Response to reviews of "Nonnegative Low-rank Sparse Component Analysis"

J.E. Cohen and Nicolas Gillis

February 2019

Foreword : Keep in mind that in the ICASSP publication process, there is no direct exchange between reviewers and authors. This means that the following answers have not been examined by the reviewers. In particular, Reviewer 2 attacked the validity of the contribution, but no further exchanges are available to discard with certitude the suspicions that were raised. We hope the following answers will nonetheless strengthen the credibility of our work.

Reviewer 1

Comment on Clarity of Presentation: I had a long text here which disappeared due to a bug in the homepage, sorry for not rewriting it, but the key points were: Terminology should be explained to non-specialists. What is a "data-point", "clean spectral bands" "spectral signature" etc. Initially size of k r and d is not known, making it hard to understand the setup. For example, does rank(D)=r mean that D has full rank or low rank?? In the final example, the dimensions of B seem to be off. Check the paper for minor misprints (brute fore, an suboptimal....)

We thank the reviewer for his remarks. The typos were corrected before final submission to the conference; in particular, the dimensions of B in the final example were indeed wrong. Also, we have tried to clarify the setup, within the limit of pages provided for the paper.

The paper presents an interesting problem which there seem to be a fair amount of prior work on. The authors cite theory produced by themselves in a more general context, but there is no noteworthy mathematical theory in the present paper, except for a small example of limited interest.

Indeed, theory is not the focus of the present paper. However, the presented example provides a first idea on how the problem of identifiability of nonnegative LRSCA could be approached, using intersections of polytopes. As far as we know, although indeed of limited scale, this is the very first result on identifiability of Dictionary Learning with nonnegativity.

So the value of the paper is in the suggested algorithms and numerical observations. The paper divides previous work in two categories, alternating optimization AO with compressed sensing tricks (which I know well), and something they call subspace clustering methods (which I do not know). AO is a non-convex strategy and it is known that it may not converge to a global optimal solution, but still it is used and it seems convergence to local optima has been proved. Prior work has used AO together with standard compressed sensing tricks (which are suboptimal and lead to a bias) to solve the problem at hand. The authors suggest doing the same but instead of compressed sensing solve brute force the combinatorial problem, arguing that often this is feasible due to low dimensionality.

This is all correct in my opinion. The low-dimensionality is of course relative to the computational power.

Here is my main problem with the paper: This is a pretty obvious point, compressed sensing was invented to deal with problems where the dimensionality prohibits this approach practically. So the fact that their algorithm outperforms the one with the compressed sensing in it is pretty obvious, and needs no numerical section to convince me.

I quite disagree with the comment somehow. Of course we expected the brute force algorithm (which itself is of course not the contribution here) to perform better than heuristics, but we did not expect such a gap in performances (see comments of reviewer 2 for a lengthy discussion on Figure 2). In fact, we first worked on the brute force algorithm to provide an upper bound on what could be achieved, in order to design heuristics. In short, although the improvement is somewhat expected, one of the contributions here is to quantify how much it improves with respect to heuristics (OMP) and relaxation (LASSO).

One interesting observation is that despite solving each update optimally in AO, their algorithm still goes to a local minima sometimes, and this leads the authors to suggest another algorithm based on the other category

(subspace clustering). Apparently this works better but the authors write that it is slower, which makes me skeptical to it for anything but problems where involved dimensions k and r are very small. For this case however, it may well be that their algorithm NOLRAK is a top contribution, I have no clue. In the numerical section a relevant and interesting real life application with low r and k is presented (an image classification problem).

Indeed this is the main contribution. However, it seems we were not clear in the first version of the paper: NOLRAK is **not** as slow as the brute force algorithm. In fact, for each iteration, the cost is driven by the computation of a NMF. Therefore, the complexity of NOLRAK is not combinatorial like that of the brute force ESNA. We have clarified this part in the revised paper.

Reviewer 2

This paper describes non-negative matrix factorization. This problem can be considered as a special case of dictionary learning. Here, the special case means the non-negative dictionary matrix D and non-negative as well as sparse coefficient matrix B. If the dictionary matrix D is given, the problem setting becomes the same as the compressed sensing with multiple measurement vectors and non-negative constraints on a sensing matrix D and coefficients B. The problem that the authors are considering in the paper is interesting. However, there are some very doubtful and suspicious simulation results. I am personally very doubtful of the correction of this paper, especially the simulation results. I address my doubtful concerns in the following comments. In summary, I am not sure of this paper to be accepted or not to IEEE ICASSP 2019 conference.

We thank the reviewer for his comments. The reviewer's concerns about Figure 2 have been, as much as possible, investigated (see answer below). We hope we have clarified as much as possible the presented results, as well as the modifications to the LASSO-HALS that have been made accordingly to improve its performance. Still, the conclusion we draw after this careful revision is rather unchanged (although LASSO-HALS is in a much better spot): both the brute-force algorithm and the subspace clustering algorithms vastly outperform heuristics-based methods in some settings, for an exact decomposition problem. We will be happy to further discuss during the conference, since there is probably more to investigate here.

1. My biggest concern of this paper is that the simulation result in Fig. 2 of the paper is very suspicious. The successful rate of LASSO-HALS and the NMF-HALS are all zeros in terms of Quasi perfect reconstruction of D. This part is really suspicious. From my personal experience on compressed sensing, LASSO (in noisy case) or (L1 minimization in the non-noise case) can provide a quite good signal recovery result. For example, when d=125, k=3, which is a quite sparse case, as long as the matrix D is correct, LASSO (or L1 minimization) will work well definitely with non-negative constraint. But, the result of Fig.2 shows the completely opposite results. I am very suspicious about the correctness of the results of Fig.2, which is the main simulation result. Also, When we check the both simulation results of Quasi perfect reconstruction of D and average relative MSE on D with k=3, the results from LASSO-HALS, NMF-HALS, a.set NNOMP, and SuNNOMP are all similar in terms of the average relative MSE on D with k=3. But, in the Quasi perfect reconstruction of D, the successful rate of LASSO-HALS and NMF-HALS are zero, while that of a.set NNOMP and SuNNOMP are quite good. The inconsistency in the simulation results makes me to ask about the correctness of the simulation results.

We provide a complete answer discussing previous choices, possible issues, and modifications that were made to the simulations accordingly.

0.0.1 Near Perfect recovery

We chose to illustrate the performance of the various tested algorithms using two criteria: reconstruction error, and a "near perfect recovery" test. The reason we did this is that the distribution of the error is bimodal for all methods:

- sometimes the algorithms converge to a solution with tiny reconstruction error but large errors on the columns of the dictionary (typically larger than 10^{-2} relative error). This happens when a local minimum is reached, for which the data is almost reconstructed although estimated atoms are far from the true ones.
- sometimes both reconstruction error and relative error on the columns of the dictionary are vanishing.

Of course, the relative reconstruction error on D is driven by the simulations with large reconstruction errors. The fact that some methods have reconstruction error vanishing to machine precision rather, than to

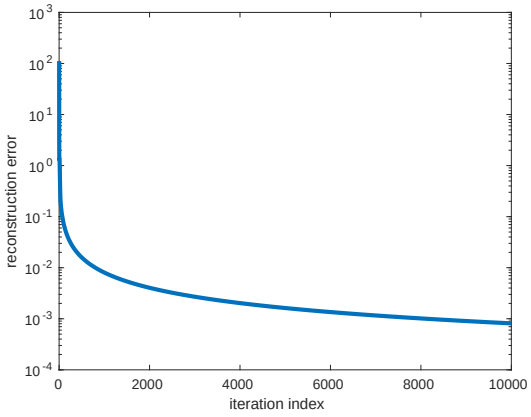

Figure 1: An instance of reconstruction error using LASSO-HALS with $d = 20$ and $k = 2$. Regularization parameter λ was set to 10^{-1}.

10^{-4} for instance, is invisible on the right part of Figure 2. This is why it is possible for a method to have in average a small MSE on D but still a poor "quasi perfect reconstruction" score.

The difficulty here is that the LASSO-HALS error on columns of D, when both reconstruction error and dictionary reconstruction error are vanishing, does not reach machine precision. Rather, it oscillates around $[10^{-3}, 10^{-5}]$ in our experiments. This can be explained by the fact that the decrease of the fitting term $\|M - DB\|_F^2$ is quite slow using this algorithm. We have plotted the reconstruction error in Figure 1 below to illustrate this fact. This may be explained by the fact that the ℓ_1 norm induces a small bias in coefficients, which deteriorates precision slightly.

We decided to use 10^{-4} as a "quasi perfect reconstruction" reconstruction criterion. Having a fixed threshold is not ideal, since the dictionary reconstruction error is not clearly distributed in two clusters separated by 10^{-4}. However, it was a simple compromise to illustrate the performance of the algorithms. For the revised version, we decided to keep this threshold.

However, looking back on this problem, a better way to quantify exact recovery would be actually to monitor the locus of coefficients B, and count the number of miss-classified samples; since ICASSP prohibits to significantly change the contents of a paper from its originally submitted version, this will be kept for a longer version.

0.0.2 Code modifications

We decided to change the implementation of LASSO-HALS in ways that improved its performance. In particular:

- we used in the first version a modified accelerated HALS. To make things simpler, the algorithm now switches from estimating D to B if a maximal number of iterations is reached, or if the error stops decreasing significantly. This avoids some instabilities.
- we removed the normalization of D inside the algorithm. This **slightly** deteriorated the results for some reason, and we apologize for introducing this bias (the variation is minor, but still this improves a little bit).
- we observed that convergence was typically very slow (input figure). Therefore, we increased the number of outer iterations by a factor 10 from the first version. This helps in reducing the dictionary reconstruction error for LASSO-HALS by a large margin. Actually, using 10 times more iterations, the 10^{-4} error threshold was fairer to compare LASSO-HALS with other methods.

0.0.3 Proving the correctness of our simulation

Since the question attacked the correctness of the simulation, we below discuss how we chose the λ parameter, and prove the correctness of the algorithm implementation to exactly reconstruct coefficients B when the true D is provided.

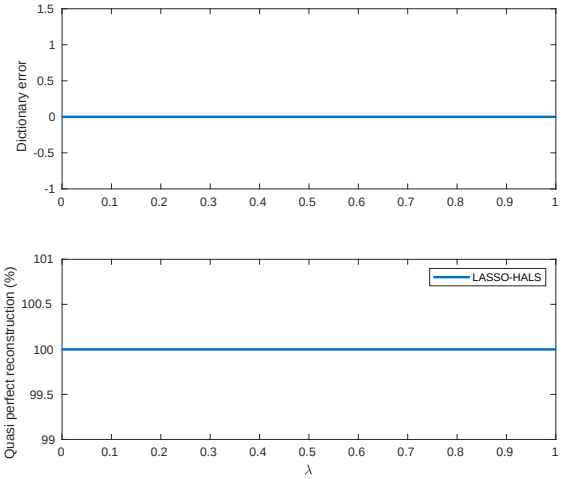

Figure 2: Quasi perfect reconstruction percentage and dictionary MSE, when the true dictionary is known. Here, $d = 10$ and $k = 3$.

d	$k = 3$ empirical best λ	$k = 2$ empirical best λ
125	$[10^{-3}, 10^{-2}]$	$[10^{-3}, 0.1]$
50	$[10^{-2}, 0.2]$	$[10^{-3}, 0.1]$
25	$[10^{-2}, 0.1]$	$[10^{-3}, 0.1]$
20	$[2 \times 10^{-3}, 10^{-2}]$	$[0.1, 0.2]$
10	0.1	$[0.1, 0.2]$
5	$[0.1, 0.2]$	0.2
4	$[10^{-2}, 0.1]$	0.1

Table 1: 10 realisations of $M = DB$ where drawn to estimate the perfect lambda for each scenario using the ground truth for evaluation. We then used these "oracle" λ for producing the revised Figure 2.

Simulation 1 : perfect initial dictionary Here we illustrate the fact that, given the true dictionary, the LASSO problem solved with HALS is very efficient. We only show this for the case $d = 10$, which seems difficult in the dictionary learning case where D is unknown. Figure 2 shows something already well-known in the literature: the LASSO problem is efficiently solved in the nonnegative case using algorithms such as Hierarchical ALS. The reconstruction error, on $N = 100$ trials, is always close to machine error. This proves that the code performs as expected (there is no hidden bug or weird behavior). This also illustrates the fact that, in Nonnegative Dictionary learning, the difficulty is that both the dictionary and the coefficients are unknown. Using intuition stemming from sparse coding only is therefore misleading.

Simulation 2 : Finding the good lambda for each dimensional setting. We used the following table to chose the best regularization parameter λ.

0.0.4 The new "Figure 2" of the manuscript

We have remade "Figure 2" accounting for all the preciously mentioned changes, see Figure 3. We let the reviewer observe that we could present results in a way that made LASSO-HALS on par with other AO methods. However, this now hides the fact that the precision of LASSO-HALS when it succeeds in identifying the right atoms is poorer than when using other methods. The code will be available online for anyone to reproduce these experiments.

Also, since we took special care in implementing and optimizing the LASSO-HALS, and in particular the choice of λ was made knowing the exact solution, the performance of LASSO-HALS now is on par with the expectations of the reviewer.

2. The authors propose two algorithms to solve the optimization problem (1) of the paper. The novelty of the algorithms is not high. For ESNA, as the authors mentioned in the paper, it uses the burst-force approach, which has exponential complexity. Therefore, ESNA can be used for small r and k. Unlike ESNA, the NOLRAK can be used for large-scale problems. However, the NOLRAK is the special case of K-subspace in order to handle non-negative constraint. Therefore, in algorithm-wise, the novelty of the paper is not so significant.

This is the reviewer's opinion. We agree ESNA is nothing new. However, NOLRAK (which is not really

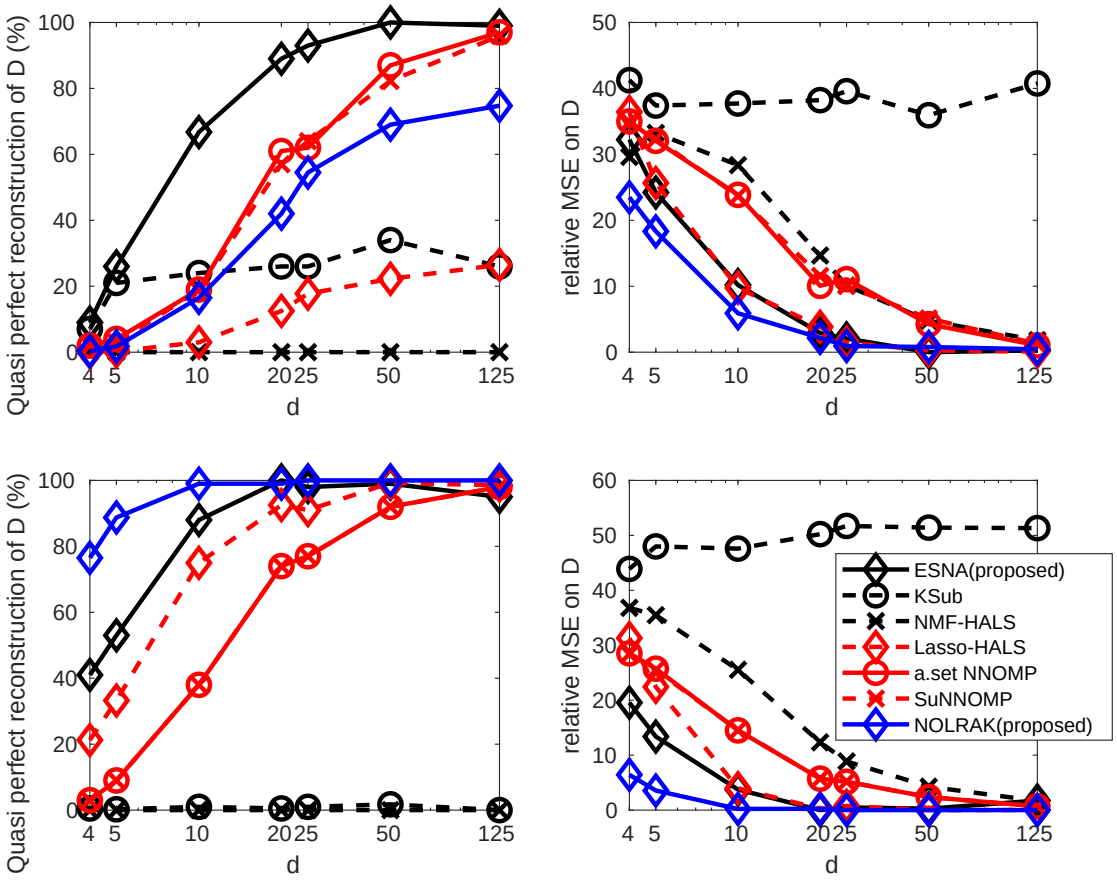

Figure 3: The revised Figure 2 of the manuscript

a trivial extension of K-Subspace, only the concept of subspace clustering is similar) is an original and new contribution, which impact is yet to be studied. Also, the performance gap between ESNA and heuristics is quite unexpected in our opinion.

Minor comments: 1. in Section 2, "Fig. 2" should be changed to "Fig. 1". 2. in the introduction of ESNA, "burst-fore approach" –> "burst-force approach"

We thank the reviewer for these remarks, we have made the required changes in the manuscript.

Reviewer 3

General Comments to Authors:

- *you define SCA and K-NMF in the exact case (M=DB) in which case there might not be a solution for arbitrary 'r'. Is that ok ?*
- *your define K-NMF in the exact case but solve for the approximate case (M ≈DB). Is that consistent ?*
- *it's not clear how you solve for D in ESNA*
- The existence of a solution is indeed assumed. In practice, only an approximate decomposition is sought, but this is outside of the scope of this paper.
- Yes, this is consistent. Indeed, to obtain the exact decomposition, it is sufficient to minimize any norm of the difference $M - DB$, since all norms are equivalent in finite dimension. So finding the solution to $\min\|M - DB\|_F^2$ under appropriate constraints is sufficient for finding the solution. If such a solution exists, the cost will decrease to zero. Of course, there might be other ways to obtain the solution to the exact factorization, but the solution will also minimize this cost function.
- We have improved the explanation of the computation of D. We would have liked to input a full algorithm as is usually done. However, we have too much content to describe, and using a fourth of a page for this would be too much. We are fully aware that the description of the algorithms is somewhat difficult to follow, and we apologize for the inconvenience.

typos: LRSCA existS I guess B should be of size 6x94249 in Spectral Unmixing section

Thanks you for pointing these typos.

Reviewer 4

This article proposes two algorithms to perform non negative low rank sparse component analysis. These algorithms seem well founded and tested. Some work is needed in order to clarify this contribution.

The authors do not define what is a data point in their factorization model (is a coefficient of M, a column, a row?, what is m_j?).

We have now precised in the manuscript that a data point is a sample point, which is nothing more than a column m_j of M.

I find the informal discussion on identifiability for d=3 not well written. Either give a real precise demonstration for this case or remove this part. It is obvious that identifiability can only be easier when you restrict the model.

We are sorry that the reviewer thinks this section is poorly written. In spite of our efforts, it is somewhat difficult to make it clearer with only using a single column. However, we do not wish to remove this section. As mentioned in the answer to Reviewer 1, this is the first time an explicit proof is given for the identifiability of Nonnegative LRSCA, and it uses tools that are not standard in the community. We agree the contribution is particularly tiny since we only consider a single scenario, but we hope this will lead to other contributions and discussions. This is not, of course, the core contribution of the paper.

The description of NOLRAK is not very well written. Writing clearly the algorithm and then commenting on the steps would be better. Maybe do the same for the first algorithm.

As stated in answer to reviewer 3, we wish we could add such a formal description of the algorithms, however the size limitation is prohibitive. We understand the idea of the reviewer, that removing the first section on identifiability frees enough space to add the algorithms, but we prefer to keep the section on identifiability.