1 & \bar{\boldsymbol{\varphi}}_2\end{pmatrix} = \begin{pmatrix}\cos\bar{\theta} & \cos\left(-\bar{\theta}\right)\\ \sin\bar{\theta} & \sin\left(-\bar{\theta}\right)\end{pmatrix} \tag{8}$$

$$\hat{\boldsymbol{\Phi}} = \begin{pmatrix}\hat{\boldsymbol{\varphi}}_1 & \hat{\boldsymbol{\varphi}}_2\end{pmatrix} = \begin{pmatrix}\cos(\alpha + \hat{\theta}) & \cos(\alpha - \hat{\theta})\\ \sin(\alpha + \hat{\theta}) & \sin(\alpha - \hat{\theta})\end{pmatrix} \tag{9}$$

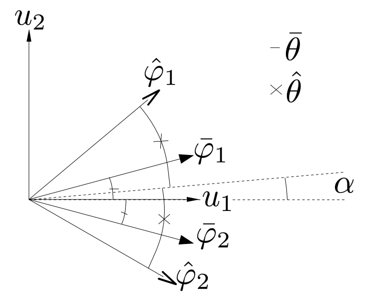

Fig. 1. Decorrelation of two atoms. For the optimal decorrelation we would have $\alpha = 0$ and the pair $(\hat{\boldsymbol{\varphi}}_1, \hat{\boldsymbol{\varphi}}_2)$ would be symmetric with respect to $\mathbf{u}_1$.

We can then express the different constraints:

$$|\langle\hat{\boldsymbol{\varphi}}_1, \hat{\boldsymbol{\varphi}}_2\rangle| = |\cos 2\hat{\theta}| \leq \bar{\mu} \tag{10}$$

$$\|\bar{\boldsymbol{\varphi}}_1 - \hat{\boldsymbol{\varphi}}_1\|_2^2 = 2 - 2\cos(\bar{\theta} - \hat{\theta} - \alpha) \tag{11}$$

$$\|\bar{\boldsymbol{\varphi}}_2 - \hat{\boldsymbol{\varphi}}_2\|_2^2 = 2 - 2\cos(\bar{\theta} - \hat{\theta} + \alpha) \tag{12}$$

$$\left\|\bar{\mathbf{\Phi}} - \hat{\mathbf{\Phi}}\right\|_{\text{FRO}}^2 = 4 - 4\cos(\bar{\theta} - \hat{\theta})\cos(\alpha) \tag{13}$$

If we assume without loss of generality that $\cos(\bar{\theta} - \hat{\theta}) > 0$, then the cost function (13) is minimal for $\alpha = 0$ and $\hat{\theta}$ as close to $\bar{\theta}$ as possible: Problem (6) is solved by rotating $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ symmetrically with respect to their mean until their correlation reaches $\bar{\mu}$. The angle $\hat{\theta}$ is the angle that reaches the equality in Equation (11):

$$\cos 2\hat{\theta} = \bar{\mu} \tag{14}$$

$$\hat{\theta} = \frac{\arccos\bar{\mu}}{2} \tag{15}$$

and the dictionary $\hat{\Phi}$ is given by Equation (9).

3.4. General case

In the general case, the previous method provides the steepest descent direction if only one pair of atoms reaches the maximal correlation. However one can easily prove that the coherence function is non-convex with respect to $\mathbf{\Phi}$ so following a steepest descent does not guarantee to find a global minimum. Instead of a descent method, we chose to decorrelate the dictionary by iterating decorrelations of pairs of atoms. The core idea is simple: as long as there are any atoms with correlation higher than $\bar{\mu}$, select a pair of them and decorrelate them with the method explained in Section 3.3.

However, decorrelating two atoms can potentially change correlations with other atoms in the dictionary, so finding the next pair would require to update the correlations after each pair decorrelation. We speed up the process by decorrelating some pairs in parallel. Instead of selecting one pair of atoms at a time, we partition the whole dictionary into high correlation pairs (and single atoms that do not need to be modified), decorrelate all those pairs and only then update the correlations. This is detailed on Algorithm 1.

Algorithm 1 $\mathbf{\Phi}$ = decorrelate($\bar{\mathbf{\Phi}}$,$\bar{\mu}$)

```
$\mathbf{\Phi} \leftarrow \bar{\mathbf{\Phi}}$
while $\mu(\mathbf{\Phi}) > \bar{\mu}$ do
    $E$ = partition($\mathbf{\Phi}, \bar{\mu}$)
    for $\forall(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j) \in E$ do
        decorrelate_pair($\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j$)
    end for
end while
```

The partitioning is performed in a greedy way. We start with the whole dictionary, group the pair with the highest correlation together and remove it from the set of considered atoms until there are no pairs left with correlation higher than $\bar{\mu}$. It is detailed in Algorithm 2.

Algorithm 2 E =partition($\bar{\mathbf{\Phi}}, \bar{\mu}$)

```
$\mathbf{\Phi} \leftarrow \bar{\mathbf{\Phi}}$
$E \leftarrow \emptyset$
while $\mu(\mathbf{\Phi}) > \bar{\mu}$ do
    $(i, j) = \text{argmax}\ \left|\mathbf{\Phi}^{\mathcal{H}}\mathbf{\Phi} - I\right|$
    $\mathbf{\Phi} \leftarrow \mathbf{\Phi} \setminus \{\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j\}$
    $E \leftarrow E \cup \{(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j)\}$
end while
```

4. NUMERICAL EXPERIMENTS

We tested the incoherent dictionary learning algorithm in order to assess if it can provide us with a dictionary for sparse representation that exhibits minimal coherence and good approximation quality. The test signal we used is a 16 kHz guitar recording that is part of the test data included in SMALLbox [8] [1], a Matlab toolbox for testing and benchmarking dictionary learning algorithms that we used in our evaluation. A musical audio signal was chosen because previous informal experiments resulted in K-SVD learning a highly coherent dictionary for this type of data.

We divided the recording into 50% overlapping blocks of 256 samples (corresponding to 16ms) with rectangular windows and arranged the resulting vectors as columns of the training data matrix $\mathbf{S}$. Then, we initialized three twice overcomplete dictionaries for sparse representation using respectively 1) randomly chosen subset of the training data $\mathbf{S}$, 2) over-complete DCT and 3) over-complete Gabor frames. We run the K-SVD dictionary learning algorithm for 20 iterations, allowing for 12 non-zero coefficients in each representation (which corresponds to about 5% of active elements if compared with the ambient dimension D).

[1] http://small-project.eu/software-data/smallbox

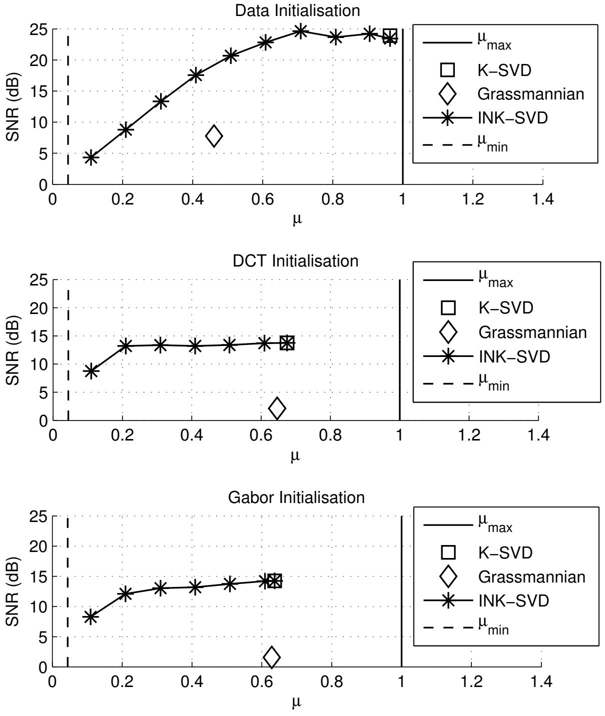

Fig. 2. Signal to noise ratio as a function of the coherence value for different choices of dictionary initialization and decorrelation functions. The levels $\mu_{max} = 1$ and $\mu_{min} = \sqrt{(M-D)/D(M-1)}$ indicate the maximum and minimum coherence attainable by a $D \times M$ dictionary.

We included the proposed INK-SVD decorrelation algorithm and compared it with the Grassmannian method detailed in section 3.1, using the implementation presented in [7, p.30].

Figure 2 depicts the results of the experiment. The first plot on the top presents the SNR achieved when the dictionary is initialized with random examples from the training data. We can note that, while K-SVD achieves a good approximation quality, it does it at the expense of high coherence $\mu \approx 0.95$. On the other hand, INK-SVD is able to achieve a lower coherence $\mu = 0.5$ while maintaining a SNR$> 20dB$ and, after this value, the approximation quality drops linearly with the mutual coherence. The Grassmannian method achieves a correlation $\mu \approx 0.45$, but with a worst SNR $\approx 8dB$.

The other two plots corresponding to DCT and Gabor initializations display overall a poorer approximation quality. In these cases, Grassmannian fails to significantly decorrelate the dictionaries and achieves a very poor SNR, while INK-SVD is able to decorrelate the dictionaries up to $\mu = 0.2$ with a small loss in approximation accuracy.

5. CONCLUSION

We provided an algorithm to learn a dictionary with fixed coherence from training data. The coherence itself is a parameter of the algorithm that can be tuned to fit the needs of an application. Experiments on musical sound have shown that our algorithm can significantly reduce the coherence of the dictionary while almost preserving the approximation quality.

The proposed decorrelation method is generic and can be used in other contexts. In our experiments it even proved better than the Grassmannian method for coherence minimization. Yet there are still many theoretical questions to be answered, the biggest one being obtaining convergence guarantees depending on the coherence threshold.

It might also be possible to improve the approximation quality by specializing the decorrelation for the learning task. In the decorrelation problem (6), one could replace the distance to the original dictionary by the dictionary learning error (2).

6. REFERENCES

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

[2] J.A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[3] M. Yaghoobi, L. Daudet, and M.E. Davies, "Parametric dictionary design for sparse coding," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4800–4810, Dec. 2009.

[4] I. Ramirez, F. Lecumberry, and G. Sapiro, "Universal priors for sparse modeling," in *Proc. CAMSAP'09*, Dec. 2009, pp. 197–200.

[5] 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, Nov 2006.

[6] Joel A. Tropp, Inderjit S. Dhillon, Robert W. Jr. Heath, and Thomas Strohmer, "Designing structured tight frames via an alternating projection method," *IEEE Transactions on Information Theory*, vol. 51, no. 1, pp. 188–209, Jan. 2005.

[7] Michael Elad, *Sparse and Redundant Representations*, Springer, 2010.

[8] Ivan Damnjanovic, Matthew E. P. Davies, and Mark D. Plumbley, "SMALLbox - an evaluation framework for sparse representations and dictionary learning algorithms," in *Proc. LVA/ICA'10*, 2010, pp. 418–425.