Chasing butterflies: In search of efficient dictionaries
Luc Le Magoarou, Rémi Gribonval

To cite this version:
Luc Le Magoarou, Rémi Gribonval. Chasing butterflies: In search of efficient dictionaries. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr 2015, Brisbane, Australia. 10.1109/ICASSP.2015.7178579 . hal-01104696v2

HAL Id: hal-01104696
https://hal.archives-ouvertes.fr/hal-01104696v2
Submitted on 23 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CHASING BUTTERFLIES: IN SEARCH OF EFFICIENT DICTIONARIES

Luc Le Magoarou, Rémi Gribonval
Inria
Centre Inria Rennes - Bretagne Atlantique

ABSTRACT

Dictionary learning aims at finding a frame (called dictionary) in which training data admits a sparse representation. Traditional dictionary learning is limited to relatively small-scale problems, because dense matrices associated to high-dimensional dictionaries can be costly to manipulate, both at the learning stage and when using this dictionary for tasks such as sparse coding. In this paper, inspired by usual fast transforms, we consider a dictionary structure allowing cheaper manipulation, and we propose a learning algorithm imposing this structure. The approach is demonstrated experimentally with the factorization of the Hadamard matrix and on image denoising.

Index Terms— Sparse representations, dictionary learning, low complexity, image denoising

1. INTRODUCTION

Sparse representations using dictionaries are a common way of providing concise descriptions of high-dimensional vectors. With an appropriate dictionary $\mathbf{D}$, one can find a sparse representation vector $\gamma$ such that a vector $x$ of interest can be approximated as $x \approx \mathbf{D}\gamma$.

Traditionally, one can distinguish two ways of choosing a good dictionary for the data at hand. The dictionary can be either an analytic dictionary derived from a mathematical formula, such as the Fourier and Hadamard transforms or wavelets, or a learned dictionary, that is automatically inferred from training data. Analytic dictionaries are in general computationally efficient because of associated fast algorithms (e.g., the Fast Fourier Transform (FFT) or the Discrete Wavelet Transform (DWT))

In this paper the goal is to design dictionaries that are as well adapted to the data as learned dictionaries, while as fast to manipulate, and we propose a learning algorithm imposing this structure. The approach is demonstrated experimentally with the factorization of the Hadamard matrix and on image denoising.

Diagonal matrices, each of its columns fixed in (2) (let us call it $\mathbf{Q}$). Typically $n \gg d \geq 2$. In this framework, multi-layer sparse dictionary learning boils down to a factorization of the data matrix $\mathbf{X}$ into $M + 1$ sparse factors (the rightmost factor being the sparse representation matrix $\mathbf{Y}$) while multi-layer sparse dictionary factorization reduces to a factorization of the dictionary $\mathbf{D}$ into $M$ sparse factors. These two problems are very similar, and introducing the matrix $\mathbf{Y}$ to embody the matrix to factorize ($\mathbf{X}$ or $\mathbf{D}$) and the integer $Q$ to embody the desired number of factors, we propose to consider one general optimization problem:

Minimize

$$\sum_{j=1}^{Q} ||\mathbf{Y} - \prod_{j=1}^{Q} \mathbf{S}_j ||_F^2 + \sum_{j=1}^{Q} g_j(\mathbf{S}_j),$$ (2)

where the $g_j$s are penalties promoting sparsity and possibly additional constraints, and the factors $\mathbf{S}_j$s are of compatible sizes. We call (2) the Multi-layer Sparse Approximation (MSA) problem.

Related work. Similar matrix factorization problems have been studied recently in several domains. In dictionary learning, two main lines of work have begun to explore this way. In [4], the authors propose the sparse-KSVD algorithm (KSVD) to learn a dictionary whose atoms are sparse linear combinations of atoms of a so-called base dictionary. The base dictionary should be associated with a fast algorithm (it takes the form (1)), so that the whole learned dictionary can be efficiently stored and manipulated. It can be seen as having the $Q \approx 2$ leftmost factors fixed in (2) (let us call it $\mathbf{D}_{base}$), the second factor being the sparse representation of the dictionary over the base
dictionary ($D = D_{base}S_2$), and the first being the sparse representation matrix $\Gamma = S_1$. One drawback with this formulation is that the learned dictionary is highly biased toward the base dictionary, which decreases adaptability to the training data. In [3], the authors propose to learn a dictionary in which each atom is the composition of several circular convolutions using sparse kernels with known supports, so that the dictionary is fast to manipulate. Their problem can be seen as [2], with the $g_j$s corresponding to the $M$ leftmost factors imposing sparse circulant matrices. This formulation is limited in nature to the case where the dictionary is well approximated by a product of sparse circulant matrices.

In statistics, a related problem is to approximately diagonalize a covariance matrix by a unitary matrix in factorized form [1], which can be addressed greedily [6] using a fixed number of elementary Givens rotations. Here we consider a richer family of sparse factors and leverage recent non-convex optimization techniques. In machine learning, similar models were explored with various points of view. For example, sparse multi-factor NMF [8] can be seen as solving problem [2] with the Kullback-Leibler divergence as data fidelity term and all $S_j$s constrained to be non-negative. Optimization relies on multiplicative updates, while we rely on proximal iterations. In the context of deep neural networks, identifiability guarantees on the network structure have been established with a generative model where consecutive network layers are spuriously connected at random, and non-linearities are neglected [9, 10]. The network structure in these studies matches the factorized structure [1], with each of the leftmost factors representing a layer of the network and the last one being its input. Apart from its hierarchical flavor, the identification algorithm in [9, 10] differs from the proximal method proposed here.

An objective somewhat related to that of having a computation-efficient dictionary is that of being able to rapidly compute the leftmost factors representing a layer of the network and the last one being the input. This results in the optimization problem:

$$\min_{\lambda, S_1, \ldots, S_Q} \Psi(S_1, \ldots, S_Q, \lambda) := \frac{1}{2} \|Y - \lambda \sum_{j=1}^{Q} S_j^T \|_F^2 + \sum_{j=1}^{Q} \delta_{E_j}(S_j),$$

with $E_j = \{ A \in \mathbb{R}^{c_j \times c_j} : \|A\|_0 \leq p_j, \|A\|_F = 1 \}$.

### 3.2. Coping with the scaling ambiguity

To avoid scaling ambiguities, it is common [5, 8] to normalize the factors and introduce a multiplicative scalar $\lambda$ in the data fidelity term. This results in the optimization problem:

$$\min_{\lambda, S_1, \ldots, S_Q} \Psi(S_1, \ldots, S_Q, \lambda) := \frac{1}{2} \|Y - \lambda \sum_{j=1}^{Q} S_j^T \|_F^2 + \sum_{j=1}^{Q} \delta_{E_j}(S_j),$$

### 3.3. Proposed algorithm

The optimization problem [4] is highly non-convex, and the sparsity enforcing part is non-smooth. We leverage recent advances in non-convex optimization to express an algorithm with convergence guarantees to a stationary point of the problem. In [13], the authors consider cost functions depending on $N$ blocks of variables:

$$\Psi(x_1, \ldots, x_N) := H(x_1, \ldots, x_N) + \sum_{j=1}^{N} f_j(x_j),$$

where the function $H$ is smooth, and the $f_j$s are proper and lower semi-continuous. To handle this cost function, the authors propose an algorithm called Proximal Alternating Linearized Minimization (PALM [13], that updates alternatively each block of variable by a proximal (or projected in our case) gradient step.

**Projection operator $P_{E_j}$.** With $E_j$ defined as in Section 3.2, the projection operator $P_{E_j}$ simply keeps the $p_j$ greatest entries (in absolute value) of its argument, sets all the other entries to zero, and then normalizes its argument so that it has unit norm.

**Gradient and Lipschitz moduli.** To specify the iterations of PALM specialized to our problem, we fix the iteration $i$ and the factor $j$, and denote $S_j^i$ the factor that we are updating. Let $L := \prod_{k=1}^{M} S_j^i$ what is on its left and $R := \prod_{k=0}^{i-1} S_j^i$ what is on its right (with the convention $S_0 = I$). With these notations we have, when updating the $j$th factor $S_j$:

$$\nabla_{S_j} H(S_1^i, \ldots, S_j^{i+1}, S_{j+1}^i, \ldots, S_Q^i, \lambda) = \lambda^T (\lambda^T L S_j^i R - Y) R^T,$$

which has a Lipschitz modulus with respect to $\|S_j\|_F$:

$$L_j(L, R, \lambda) = (\lambda^T)^2 \|R\|_2^2 + \|L\|_2^2.$$

Once the first $Q$ factors are updated we need to update $\lambda$. Denoting $Y = \prod_{k=1}^{Q} S_k^i$, we get:

$$\nabla_{\lambda} H(S_1^i, \ldots, S_Q^i, \lambda) = \lambda^T (\lambda^T Y - Y^T Y),$$

which has the Lipschitz modulus with respect to $\|S_1\|_F$:

$$L_j(L, R, \lambda) = (\lambda^T)^2 \|R\|_2^2 + \|L\|_2^2.$$

An explicit version of the algorithm is given in Algorithm [1] in which the factors are updated alternatively by a projected gradient step (line 6) with a stepsize controlled by the Lipschitz modulus of the gradient (line 5). We can solve for $\lambda$ directly at each iteration.
Indeed, when general. This observation suggested to adopt a hierarchical strategy.
and allowing more non-zero entries per factor led to better results in
was very often not satisfactory (high data fidelity term).
also the product
dictionary factorization
the data fidelity term of the cost function is small, there is no guar-
ever, while we are primarily interested in stationary points where
Data matrix or dictionary Y; desired number of factors Q;
Input: Data matrix or dictionary Y; desired number of factors Q;
Output: The estimated factorization:
3.4. Practical strategy
Algorithm I presented above factorizes a data matrix into Q sparse
factors and converges to a stationary point of problem (4).
However, while we are primarily interested in stationary points where
the data fidelity term of the cost function is small, there is no guar-
antee that the algorithm converges to such a stationary point.
This was observed on a very simple experiment where Algorithm I
was used for a dictionary factorization task, with a dictionary Y = D
with a known factorization in M factors: D = \prod_{j=1}^{M} S_j, such as
the Hadamard dictionary. The naive approach consisted in taking
directly Q = M in Algorithm I and setting the constraints so as to
reflect the actual sparsity of the true factors. This simple strategy
performed quite poorly in practice, and the attained local minimum
was very often not satisfactory (high data fidelity term).
We noticed experimentally that taking fewer factors (Q small)
and allowing more non-zero entries per factor led to better results in
general. This observation suggested to adopt a hierarchical strategy.
Indeed, when Y = \prod_{j=1}^{Q} S_j is the product of Q sparse factors, it is
also the product Y = T_1 S_1 of 2 factors with T_1 = \prod_{j=2}^{Q} S_j, so
that S_1 is sparser than T_1. Our strategy is then to factorize the input
matrix Y in 2 factors, one being sparse (corresponding to S_1), and
the other less sparse (corresponding to T_1). The process can then
be repeated on the less sparse factor T_1, and so on until we attain
the desired number Q of factors. Denoting T_k = \prod_{j=k+1}^{Q} S_j, a
simple calculation shows that if we expect each S_j to have roughly
O(h) non-zero entries per row, then T_k cannot have more than
O(h^{Q-(k+1)}) non-zero entries per row. This suggests to decrease
exponentially the number of non-zero entries in T_k with k. This
strategy turns out to be surprisingly effective and the attained local
minima are very good, as illustrated in the next section. In a sense,
this hierarchical approach is similar to learning a deep neural net-
work layer by layer, as done in [9, 10].

Our preliminary theoretical analysis of the proposed approach suggests that the results of [9, 10]
could in fact be leveraged to explain the observed empirical success.

The proposed hierarchical strategy is summarized in Algo-

rithm 2 where we need to specify at each step the constraint sets
related to the two factors. For that, let us introduce some notation: 
\mathcal{E}_k will be the constraint set for the right factor and \mathcal{E}_k the one
for the left factor at the kth factorization. The global optimization step
(line 5) is done by initializing palm4MSA with the current values of
\{S_j\}_{j=1}^{k} and T_k. It is here to keep an attach to the data matrix
Y. Roughly we can say that line 3 of the algorithm is here to yield
complexity savings, whereas line 5 is here to improve data fidelity.

Algorithm 2 Hierarchical factorization
Input: The data matrix or dictionary Y, the desired number of fac-
tors Q and the constraint sets \mathcal{E}_k and \mathcal{E}_k, k \in \{1 \ldots Q-1 \}.
1: T_0 \leftarrow Y
2: for k = 1 to Q - 1 do
3: Factorize the residual T_{k-1} into 2 factors: 
\lambda \mathcal{E}_k \{ F_2, F_1 \} = \text{palm4MSA}(T_{k-1}, \mathcal{E}_k, \ldots )
4: T_k \leftarrow \lambda F_2 and S_k \leftarrow F_1
5: Global optimization: 
\lambda \mathcal{E}_k \{ S_j \}_{j=1}^{k} = \text{palm4MSA}(Y, k + 1, \mathcal{E}_k, \mathcal{E}_k, \ldots )
6: end for
7: S_0 \leftarrow T_{Q-1}
Output: The estimated factorization \lambda \{ S_j \}_{j=1}^{Q}.

In the case of structured dictionary learning, the first factor S_1
corresponds to the sparse coefficients matrix \Gamma, and the first iteration
of factorization (k = 1) amounts to classical dictionary learning. It
can thus be done with any classical dictionary learning method, such
as K-SVD. Moreover, to avoid manipulating the coefficient matrix
(which is often way bigger than the other factors) one can keep it
fixed, and update it only by Orthogonal Matching Pursuit (OMP)
[15] after the global optimization step (between lines 5 and 6). This
simple modification of Algorithm 2 makes the hierarchical strategy
look like a traditional dictionary learning algorithm [3] where update
of the coefficients and update of the dictionary alternate, except that
in the present case the dictionary update is different and seeks to
make complexity savings (reducing RC), and the coefficient update
can be done efficiently taking advantage of the dictionary structure.

4. EXPERIMENTS
4.1. Dictionary based image denoising
Our experimental scenario for the structured dictionary learning task
follows the workflow of a classical dictionary based image denoi-
sing. First, n = 10000 patches x_i of size 8 x 8 (dimension d = 64)
are randomly picked from an input 512 \times 512 noisy image (with
PSNR= 22.1dB), and a dictionary is learned on these patches. Then
the learned dictionary is used to denoise the entire input image by
computing the sparse representation of all its patches in the dictio-
nary using OMP allowing each patch to use 5 dictionary atoms.
The image is reconstructed by averaging the overlapping patches.
Various dictionary learning methods were compared, each learning a
time four times complete dictionary (D \in \mathbb{R}^{64 \times 2560}).

Settings of our algorithm. We tested several configurations for Al-
gorithm 2 and we present here one that exhibits a good tradeoff be-
 tween relative complexity (RC) and adaptation to the data (we call
it MSA for Multi-layer Sparse Approximation). Inspired by usual
fast transforms, we chose a number of factors Q close to the log-
arithm of the signal dimension d = 64, here Q = 5. The sizes of
the factors are: \Gamma = S_1 \in \mathbb{R}^{256 \times 10000}, S_2 \in \mathbb{R}^{64 \times 256}
and \mathbf{S}_3, \ldots, S_Q \in \mathbb{R}^{64 \times 64} Algorithm 2 was used with the modifica-
tions presented in the previous section i.e. the first factorization (k =
1 in Algorithm 2) is done by KSVD [16] and S_1 was updated only
by OMP allowing each patch to use 5 dictionary atoms. In this setting, \( \mathcal{E}_1 \) and \( \mathcal{E}_1 \) do not need to be specified, but implicitly correspond to \( \mathcal{E}_1 = \{ \mathbf{A} \in \mathbb{R}^{64 \times 256}, ||\mathbf{a}_i||_0 \leq 5 \land i \in \{1 \ldots 10000\} \} \) and \( \mathcal{E}_1 = \{ \mathbf{A} \in \mathbb{R}^{64 \times 256}, ||\mathbf{a}_i||_0 = 1 \land i \in \{1 \ldots 256\} \} \). Regarding the other factorizations \( k > 1 \) in Algorithm 2 we allowed in each factor 4 non-zero entries per column in average, so that the considered constraint sets were: \( \mathcal{E}_2 = \{ \mathbf{A} \in \mathbb{R}^{64 \times 256}, ||\mathbf{a}_i||_0 \leq 1024, ||\mathbf{a}_f||_0 = 1 \} \) and for \( k \in \{3 \ldots Q - 1\} \): \( \mathcal{E}_k = \{ \mathbf{A} \in \mathbb{R}^{64 \times 64}, ||\mathbf{a}_i||_0 \leq 256, ||\mathbf{a}_f||_0 = 1 \} \). In order to decrease RC for the learned dictionary at each new factorization, the number of non-zero entries in the residual was divided by 2 at each step, starting from \( p = 1.3 \times 2048 \), leading to, for \( k \in \{2 \ldots Q - 1\} \): \( \mathcal{E}_k = \{ \mathbf{A} \in \mathbb{R}^{64 \times 64}, ||\mathbf{a}_i||_0 \leq \frac{p}{2^k}, ||\mathbf{a}_f||_0 = 1 \} \). The stopping criterion for \( \text{palm4MSA} \) was a number of iterations \( N_t = 50 \).

**Baselines.** We compared Algorithm 2 with the following methods. All methods involve a coefficient update step which is performed using OMP allowing each patch to use 5 dictionary atoms:

- **KSVD** [16]. We used the implementation described in [17], running 50 iterations (empirically sufficient to ensure convergence).
- **Online Dictionary Learning (ODL)** [18], running 200 iterations (empirically sufficient to ensure convergence).
- **Sparse KSVD (KSVDS)** [4], a method that seeks to bridge the gap between learned dictionaries and analytic dictionaries. The implementation of [4] is used, with \( \mathbf{D}_{\text{base}} \) the four-times overcomplete 2D-Discrete Cosine Transform (DCT) matrix and 50 iterations, ensuring convergence in practice. The columns of the estimated dictionary \( \mathbf{D} \) are \( \ell_0 \)-sparse in \( \mathbf{D}_{\text{base}} \).

**Performance measures.** The computational efficiency of the dictionary is measured through the Relative Complexity (RC) quantity introduced in Section 3.3. For the KSVDS method, we used the formula provided in [4] to compute RC. The quality of approximation at the learning stage and at the denoising stage is expressed using the classical PSNR measure (even though at the learning stage we are manipulating patches and not entire images).

**Discussion of the results.** Table 1 summarizes the performance of the different methods. First of all, the proposed method is able to provide dictionaries that are much more computationally efficient than the others (RC = 0.13). Second, and perhaps more surprisingly, while the proposed method performs worse than the others at the learning stage, it outperforms them at the denoising stage. This indicates that its generalization properties are better [19 20 21], which is consistent with the lower number of parameters to learn compared to KSVD and ODL. Regarding KSVDS, its poorer performance may be explained by the fact that it is highly biased toward its base dictionary. Moreover, we suspect the proposed factorized structure to have the interesting property of being more intrinsically unable to fit noise than standard dictionaries.

### 4.2 Retrieving the fast Hadamard transform

Here is presented an experiment regarding the dictionary factorization task. Consider a data matrix \( \mathbf{Y} = \mathbf{D} \) with a known factorization in \( M \) factors, \( \mathbf{D} = \prod_{i=1}^{M} \mathbf{S}_i \); in Section 3.4 we evoked the failure of Algorithm 1 for this factorization problem. In contrast, Figure 2 illustrates the result of the proposed hierarchical strategy (Algorithm 2) with \( \mathbf{D} \) the Hadamard dictionary in dimension \( n = 32 \). The obtained factorization is exact and as good as the reference one in terms of complexity savings \( \left( 2n \log n - \text{non-zero entries in the factors} \right) \), so in a way the chased butterflies were caught. The running time is less than a second. Factorization of the Hadamard matrix in dimension up to \( n = 1024 \) showed identical behaviour, with running time \( O(n^2) \) up to ten minutes.

### 5. CONCLUSION

We proposed a matrix factorization framework with convergence guarantees that provides a flexible tradeoff between computational efficiency and data fidelity. It shows promising results in image denoising and is able to automatically retrieve an ideally compact factorization of the fast Hadamard transform. Besides the obvious need to better understand the role of its parameters in the control of the desired tradeoff, a particular challenge will be to leverage the gained complexity to speed up the factorization process itself, in order to efficiently learn efficient dictionaries. Moreover, the algorithm underlying the factorization is very general, and it would be interesting to take advantage of its versatility in order to incorporate other structures in the factors. This line of work could lead for example to learn data-driven efficient wavelets on graphs [23], by constraining the support of the factors. Finally, a concern in the numerical analysis community is to be able to approximate certain integral operators with a butterfly structure [24 25]. The method proposed in this paper could help finding automatically such approximate fast transforms.

<table>
<thead>
<tr>
<th>Learning (PSNR)</th>
<th>Denoising (PSNR)</th>
<th>Complexity (RC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSVD</td>
<td>24.71</td>
<td>27.55</td>
</tr>
<tr>
<td>ODL</td>
<td>24.62</td>
<td>27.51</td>
</tr>
<tr>
<td>KSVDS</td>
<td>24.16</td>
<td>27.64</td>
</tr>
<tr>
<td>MSA</td>
<td>23.63</td>
<td>29.38</td>
</tr>
</tbody>
</table>

**Table 1.** Image denoising results, averaged over the standard image database taken from [23] (12 standard grey 512 × 512 images). The best result of each column is bold.

**Fig. 1.** Example of denoising result. It is a zoom on a small part of the “house” standard image.

**Fig. 2.** Hierarchical factorization of the Hadamard matrix of size 32 × 32. The matrix is iteratively factorized in 2 factors, until we have \( Q = 5 \) factors, each having \( p = 64 \) non-zero entries.
6. REFERENCES


