ℓ_1/ℓ_2 regularized non-convex low-rank matrix factorization

Paris V. Giampouras*, Athanasios A. Rontogiannis*, Konstantinos D. Koutroumbas*
*IAASARS, National Observatory of Athens, GR-15236, Penteli, Greece
email: {parisg,tronto,koutroum}@noa.gr

***Abstract*—Low-rank matrix factorization plays a key role in a plethora of problems commonly met in machine learning applications dealing with big data as it reduces the size of the emerging optimization problems. In this work we introduce a novel low-rank promoting regularization function which gives rise to an algorithm that induces sparsity jointly on the columns of the matrix factors. Apart from the reduced computational complexity requirements it offers, the new algorithm also provides a *basis* of the sought low-rank subspace.**

I. Introduction

Low-rank matrix factorization has been at the heart of a great many problems dealing with big data. In mathematical terms it can be formulated via the following minimization problem:

$$\underset{\mathbf{X},\mathbf{W}}{\operatorname{argmin}} \quad l(\mathbf{Y}, \mathbf{X}\mathbf{W}^T) + \delta \|\mathbf{X}\mathbf{W}^T\|_{\mathcal{S}_p}^p \tag{1}$$

where $l(\cdot)$ denotes the function that measures the distance between data matrix $\mathbf{Y} \in \mathcal{R}^{N \times K}$ and its low-rank representation $\mathbf{X}\mathbf{W}^T$; with $\mathbf{X} \in \mathcal{R}^{N \times L}$, $\mathbf{W} \in \mathcal{R}^{K \times L}$ standing for the coefficients' matrix and the subspace matrix, respectively, (with $L \ll \min(K, N)$) and $\delta > 0$ being a regularization parameter. In the low-rank setting it is assumed that the data matrix $\mathbf{Y}$ can be well represented in a space of an unknown dimension $r \leq L$. In light of this, the second term of (1) denotes the low-rank inducing *Schatten-p* quasi-norm applied on the matrix $\mathbf{X}\mathbf{W}^T$. Recently, a wealth of algorithms have appeared differing on the selection of p, [1]. Note that the choice of p affects the convexity of problem (1), e.g. for $p = 1$ (which corresponds to the nuclear norm) the minimization problem is convex w.r.t. the product $\mathbf{X}\mathbf{W}^T$, while for $p < 1$ the problem is non-convex.

In the matrix factorization setting, (1) is replaced by a relaxed minimization problem, which arises by utilizing upper-bounds of the low-rank promoting terms [2]. Focusing on the nuclear norm, the celebrated tight upper bound of the nuclear norm defined as

$$\|\mathbf{X}\mathbf{W}^T\|_* \equiv \inf \sum_{l=1}^{L} \|\boldsymbol{x}_l\|_2 \|\boldsymbol{w}_l\|_2 \equiv \inf \frac{1}{2} \sum_{l=1}^{L} \left(\|\boldsymbol{x}_l\|_2^2 + \|\boldsymbol{w}_l\|_2^2 \right) \tag{2}$$

where $\boldsymbol{x}_l$ and $\boldsymbol{w}_l$ denote columns of $\mathbf{X}$ and $\mathbf{W}$, respectively, has been widely applied in numerous works offering stimulating results [3]–[5]. Recently, generalized approaches have been put forth corresponding to the non-convex scenario, where $p \in [0, 1)$, [6]. Such methods, sacrifice global optimality guarantees, [3], in favor of better estimation results.

II. Proposed problem formulation

Along those lines, we herein propose a novel non-convex low-rank promoting term which stems from the group-sparsity ℓ_1/ℓ_2 norm. The key idea, first introduced in [7], is to relate column sparsity imposition of $\mathbf{X}$ to that of column sparsity of the subspace matrix $\mathbf{W}$. The proposed minimization problem (different from that introduced in [7]) is expressed as follows:

$$\underset{\mathbf{X},\mathbf{W}}{\operatorname{argmin}} \quad l(\mathbf{Y}, \mathbf{X}\mathbf{W}^T) + \delta \sum_{l=1}^{L} \sqrt{\|\boldsymbol{x}_l\|_2^2 + \|\boldsymbol{w}_l\|_2^2} \tag{3}$$

Remark 1: The (non-smooth) low-rank promoting term of (3), induces non-separability w.r.t. the columns of $\mathbf{X}$, $\mathbf{W}$*. This, allows for some of the* L *terms of the summation corresponding to the* ℓ_2 *norms of the coupled vectors* $\left[\begin{smallmatrix} \boldsymbol{x}_l \\ \boldsymbol{w}_l \end{smallmatrix}\right]$ *to be shrunk.*

Next we present a minimization algorithm for solving problem (3) using the square of the Frobenious norm as the distance metric function $l(\cdot)$.

III. Proposed Minimization Algorithm

It is easily observed that an exact alternating minimization of (3) w.r.t. columns $\boldsymbol{x}_l$ and $\boldsymbol{w}_l$ is infeasible due to the abovementioned non-separability of the proposed low-rank promoting term. Moreover, non-smoothness induces serious obstacles in the pursuit of stationary points. Taking into account the above restrictions and following the block successive minimization framework of [8], we alternatingly update blocks $\boldsymbol{x}_l$ and $\boldsymbol{w}_l$ by minimizing appropriately defined for $\boldsymbol{x}_l$'s and $\boldsymbol{w}_l$'s upper bound functions u_l's considering the remaining blocks fixed to their latest available (at iteration i) values. For instance u_ls corresponding to updates of $\boldsymbol{x}_l$'s (similar functions are defined for $\boldsymbol{w}_l$'s) are given below:

$$\begin{aligned} u_l(\boldsymbol{x}_l) =& \frac{1}{2} \|\mathbf{Y} - \boldsymbol{x}_l \boldsymbol{w}_l^{iT} - \mathbf{X}_{\neg l}^i \mathbf{W}_{\neg l}^{iT}\|_F^2 \\ &+ \frac{\delta}{2} \sum_{l=1}^{L} \left(\frac{\|\boldsymbol{x}_l\|_2^2 + \|\boldsymbol{w}_l^i\|_2^2 + \eta^2}{\sqrt{\|\boldsymbol{x}_l^i\|_2^2 + \|\boldsymbol{w}_l^i\|_2^2 + \eta^2}} + \sqrt{\|\boldsymbol{x}_l^i\|_2^2 + \|\boldsymbol{w}_l^i\|_2^2 + \eta^2} \right) \end{aligned} \tag{4}$$

with η a small constant (introduced to guarantee smoothness). The proposed algorithm is presented in Algorithm 1.

Proposition 1: The sequence of $\{\mathbf{X}^i, \mathbf{W}^i\}$ *generated by Algorithm 1 converges to a stationary point (local minimum) of the cost function defined in (3)* **Proof:** Can be proved utilizing Theorem 1 of [8].

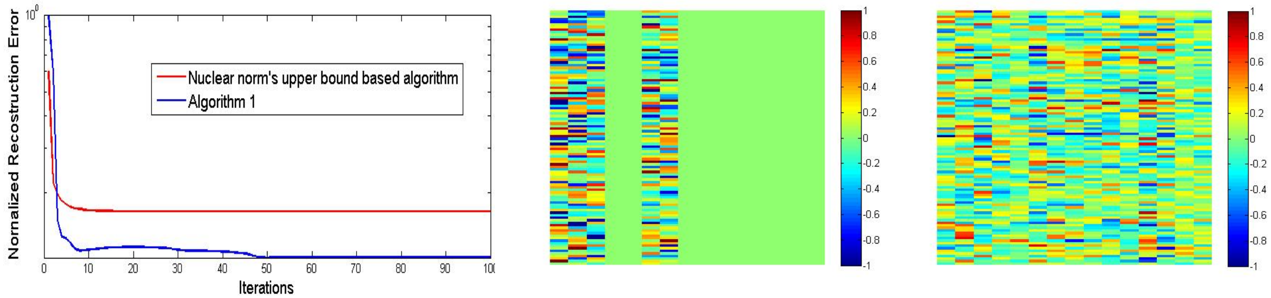
IV. Experiments

A. Synthetic Data

We randomly generate a subspace matrix $\mathbf{U} \in \mathcal{R}^{100 \times 5}$ and and coefficients matrix $\mathbf{V} \in \mathcal{R}^{500 \times 5}$ for producing data matrix $\mathbf{Y} = \mathbf{U}V^T$ which is contaminated by additive Gaussian i.i.d noise of $\sigma = 10^{-2}$. Performance of Algorithm 1 is compared to that of an alternating regularized least squares algorithm arising by the utilization of the nuclear norm's upper bound defined in (2). In the absence of knowledge of the true rank both the tested algorithms are initialized with $L = 15$. Fig. 1 shows the reconstruction error per iteration and the structure of the estimated subspace matrices.

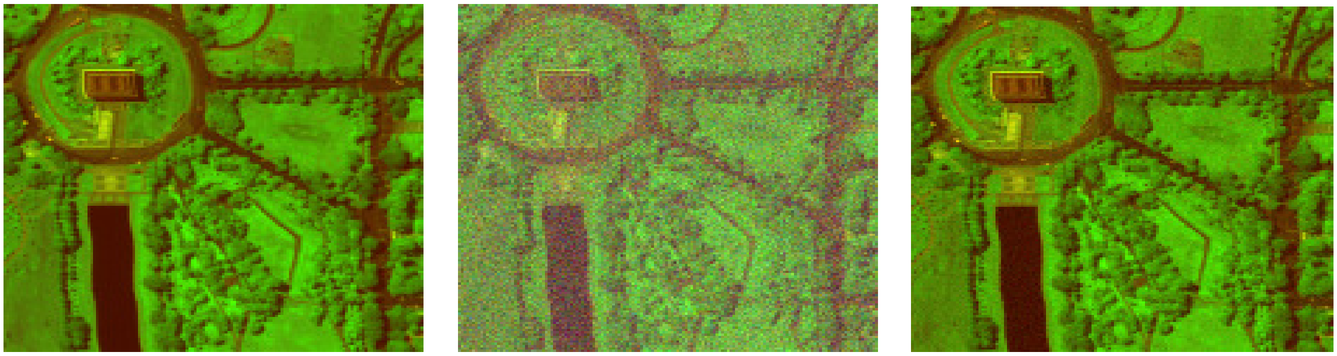
B. Real Hyperspectral dataset

The efficiency of the proposed algorithm is now tested in the denoising problem of a real hyperspectral dataset i.e. Washighton DC hyperspectral image (HSI) captured by HYDICE at 191 contiguous spectral bands. The region of interest consists of 150×150 pixels. The true image (Fig. 2a) is corrupted by additive Gaussian noise of $\sigma = 8 \times 10^{-2}$ resulting to the noisy version shown in Fig. 2b. The reconstructed HSI is given in Fig. 2c. The initial rank L is set to 45. In Fig. 3, the corresponding structural similarity indexes (SSIMs) per band obtained by the proposed Algorithm 1 and the nuclear norm's upper bound based algorithm are presented.



a) Normalized Reconstruction Error ($\frac{\|\mathbf{Y}_{true}-\hat{\mathbf{X}}\hat{\mathbf{W}}^T\|_F}{\|\mathbf{Y}_{true}\|_F}$) b) $\hat{\mathbf{W}}$, Algorithm 1 c) $\hat{\mathbf{W}}$, Nuclear norm's upper-bound algorithm

Fig. 1. Simulation results. It can be easily seen, the proposed algorithm achieves lower reconstruction error than that of the nuclear norm based relevant algorithm. From b) it is observed that Algorithm 1 not only imposes low-rankness, but also it converges to a basis of the true subspace. This is carried out by zeroing columns of $\mathbf{X}$ and $\mathbf{W}$ jointly.



a) true HSI b) noisy HSI, SNR = 9dB c) reconstructed HSI

Fig. 2. Results on Washighton DC HSI. False color images (bands 10,60 and 160). From a visual inspection of Fig. 2c, it is clear that the proposed algorithm is proven adept at recontructing the true image with high accuracy. This is attributed to the fact that it efficiently exploits the low-rank nature of the HSI via the novel ℓ_2/ℓ_1 norm based low-rank promoting term. It should be noted that both the estimated $\hat{\mathbf{X}}$ and $\hat{\mathbf{W}}$ consist of 4 nonzero columns, i.e, 41 columns have been zeroed.

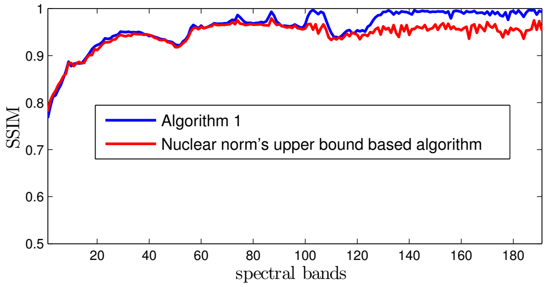


Fig. 3. SSIMs of the proposed Algorithm 1 and the nuclear norm's upper bound based algorithm. Algorithm 1 (blue line) outperforms (higher SSIM) the nuclear norm's upper bound based algorithm in several bands.

Algorithm 1 The proposed Alternating Iterative Reweighted Least Squares type algorithm

```
**Initialize** $\mathbf{W}^0, \mathbf{X}^0, \delta$
**for** $i = 1, 2, \ldots$
 $l = 1, 2, \ldots, L$
$\hat{\boldsymbol{x}}_l \equiv \left(\boldsymbol{w}_l^{i,T}\boldsymbol{w}_l^i + \frac{\delta}{\sqrt{\boldsymbol{x}_l^{T,i}\boldsymbol{x}_l^i + \boldsymbol{w}_l^{T,i}\boldsymbol{w}^il + \eta^2}}\right)^{-1} \left(\mathbf{Y} - \mathbf{X}_{\neg l}^i\mathbf{W}_{\neg l}^{i,T}\right)\boldsymbol{w}_l^i$
$\hat{\boldsymbol{w}}_l = \left(\boldsymbol{x}_l^{i,T}\boldsymbol{x}_l^i + \frac{\delta}{\sqrt{\boldsymbol{x}_l^{T,i}\boldsymbol{x}_l^i + \boldsymbol{w}_l^{T,i}\boldsymbol{w}^rl + \eta^2}}\right)^{-1} \left(\mathbf{Y} - \mathbf{X}_{\neg l}^i\mathbf{W}_{\neg l}^{i,T}\right)^T\boldsymbol{x}_l^i$
**end**
```

References

[1] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 4130–4137.

[2] F. Shang, Y. Liu, and J. Cheng, "Tractable and scalable schatten quasi-norm approximations for rank minimization," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Gretton and C. C. Robert, Eds., vol. 51. Cadiz, Spain: PMLR, 09–11 May 2016, pp. 620–629. [Online]. Available: http://proceedings.mlr.press/v51/shang16.html

[3] F. R. Bach, J. Mairal, and J. Ponce, "Convex sparse matrix factorizations," *CoRR*, vol. abs/0812.1869, 2008. [Online]. Available: http://arxiv.org/abs/0812.1869

[4] M. Mardani, G. Mateos, and G. B. Giannakis, "Subspace learning and imputation for streaming big data matrices and tensors," *IEEE Transactions on Signal Processing*, vol. 63, no. 10, pp. 2663–2677, May 2015.

[5] B. Haeffele, E. Young, and R. Vidal, "Structured low-rank matrix factorization: optimality, algorithm, and applications to image processing," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 2007–2015.

[6] F. Shang, Y. Liu, and J. Cheng, "Unified scalable equivalent formulations for schatten quasi-norms," *arXiv preprint arXiv:1606.00668*, 2016.

[7] V. Y. F. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the /spl beta/-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, July 2013.

[8] M. Hong, M. Razaviyayn, Z. Q. Luo, and J. S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: with applications in machine learning and signal processing," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 57–77, Jan 2016.