Alternating Iteratively Reweighted Least Squares Minimization for Low-Rank Matrix Factorization

Paris V. Giampouras, Athanasios A. Rontogiannis, *Member, IEEE*, and Konstantinos D. Koutroumbas

***Abstract*—Nowadays, the availability of large-scale data in disparate application domains urges the deployment of sophisticated tools for extracting valuable knowledge out of this huge bulk of information. In that vein, low-rank representations (LRRs), which seek low-dimensional embeddings of data have naturally appeared. In an effort to reduce computational complexity and improve estimation performance, LRR has been viewed via a matrix factorization (MF) perspective. Recently, low-rank MF (LRMF) approaches have been proposed for tackling the inherent weakness of MF, i.e., the unawareness of the dimension of the low-dimensional space where data reside. Herein, inspired by the merits of iterative reweighted schemes for sparse recovery and rank minimization, we come up with a generic low-rank promoting regularization function. Then, focusing on a specific instance of it, we propose a regularizer that imposes column-sparsity jointly on the two matrix factors that result from MF, thus promoting low-rankness on the optimization problem. The low-rank promoting properties of the resulting regularization term are brought to light by mathematically showing that it is actually a tight upper bound of a specific version of the weighted nuclear norm. The problems of denoising and matrix completion are redefined according to the new LRMF formulation and solved via efficient alternating iteratively reweighted least squares type algorithms. Theoretical analysis of the algorithms regarding the convergence and the rates of convergence to stationary points is provided. The effectiveness of the proposed algorithms is verified in diverse simulated and real data experiments.**

***Index Terms*—Matrix factorization, low-rank, iteratively reweighted, alternating minimization, matrix completion.**

I. Introduction

LOW-RANK representation (LRR) of data has recently attracted great interest since it appears in a wide spectrum of research fields and applications, such as signal processing, machine learning, quantum tomography, etc., [1]. LRR shares similar characteristics with sparse representation and hence is in principle formulated as a NP-hard problem, [2]. Convex relaxations have played a remarkable role in the course of making the problem tractable. In that respect, the nuclear norm has been extensively applied offering favorable results and a solid theoretical understanding, [3]. However, in the case of high-dimensional and large-scale datasets, conventional convex LRR approaches are confronted with inherent limitations related to their high computational complexity, [4].

To overcome these limitations matrix factorization (MF) methods have been introduced lately. MF gives rise to non-convex optimization problems and hence its theoretical understanding is a much more challenging task. Notably, a great effort has been recently devoted towards deriving a comprehensive theoretical framework of MF with the goal to reach to global optimality guarantees, [5]–[8]. MF presents significant computational merits by reducing the size of the emerging optimization problems. Thus, it leads to optimization algorithms of lower computational complexity as compared to relevant convex approaches. In addition, MF lies at the heart of a variety of problems dealing with the task of finding low-dimensional embeddings. In that respect, ubiquitous problems such as clustering, [9], blind source separation, matrix completion, [10], etc., have been seen in the literature through the lens of MF. MF entails the use of two matrix factors with a fixed number of columns, which, in the most favorable case, coincides with the rank of the sought matrix. However, the rank of the matrix, which is usually much less than its dimensions, is unknown a priori.

In light of this, a widespread approach is based on the following premise: overstate the number of columns of the matrix factors and then penalize their rank by using appropriate low-rank promoting regularizers. Along those lines, various regularizers have been recently proposed. Amongst them the most popular one is the variational characterization of the nuclear norm (proven to be a tight upper-bound of it) defined as the sum of the squared Frobenious norms of the factors [11]. More recently, generalized versions of this approach have come to the scene. In that respect, in [12]–[14], tight upper-bounds of the low-rank promoting Schatten-p quasinorms were presented under a general framework. In [15], an alternative approach for promoting low-rankness via non-convex MF was described. The novelty of that approach comes from the incorporation of additional constraints on the matrix factors giving thus rise to an interesting low-rank structured MF framework. In [4], a fast algorithm based on the above-mentioned variational characterization of the nuclear norm is presented. The derived algorithm is amenable to handling incomplete big-data, contrary to conventional convex and other non-MF based approaches. It should

Manuscript received April 3, 2018; revised August 31, 2018 and November 16, 2018; accepted November 17, 2018. Date of publication November 29, 2018; date of current version December 14, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yue Rong. This work was supported by the project PROTEAS II—Advanced Space Applications for Exploring the Universe of Space and Earth (MIS 5002515), which is implemented under the Action Reinforcement of the Research and Innovation Infrastructure, funded by the Operational Programme Competitiveness, Entrepreneurship and Innovation (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund). *(Corresponding author: Paris V. Giampouras.)*

The authors are with the Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Penteli 15236, Greece (e-mail: parisg@noa.gr; tronto@noa.gr; koutroum@space.noa.gr).

Digital Object Identifier 10.1109/TSP.2018.2883921

be noted that common characteristic of the low-rank MF methods mentioned above is the following: although the rank of the product of the matrix factors may decrease as a result of the penalization process, the number of columns of the matrix factors (which has initially been overstated) remains fixed throughout the execution of the minimization algorithms. Hence, the per iteration complexity remains unaltered, albeit the rank of the matrix factors may potentially decrease gradually to a large degree as the algorithms evolve.

In the current work, motivated by the latter (possibly undesirable in large-scale data applications) issue, we propose a novel generic formulation for non-convex low-rank MF. To this end, recent ideas stemming from iterative reweighted approaches for sparse recovery, [16] and low-rank matrix estimation, proposed in [17]–[19] as efficient alternatives for nuclear norm minimization, are now extended to the MF framework. This way, we come up with a novel alternating reweighted scheme for low-rank promotion in MF problems. As is shown, the recent low-rank MF schemes proposed in [12] can be cast as special occasions of the proposed formulation by suitably selecting the reweighting matrices applied on the matrix factors. Going one step further, we propose the selection of a common *diagonal* reweighting matrix that couples the matrix factors and leads to a joint column sparsity promoting regularization term, [20], [21]. In doing so, low-rank promotion now reduces to the task of jointly annihilating columns of the matrix factors. Interestingly, the resulting term is proven to be a tight upper bound of the *weighted nuclear norm* (upon appropriately selecting the weights), whose enhanced low-rank promoting properties have been recently reported in the literature, [19].

In an effort to better highlight the efficiency and ubiquity of the proposed low-rank MF formulation, we address two popular problems in the signal processing and machine learning literature, namely denoising and matrix completion. These problems are accordingly formulated in Section II. Then by exploiting the block successive upper bound minimization (BSUM) concept, [22], we minimize the arising non-smooth and non-separable objective functions in Section III. This is achieved by introducing appropriate upper bound functions for each subproblem related to the matrix factors, where minimization leads to closed-form analytical expressions thereof. In this regard, novel iteratively reweighted least squares (IRLS) type denoising and matrix completion algorithms are devised that rely exclusively on efficient matrix-wise updates. In addition, to further reduce complexity, we may incorporate a column pruning procedure that removes the matrix factor columns whose power has become negligible, thus reducing the size of the optimization problems towards that of the actual rank of the sought matrix. The connection of the proposed schemes with previously reported IRLS algorithms is established in Section IV. Analysis regarding the convergence of the algorithms to stationary points and their rates of convergence are given in Section V. In Section VI, the merits of the proposed algorithms in terms of estimation performance and computational complexity, compared to relevant state-of-art algorithms, are illustrated on simulated and real data experiments. In order to test the effectiveness of the algorithms on real applications involving large-scale data, the problems of hyperspectral image denoising and matrix completion in movies recommender systems are employed. Finally, Section VII concludes this work.

Notation: Matrices are represented as boldface uppercase letters, e.g., $\mathbf{X}$, and, column vectors as boldface lowercase letters, e.g., $\mathbf{x}$, while the i-th component of vector $\mathbf{x}$ is denoted by x_i and the ij-th element of matrix $\mathbf{X}$ by x_{ij}. Moreover, ^{T} denotes transposition, $\mathbf{I}_m$ is the $m \times m$ identity matrix and $\mathbf{0}$ is a zero matrix with respective dimensions, $\mathrm{rank}(\mathbf{X})$ is the rank of $\mathbf{X}$, $\mathrm{tr}\{\mathbf{X}\}$ denotes the trace of matrix $\mathbf{X}$, $\mathrm{diag}(\mathbf{x})$ is a diagonal matrix with the elements of vector $\mathbf{x}$ on its diagonal, $\boldsymbol{\sigma}(\mathbf{X})$ is the vector of the singular values of $\mathbf{X}$ arranged in a non-ascending order, $\|\cdot\|_p$ is the standard ℓ_p vector norm, $\|\mathbf{X}\|_* = \mathrm{tr}(\sqrt{\mathbf{X}^T\mathbf{X}}) = \sum_{i=1}^{rank(\mathbf{X})} \sigma_i(\mathbf{X})$, denotes the nuclear norm, $\|\mathbf{X}\|_{*,\mathbf{w}} = \sum_{i=1}^{rank(\mathbf{X})} w_i\sigma_i(\mathbf{X})$, is the weighted nuclear norm and $\|\mathbf{X}\|_{\mathcal{S}_p} = \|\boldsymbol{\sigma}(\mathbf{X})\|_p$ is the Schatten-p norm, $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j x_{ij}^2}$, stands for the Frobenius norm. $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. Also, $\mathcal{R}^{m\times n}$ stands for the $m \times n$-dimensional Euclidean space and $\otimes$ denotes the kronecker product operation.

II. Low-rank Matrix Factorization

Low-rank matrix estimation per se has been addressed by a wealth of different approaches, lending itself to disparate applications. Focusing on the task of recovering low-rank matrices from linear measurements, we come up with the ubiquitous affine rank minimization problem, [3], which is formulated as follows,

$$\min\left[\mathrm{rank}(\mathbf{X})\right] \quad \text{s.t.} \quad \mathcal{A}(\mathbf{X}) = \mathbf{b}, \tag{1}$$

where $\mathcal{A}$ denotes the linear operator that maps $\mathbf{X} \in \mathcal{R}^{m\times n}$ to $\mathbf{b} \in \mathcal{R}^l$. Problem (1) is tantamount to solving the ℓ_0 minimization problem on the singular values of $\mathbf{X}$ and hence is NP-hard. To this end various relaxation schemes have come to the scene in literature, many of which are based on the Schatten-p quasinorm [19], [23]. The Schatten-p quasinorm is defined as

$$\|\mathbf{X}\|_{\mathcal{S}_p} = \|\boldsymbol{\sigma}(\mathbf{X})\|_p, \tag{2}$$

with $0 < p \leq 1$. As is known, for $p = 1$, the Schatten-p quasinorm reduces to the well-known nuclear norm $\|\mathbf{X}\|_*$, which has been proven to be the convex envelope of the rank [2]. Schatten-p quasinorms have played a significant role in numerous cases involving the rank minimization problem of (1) reformulating it as

$$\min \|\mathbf{X}\|_{\mathcal{S}_p}^p \quad \text{s.t.} \quad \mathcal{A}(\mathbf{X}) = \mathbf{b}. \tag{3}$$

Nowadays, Schatten-p quasinorm based minimization has been seen via a more intriguing perspective i.e. using an iterative reweighting approach. In this vein, inspired by the IRLS method used in place of ℓ_1 norm minimization for imposing sparsity, [16], in [17] and [18] the authors propose to minimize a *reweighted* Frobenious norm. The equivalence of the Schatten-p quasinorm and those minimized in [17], [18], is mathematically

expressed as follows,

$$\|\mathbf{X}\|_{\mathcal{S}_p}^p = \mathrm{tr}\{ \left(\mathbf{X}^T\mathbf{X}\right)^{p/2} \} = \mathrm{tr}\{ \left(\mathbf{X}^T\mathbf{X}\right) \left(\mathbf{X}^T\mathbf{X}\right)^{p/2-1} \}$$
$$= \mathrm{tr}\{ \left(\mathbf{X}^T\mathbf{X}\right) \mathbf{W} \} = \|\mathbf{X}\mathbf{W}^{\frac{1}{2}}\|_F^2, \quad (4)$$

where $\mathbf{W}$ is the symmetric weight matrix

$$\mathbf{W} = \left(\mathbf{X}^T\mathbf{X}\right)^{p/2-1}. \quad (5)$$

It should be noted that reweighted Frobenious norm schemes are iterative and, in each iteration the weight matrix $\mathbf{W}$ is computed from the estimate of $\mathbf{X}$ obtained in the previous iteration.

More recently, low-rank matrix estimation has also been addressed by using weighted and reweighted versions of the nuclear norm, [19]. In this regard the rank minimization problem is formulated as follows,

$$\min \|\mathbf{X}\|_{*,\mathbf{w}} \quad \text{s.t.} \quad \mathcal{A}(\mathbf{X}) = \mathbf{b}. \quad (6)$$

As shown in [19], by suitably selecting the weights w_i s in vector $\mathbf{w}$, (6) offers a generic framework for efficiently tackling various rank minimization tasks, including the Schatten-p norm minimization problem defined in (3).

It should be noted that both the reweighted Frobenius and the (re)weighted nuclear norm based schemes have been shown to offer significant merits in terms of computational complexity, estimation performance and rate of convergence.

Recently, low-rank matrix estimation has been effectively tackled using a *matrix factorization* approach. The crux of the relevant methods is that a low-rank matrix can be well represented by a product of two matrices $\mathbf{U}$ $(m \times r)$ and $\mathbf{V}$ $(n \times r)$ i.e., $\mathbf{X} = \mathbf{U}\mathbf{V}^T$ with the inner dimension r of the involved matrices quite smaller than the outer dimensions i.e., $r \ll \min(m, n)$. Needless to say that those ideas offer significant advantages when it comes to the processing of large scale and high-dimensional datasets (where both m and n are huge) by reducing the size of the involved variables, thus decreasing both the storage space required from $\mathcal{O}(mn)$ to $\mathcal{O}\left((m+n)r\right)$ as well as the computational complexity of the algorithms used. However, a downside of this approach is that an additional variable is brought up i.e., the inner dimension r of the factorization. The task of finding the actual r (which coincides with the rank of matrix $\mathbf{X}$) is relevant to the rank minimization problem and is referred in the literature also as dimensionality reduction, model order selection, etc.

The latter has given rise to methods that select r based on the minimization of various criteria such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the minimum distance length (MDL), [24], etc. However, these methods can be computationally expensive especially in large scale datasets, since they require multiple runs using different values for r. Modern approaches termed low-rank matrix factorization (LRMF) techniques, [15], hinge on the following philosophy: a) overstate the rank r of the product with $d \geq r$ and then b) impose low-rankness thereof by utilizing appropriate norms. This rationale has given rise to LRMF techniques that solve the following,

$$\min \left[\mathrm{rank}(\mathbf{U}\mathbf{V}^T)\right] \quad \text{s.t.} \quad \mathcal{A}(\mathbf{U}\mathbf{V}^T) = \mathbf{b}. \quad (7)$$

Problem (7) has been addressed by different ways in the literature. Among other approaches, the tight upper-bound of the nuclear norm defined as

$$\|\mathbf{X}\|_* = \min_{\mathbf{U}\in\mathcal{R}^{m\times d}, \mathbf{V}\in\mathcal{R}^{n\times d}, \mathbf{X}=\mathbf{U}\mathbf{V}^T} \|\mathbf{U}\|_F \|\mathbf{V}\|_F$$
$$= \min_{\mathbf{U}\in\mathcal{R}^{m\times d}, \mathbf{V}\in\mathcal{R}^{n\times d}, \mathbf{X}=\mathbf{U}\mathbf{V}^T} \frac{1}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \quad (8)$$

is the most popular, [11]. In fact, minimization of (8) favors low-rankness on $\mathbf{U}$ and $\mathbf{V}$ by inducing smoothness on these matrices. In [12] and [25], the authors derive the tight upper-bounds for all Schatten-p quasinorms with $0 < p \leq 1$, ([25, Th. 1]) i.e.,

$$\|\mathbf{X}\|_{\mathcal{S}_p}^p = \min_{\mathbf{U}\in\mathcal{R}^{m\times d}, \mathbf{V}\in\mathcal{R}^{n\times d}, \mathbf{X}=\mathbf{U}\mathbf{V}^T} \|\mathbf{U}\|_{\mathcal{S}_{2p}}^p \|\mathbf{V}\|_{\mathcal{S}_{2p}}^p$$
$$= \min_{\mathbf{U}\in\mathcal{R}^{m\times d}, \mathbf{V}\in\mathcal{R}^{n\times d}, \mathbf{X}=\mathbf{U}\mathbf{V}^T} \frac{1}{2}\left(\|\mathbf{U}\|_{\mathcal{S}_{2p}}^{2p} + \|\mathbf{V}\|_{\mathcal{S}_{2p}}^{2p}\right). \quad (9)$$

Common denominator of the afore-mentioned low-rank matrix factorization approaches is their direct connection with the low-rank imposing Schatten-p quasinorms, since they provide tight upper-bounds thereof.

In this work we aspire to apply ideas stemming from *iterative reweighting methods for low-rank matrix recovery*, to this challenging low-rank matrix factorization scenario. Therefore, generalizing the above-described low-rank promoting norm upper bounds, we propose to minimize the sum of reweighted (as in (4)) Frobenious norms of the individual factors $\mathbf{U}$ and $\mathbf{V}$. Hence, the newly introduced low-rank inducing function is defined as follows,

$$h(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\left(\|\mathbf{U}\mathbf{W}_{\mathbf{U}}^{\frac{1}{2}}\|_F^2 + \|\mathbf{V}\mathbf{W}_{\mathbf{V}}^{\frac{1}{2}}\|_F^2\right) \quad (10)$$

where the weight matrices $\mathbf{W}_{\mathbf{U}}$ and $\mathbf{W}_{\mathbf{V}}$ are appropriately selected. The proposed low-rank promoting function defined in (10) is generic as it includes the previously mentioned MF-based low-rank promoting terms as special cases. Indeed, according to (4), (5) and by setting $\mathbf{W}_{\mathbf{U}} = (\mathbf{U}^T\mathbf{U})^{p-1}$ and $\mathbf{W}_{\mathbf{V}} = (\mathbf{V}^T\mathbf{V})^{p-1}$ in (10), we get the upper-bound of the Schatten-p quasinorm given in (9), while for $p = 1$, i.e., $\mathbf{W}_{\mathbf{U}} = \mathbf{W}_{\mathbf{V}} = \mathbf{I}_d$, we get the variational form of the nuclear norm defined in (8).

In the rest of this paper, we adhere to a special instance of (10) which arises by setting $\mathbf{W}_{\mathbf{U}} = \mathbf{W}_{\mathbf{V}} = \mathbf{W}$ with

$$\mathbf{W} = \mathrm{diag}\Big(\left(\|\boldsymbol{u}_1\|_2^2 + \|\boldsymbol{v}_1\|_2^2\right)^{p/2-1}, \left(\|\boldsymbol{u}_2\|_2^2 + \|\boldsymbol{v}_2\|_2^2\right)^{p/2-1},$$
$$\ldots, \left(\|\boldsymbol{u}_d\|_2^2 + \|\boldsymbol{v}_d\|_2^2\right)^{p/2-1} \Big), \quad (11)$$

where $0 < p \leq 1$ and $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$, respectively.[1] The selection of the common *diagonal* weight matrix of the factors as in (11) is not arbitrary. As we will see in Sections III and IV, this matrix leads to IRLS schemes for low-rank matrix factorization, generalizing the IRLS-p family

[1] If $\mathbf{U}, \mathbf{V}$ had orthogonal columns, $\mathbf{W}$ in (11) would be equal to $(\mathbf{U}^T\mathbf{U} + \mathbf{V}^T\mathbf{V})^{p/2-1}$, whose resemblance to (5) is evident.

of algorithms developed in [16] for sparse vector recovery. In addition by selecting a common $\mathbf{W}$ for $\mathbf{U}$ and $\mathbf{V}$, matrices $\mathbf{U}$ and $\mathbf{V}$ are implicitly coupled w.r.t. their columns. If we now substitute (11) in (10) yields

$$h(\mathbf{U},\mathbf{V}) = \frac{1}{2}\sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2}. \tag{12}$$

Surprisingly, the resulting expression coincides with the (scaled by 1/2) group sparsity inducing $\ell_{p,2}^p$ norm ($0 < p \leq 1$) of the concatenated matrix $\left[\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right]$, which for $p = 1$ reduces to the commonly used $\ell_{1,2}$ matrix norm. Intuitively, the low-rank inducing properties of the proposed in (12) joint column sparsity promoting term can be easily explained as follows. Let us consider the rank one decomposition of the matrix product $\mathbf{UV}^T$,

$$\mathbf{UV}^T = \sum_{i=1}^{d}\boldsymbol{u}_i\boldsymbol{v}_i^T. \tag{13}$$

Clearly, due to the subadditivity property of the rank, eliminating rank one terms of the summation on the right side of (13) results to a relevant decrease of the rank of the product $\mathbf{UV}^T$. Besides the previous intuitive explanation, a more rigorous theoretical justification of the low-rank promoting properties of the proposed regularizer is provided in the following Proposition.

Proposition 1: Let $\mathbf{X} = \mathbf{L}\boldsymbol{\Sigma}\mathbf{R}^T$ be the singular value decomposition of matrix $\mathbf{X} \in \mathcal{R}^{m\times n}$, where $\mathbf{L} \in \mathcal{R}^{m\times d}$, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$, with σ_is being the singular values of $\mathbf{X}$ arranged in non-increasing order and $\mathbf{R} \in \mathcal{R}^{n\times d}$. Let also $\mathbf{X} = \mathbf{UV}^T$ be an arbitrary decomposition of $\mathbf{X}$, where $\mathbf{U} \in \mathcal{R}^{m\times d}$ and $\mathbf{V} \in \mathcal{R}^{n\times d}$. The proposed regularizer defined in (12) is a tight upper bound of the weighted nuclear norm of $\mathbf{X}$, i.e.,

$$\|\mathbf{X}\|_{*,\mathbf{w}} \leq \frac{1}{2}\sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2} \tag{14}$$

where $\mathbf{w}$ contains the diagonal elements of $\mathbf{W}$ defined in (11) arranged in a non-decreasing order.

Proof: See Appendix.

Note that by Proposition 1 it becomes clear that the proposed low-rank promoting term deviates from the previous relevant regularizers given in (8) and (9), since it introduces a tight upper bound of the recently proposed weighted and reweighted variants of the nuclear norm. It should be also noted that the idea of imposing jointly column sparsity first appeared in [26], albeit in a Bayesian framework tailored to the NMF problem. In [27], the emerging via the maximum a posteriori probability (MAP) approach optimization problem boils down to the minimization of the column sparsity promoting concave logarithm function. Hence capitalizing on (12), we are led to LRMF optimization problems having the form,

$$\min_{\mathbf{U}\in\mathcal{R}^{m\times d},\mathbf{V}\in\mathcal{R}^{n\times d}} \sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2} \quad \text{s.t. } \mathcal{A}(\mathbf{UV}^T) = \mathbf{b}. \tag{15}$$

Next, the generic problem given in (15) is reformulated and solved for two important learning tasks namely a) denoising and b) matrix completion.

A. Denoising

By assuming that a) the linear operator $\mathcal{A}$ reduces to a diagonal matrix and b) our measurements $\mathbf{Y} \in \mathcal{R}^{m\times n}$ are corrupted by i.i.d. Gaussian noise, we come up with the following optimization problem,

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2} \quad \text{s.t. } \|\mathbf{Y} - \mathbf{UV}^T\|_F^2 \leq \epsilon. \tag{16}$$

where ϵ is a small positive constant. By Lagrange theorem we know that (16) can be equivalently written in the following form,

$$\{\hat{\mathbf{U}},\hat{\mathbf{V}}\} = \underset{\mathbf{U},\mathbf{V}}{\mathrm{argmin}} \frac{1}{2}\|\mathbf{Y} - \mathbf{UV}^T\|_F^2 + \lambda\sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2} \tag{17}$$

where λ denotes the Lagrange multiplier.

B. Matrix Completion

Another popular problem that follows the general model described by (15) is matrix completion, as it is widely addressed via low-rank minimization. The main premise here lies in recovering missing entries of a matrix $\mathbf{Y}$ assuming high coherence among its elements, which gives rise to a low-rank structured matrix $\mathbf{X}$. The problem is thus set up as,

$$\min[\mathrm{rank}(\mathbf{X})] \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{Y}) = \mathcal{P}_\Omega(\mathbf{X}), \tag{18}$$

where $\mathcal{P}_\Omega$ denotes the sampling operator on the set Ω of indexes of matrix $\mathbf{Y}$ where information is present. In the matrix factorization setting, the incomplete matrix $\mathbf{Y}$ is approximated by a matrix $\mathbf{X}$ expressed as $\mathbf{X} = \mathbf{UV}^T$. As mentioned above, the rank r of the reconstructed matrix $\mathbf{X}$ is generally unknown and hence it is overstated with $d \geq r$. This necessitates the penalization of the rank of the product $\mathbf{UV}^T$, which in our case takes place with the proposed low-rank promoting term giving rise to the optimization problem,

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2} \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{Y}) = \mathcal{P}_\Omega(\mathbf{UV}^T). \tag{19}$$

Considering further the existence of additive i.i.d. Gaussian noise in $\mathbf{Y}$ we get,

$$\{\hat{\mathbf{U}},\hat{\mathbf{V}}\} = \underset{\mathbf{U},\mathbf{V}}{\mathrm{argmin}} \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y}) - \mathcal{P}_\Omega(\mathbf{UV}^T)\|_F^2 + \lambda\sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2}. \tag{20}$$

As it is shown later, the simplicity and tractability of the proposed regularizer facilitates the derivation of new optimization algorithms, while the adoption of the minimization framework presented in the next section paves the way for the theoretical analysis of their convergence behavior.

III. Minimization Algorithms

Herein, we present two new efficient block coordinate minimization (BCM) algorithms for denoising and matrix completion, respectively. The alternating minimization, w.r.t. the 'blocks' $\mathbf{U}$ and $\mathbf{V}$, of the proposed low-rank promoting function defined in (12) lies at the heart of those algorithms.

Remark 1: The proposed low-rank promoting regularizer is a) non-smooth and b) non-separable w.r.t. $\mathbf{U}$ and $\mathbf{V}$.

Both the above-mentioned properties i.e., non-smoothness and non-separability induce severe difficulties in the optimization task that call for appropriate handling. More specifically, as it has been shown, [28], in BCM schemes the respective algorithms might be led to irregular points i.e., coordinate-wise minima that are not necessarily stationary points of the minimized objective function. In light of this we follow a simple smoothing approach by including a small positive constant η^2 in the proposed regularizer, which becomes,

$$\hat{h}(\mathbf{U},\mathbf{V})=\sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2+\|\boldsymbol{v}_i\|_2^2+\eta^2)^{p/2}. \tag{21}$$

This way we alleviate singular points i.e., points where the gradient is not continuous, and the resulting optimization problems become smooth. On the other hand, non-separability poses obstacles in getting closed-form expressions for the optimization variables $\mathbf{U}$ and $\mathbf{V}$. For this reason, each of the associative optimization problems is reformulated using appropriate relaxation schemes. By working in an alternating fashion, each of these schemes results in closed form expressions. Next, the proposed algorithms that solve the denoising and matrix completion problems are analytically described.

A. Denoising

In this section, we present a new algorithm designed for solving the denoising problem given in (17). To this end, let us first define the respective objective function as,

$$f(\mathbf{U},\mathbf{V})=\frac{1}{2}\|\mathbf{Y}-\mathbf{U}\mathbf{V}^T\|_F^2 + \lambda\sum_{i=1}^{d}(\|\boldsymbol{u}_i\|_2^2+\|\boldsymbol{v}_i\|_2^2+\eta^2)^{p/2}. \tag{22}$$

It is obvious that minimizing (22) alternatingly w.r.t. $\mathbf{U}$ and $\mathbf{V}$ is infeasible, since exact analytical expressions can not be obtained as a result of the non-separable nature of the regularizing term. To this end, at each iteration $k+1$ we solve two distinct subproblems i.e. a) given the latest available update $\mathbf{V}_k$ of $\mathbf{V}$, we minimize an approximate cost function w.r.t. $\mathbf{U}$ to get $\mathbf{U}_{k+1}$ and b) we use $\mathbf{U}_{k+1}$ in order to minimize another approximate cost function w.r.t. the second block variable of our problem i.e., matrix $\mathbf{V}$. Following the block successive upper-bound minimization (BSUM) philosophy, [22], [29], we minimize at each iteration local tight upper-bounds of the respective objective functions. That said, $\mathbf{U}$ is updated by minimizing an approximate second-order Taylor expansion of $f(\mathbf{U},\mathbf{V}_k)$ around the point $(\mathbf{U}_k,\mathbf{V}_k)$. Likewise, an approximate second-order Taylor expansion of $f(\mathbf{U}_{k+1},\mathbf{V})$ around $(\mathbf{U}_{k+1},\mathbf{V}_k)$ is utilized for obtaining $\mathbf{V}_{k+1}$. To be more specific $\mathbf{U}_{k+1}$ is computed by

$$\mathbf{U}_{k+1}=\underset{\mathbf{U}}{\operatorname{argmin}}\; l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k), \tag{23}$$

where,

$$l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)=f(\mathbf{U}_k,\mathbf{V}_k)+\operatorname{tr}\{(\mathbf{U}-\mathbf{U}_k)^T\nabla_{\mathbf{U}}f(\mathbf{U}_k,\mathbf{V}_k)\} + \frac{1}{2}\operatorname{vec}(\mathbf{U}-\mathbf{U}_k)^T\bar{\mathbf{H}}_{\mathbf{U}_k}\operatorname{vec}(\mathbf{U}-\mathbf{U}_k) \tag{24}$$

and $\operatorname{vec}(\cdot)$ denotes the row vectorization operator. In (24), the true Hessian $\mathbf{H}_{\mathbf{U}_k}$ of $f(\mathbf{U},\mathbf{V}_k)$ at $\mathbf{U}_k$ has been approximated by the $md\times md$ positive-definite block diagonal matrix $\bar{\mathbf{H}}_{\mathbf{U}_k}$, which is expressed as

$$\bar{\mathbf{H}}_{\mathbf{U}_k}=\mathbf{I}_m\otimes\tilde{\mathbf{H}}_{\mathbf{U}_k}, \tag{25}$$

where $\otimes$ denotes the Kronecker product operation. For reasons that will be explained later, the $d\times d$ diagonal block $\tilde{\mathbf{H}}_{\mathbf{U}_k}$ is defined as

$$\tilde{\mathbf{H}}_{\mathbf{U}_k}=\mathbf{V}_k^T\mathbf{V}_k+\lambda\mathbf{D}_{(\mathbf{U}_k,\mathbf{V}_k)} \tag{26}$$

with

$$\mathbf{D}_{(\mathbf{U},\mathbf{V})}=p\operatorname{diag}\Big((\|\boldsymbol{u}_1\|_2^2+\|\boldsymbol{v}_1\|_2^2+\eta^2)^{p/2-1}, (\|\boldsymbol{u}_2\|_2^2+\|\boldsymbol{v}_2\|_2^2+\eta^2)^{p/2-1},\ldots, (\|\boldsymbol{u}_d\|_2^2+\|\boldsymbol{v}_d\|_2^2+\eta^2)^{p/2-1}\Big). \tag{27}$$

As it is shown in Section V and the Appendix, due to the form of $\bar{\mathbf{H}}_{\mathbf{U}_k}$ from (25) and (26) and its relation to the exact Hessian $\mathbf{H}_{\mathbf{U}_k}$ of $f(\mathbf{U},\mathbf{V}_k)$ at $\mathbf{U}_k$, $l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)$ bounds $f(\mathbf{U},\mathbf{V}_k)$ from above and hence the conditions set by the BSUM framework are satisfied. Actually, the approximation of the exact Hessian by using (25) leads to a closed-from expression for updating $\mathbf{U}$ and a dramatic decrease of the required computational complexity, as it will be further explained below.

Following a similar path as above we come up with appropriate upper-bound functions for updating $\mathbf{V}$ i.e,

$$\mathbf{V}_{k+1}=\underset{\mathbf{V}}{\operatorname{argmin}}\; g(\mathbf{V}|\mathbf{U}_{k+1},\mathbf{V}_k) \tag{28}$$

with

$$g(\mathbf{V}|\mathbf{U}_{k+1},\mathbf{V}_k)=f(\mathbf{U}_{k+1},\mathbf{V}_k) + \operatorname{tr}\{(\mathbf{V}-\mathbf{V}_k)^T\nabla_{\mathbf{V}}f(\mathbf{U}_{k+1},\mathbf{V}_k)\} + \frac{1}{2}\operatorname{vec}(\mathbf{V}-\mathbf{V}_k)^T\bar{\mathbf{H}}_{\mathbf{V}_k}\operatorname{vec}(\mathbf{V}-\mathbf{V}_k) \tag{29}$$

Algorithm 1: Alternating Iteratively Reweighted Least Squares (AIRLS) Denoising Algorithm.

Input: $\mathbf{Y}, \lambda > 0$
Initialize: $k = 0, \mathbf{V}_0, \mathbf{U}_0, \mathbf{D}_{(\mathbf{U}_0, \mathbf{V}_0)}$
repeat
$\mathbf{U}_{k+1} = \mathbf{Y}\mathbf{V}_k \left(\mathbf{V}_k^T \mathbf{V}_k + \lambda \mathbf{D}_{(\mathbf{U}_k, \mathbf{V}_k)}\right)^{-1}$
$\mathbf{V}_{k+1} = \mathbf{Y}^T \mathbf{U}_{k+1} \left(\mathbf{U}_{k+1}^T \mathbf{U}_{k+1} + \lambda \mathbf{D}_{(\mathbf{U}_{k+1}, \mathbf{V}_k)}\right)^{-1}$
$k = k + 1$
until *convergence*
Output: $\hat{\mathbf{U}} = \mathbf{U}_{k+1}, \hat{\mathbf{V}} = \mathbf{V}_{k+1}$

and $\bar{\mathbf{H}}_{\mathbf{V}_k}$ being a block diagonal $md \times md$ matrix (similar to $\bar{\mathbf{H}}_{\mathbf{U}_k}$) whose $d \times d$ diagonal blocks $\tilde{\mathbf{H}}_{\mathbf{V}_k}$ are defined as

$$\tilde{\mathbf{H}}_{\mathbf{V}_k} = \mathbf{U}_{k+1}^T \mathbf{U}_{k+1} + \lambda \mathbf{D}_{(\mathbf{U}_{k+1}, \mathbf{V}_k)}. \tag{30}$$

By solving (23) and (28) we obtain analytical expressions for $\mathbf{U}_{k+1}$ and $\mathbf{V}_{k+1}$ that constitute the main steps of the proposed denoising algorithm given in Algorithm 1. As explained in Section IV, Algorithm 1 is an alternating IRLS (AIRLS) algorithm for low rank matrix factorization applied to data denoising.

B. Matrix Completion

Next the matrix completion problem, under the matrix factorization setting stated in (20), is addressed. As mentioned earlier, matrix factorization offers scalability making the derived algorithms amenable to processing big and high dimensional data. It should be emphasized that in the proposed formulation of the problem (20), the impediments arising by the low-rank promoting term (Remark 1) are now complemented by the difficulty to get computationally efficient *matrix-wise* updates for $\mathbf{U}$ and $\mathbf{V}$, due to the presence of the sampling operator $\mathcal{P}_\Omega$ in the data fitting term. That said, the objective function is now modified as

$$f(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathcal{P}_\Omega \left(\mathbf{Y} - \mathbf{U}\mathbf{V}^T\right)\|_F^2 + \lambda \sum_{i=1}^{d} (\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2 + \eta^2)^{p/2}. \tag{31}$$

As in the denoising problem, we minimize quadratic upper-bound functions based on approximate second-order Taylor expansions. In this respect, in order to get closed-form analytical expressions for $\mathbf{U}_{k+1}$ and $\mathbf{V}_{k+1}$ that involve exclusively matrix operations, we select again the upper bound functions $l(\mathbf{U}|\mathbf{U}_k, \mathbf{V}_k)$ and $g(\mathbf{V}|\mathbf{U}_{k+1}, \mathbf{V}_k)$ defined in (24) and (29), with $\bar{\mathbf{H}}_{\mathbf{U}_k}$ and $\bar{\mathbf{H}}_{\mathbf{V}_k}$ as given before, but $f(\mathbf{U}, \mathbf{V})$ now defined as in (31). The resulting efficient matrix-wise update formulas are shown in Algorithm 2, where the new AIRLS matrix completion algorithm (AIRLS-MC) is presented.

Remark 2: For $\lambda > 0$, approximation matrices $\bar{\mathbf{H}}_{\mathbf{U}_k}$ and $\bar{\mathbf{H}}_{\mathbf{V}_k}$ are always positive definite and hence invertible. In other words, both $l(\mathbf{U}|\mathbf{U}_k, \mathbf{V}_k)$ and $g(\mathbf{V}|\mathbf{U}_{k+1}, \mathbf{V}_k)$ are strictly convex and hence have unique minimizers. In addition, since approximations of the exact Hessians are used in the two block problems, we end up with quasi-Newton type update formulas for $\mathbf{U}$ and $\mathbf{V}$.

Algorithm 2: AIRLS Matrix Completion (AIRLS-MC) Algorithm.

Input: $\mathbf{Y}, \lambda > 0$
Initialize: $k = 0, \mathbf{U}_0, \mathbf{V}_0, \mathbf{D}_{(\mathbf{U}_0, \mathbf{V}_0)}$
repeat
$\mathbf{U}_{k+1} = \mathbf{U}_k - \left(\mathcal{P}_\Omega \left(\mathbf{U}_k \mathbf{V}_k^T - \mathbf{Y}\right) \mathbf{V}_k + \lambda \mathbf{U}_k \mathbf{D}_{(\mathbf{U}_k, \mathbf{V}_k)}\right) \left(\mathbf{V}_k^T \mathbf{V}_k + \lambda \mathbf{D}_{(\mathbf{U}_k, \mathbf{V}_k)}\right)^{-1}$
$\mathbf{V}_{k+1} = \mathbf{V}_k - \left(\mathcal{P}_\Omega \left(\mathbf{V}_k \mathbf{U}_{k+1}^T - \mathbf{Y}^T\right) \mathbf{U}_{k+1} + \lambda \mathbf{V}_k \mathbf{D}_{(\mathbf{U}_{k+1}, \mathbf{V}_k)}\right) \left(\mathbf{U}_{k+1}^T \mathbf{U}_{k+1} + \lambda \mathbf{D}_{(\mathbf{U}_{k+1}, \mathbf{V}_k)}\right)^{-1}$
$k = k + 1$
until *convergence*
Output: $\hat{\mathbf{U}} = \mathbf{U}_{k+1}, \hat{\mathbf{V}} = \mathbf{V}_{k+1}$

Remark 3: The gain of using matrices $\bar{\mathbf{H}}_{\mathbf{U}_k}$ and $\bar{\mathbf{H}}_{\mathbf{V}_k}$ in the approximation of the exact Hessians of $f(\mathbf{U}, \mathbf{V})$ (given either by (22) or (31)) w.r.t. $\mathbf{U}$ and $\mathbf{V}$ is twofold. Not only (as proven in the Appendix) we remain in the BSUM framework, which offers favorable theoretical properties, but also we are able to update $\mathbf{U}$ and $\mathbf{V}$ at a very low computational cost. As it can be noticed in Algorithms 1 and 2, the inversions of $\bar{\mathbf{H}}_{\mathbf{U}_k}$ and $\bar{\mathbf{H}}_{\mathbf{V}_k}$ involved in the updates of $\mathbf{U}$ and $\mathbf{V}$ reduce to the inversions of the $d \times d$ matrices $\tilde{\mathbf{H}}_{\mathbf{U}_k}$ and $\tilde{\mathbf{H}}_{\mathbf{V}_k}$ thus inducing complexity in the order of $\mathcal{O}(d^3)$. Contrary, utilization of the exact Hessians w.r.t. $\mathbf{U}$ and $\mathbf{V}$ would have given rise to inversions with much higher computational complexity i.e., $\mathcal{O}(\max(m, n) \times d^3)$.

IV. Relation to Prior Art

The proposed algorithms belong to the family of iteratively reweighted least squares minimization algorithms, which date back to the 1930's [30]. Recently, the IRLS method has been adopted for sparse vector recovery in [16], leading to an iterative algorithm that solves the following minimization problem at the $(k+1)$th iteration

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \sum_{i=1}^{m} w_i^k x_i^2 \quad \text{s.t.} \quad \mathcal{A}(\mathbf{x}) = \mathbf{b}, \tag{32}$$

where the sparse vector $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T \in \mathcal{R}^{m \times 1}$ and $w_i^k = (|x_i^k|^2 + \eta^2)^{p/2-1}$. Theoretical guarantees for sparse signal recovery have been provided in [16] for $p = 1$. To generalize, the minimization problem in (32) can be extended to promote structured (group) sparsity as follows

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \sum_{i=1}^{d} w_i^k ||\mathbf{x}_i||_2^2 \quad \text{s.t.} \quad \mathcal{A}(\mathbf{x}) = \mathbf{b}, \tag{33}$$

where now $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_d^T]^T$ is structured in d groups and $w_i^k = (||\mathbf{x}_i^k||_2^2 + \eta^2)^{p/2-1}$.

More recently, the same idea has been applied for low-rank matrix recovery in [18]. In this vein the minimization problem

is properly adjusted as,

$$\mathbf{X}_{k+1} = \arg\min_{\mathbf{X}} \text{tr}(\mathbf{W}_k \mathbf{X}^T \mathbf{X}) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{X}) = \mathbf{b}, \tag{34}$$

and $\mathbf{W}_k = (\mathbf{X}_k^T \mathbf{X}_k + \eta^2 \mathbf{I}_n)^{p/2-1}$. As explained in Section II and eq. (4), this problem is equivalent to minimizing the Schatten-p quasinorm of $\mathbf{X}$, thus promoting low-rank solutions.

To place our method in the above described framework, we rewrite our generic optimization problem, given in (15), as follows,

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{d} (\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)^{p/2-1} (\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)$$
$$\text{s.t. } \mathcal{A}(\mathbf{U}\mathbf{V}^T) = \mathbf{b}. \tag{35}$$

Then, from (35) we can define the following IRLS minimization scheme

$$\{\mathbf{U}_{k+1}, \mathbf{V}_{k+1}\} = \underset{\mathbf{U},\mathbf{V}}{\text{argmin}} \sum_{i=1}^{d} w_i^k (\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2)$$
$$\text{s.t. } \mathcal{A}(\mathbf{U}\mathbf{V}^T) = \mathbf{b}, \tag{36}$$

where $w_i^k = (\|\boldsymbol{u}_i^k\|_2^2 + \|\boldsymbol{v}_i^k\|_2^2 + \eta^2)^{p/2-1}$. This optimization task can be solved alternatingly with respect to $\mathbf{U}$ and $\mathbf{V}$ as follows,

$$\mathbf{U}_{k+1} = \underset{\mathbf{U}}{\text{argmin}} \sum_{i=1}^{d} w_i^{k,k} \|\boldsymbol{u}_i\|_2^2 \text{ s.t. } \mathcal{A}(\mathbf{U}\mathbf{V}_k^T) = \mathbf{b}, \tag{37}$$

$$\mathbf{V}_{k+1} = \underset{\mathbf{V}}{\text{argmin}} \sum_{i=1}^{d} w_i^{k+1,k} \|\boldsymbol{v}_i\|_2^2 \text{ s.t. } \mathcal{A}(\mathbf{U}_{k+1}\mathbf{V}^T) = \mathbf{b}, \tag{38}$$

where $w_i^{k,k} = (\|\boldsymbol{u}_i^k\|_2^2 + \|\boldsymbol{v}_i^k\|_2^2 + \eta^2)^{p/2-1}$ and $w_i^{k+1,k} = (\|\boldsymbol{v}_i^k\|_2^2 + \|\boldsymbol{u}_i^{k+1}\|_2^2 + \eta^2)^{p/2-1}$. It can be shown that if we consider a LS data fitting term in our objective function, the solution of the IRLS schemes (37) and (38) leads to the same exact expressions for $\mathbf{U}_{k+1}$ and $\mathbf{V}_{k+1}$ as those obtained for AIRLS in the previous section. Note that (37) and (38) hold close resemblance with the minimization problem (33) via the correspondence of the block vectors $\mathbf{x}_i$ with the column vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ respectively. Hence, as (33) imposes group sparsity on a vector quantity, (37) and (38) are expected to induce column sparsity on the matrix $\left[\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right]$ thus promoting low-rankness in a matrix factorization framework.

This key feature of the proposed algorithms let us incorporate a pruning procedure which removes the columns that are zeroed as the algorithms evolve. By doing so, the per iteration computational complexity of the algorithms is gradually reduced, and this reduction may become significant, as is also highlighted in Section VI, where empirical numerical results are presented.

V. Convergence Analysis

In this part of the paper we analyze the convergence behavior of AIRLS and AIRLS-MC as presented in Section III and ignoring the above mentioned pruning procedure which is basically an algorithmic mechanism to reduce complexity. The analysis is common for the two algorithms, since, as mentioned above, both minimize upper bound surrogate functions of the same form. Towards this, we first prove the following Lemma.

Lemma 1: The surrogate functions $l(\mathbf{U}|\mathbf{U}_k, \mathbf{V}_k)$ and $g(\mathbf{V}|\mathbf{U}_{k+1}, \mathbf{V}_k)$ minimized at each iteration of AIRLS and AIRLS-MC are tight upper-bounds of the corresponding $f(\mathbf{U}, \mathbf{V}_k)$ and $f(\mathbf{U}_{k+1}, \mathbf{V})$ with $f(\mathbf{U}, \mathbf{V})$ defined in eqs. (22) and (31) for the two algorithms, respectively.

Proof: See Appendix.

Having shown that the proposed surrogate objective functions are upper bounds of the actual ones, in Proposition 2 given below the monotonic decrease of the initial objective functions per iteration of the respective algorithms is established.

Proposition 2: The sequences of $\{\mathbf{U}_k, \mathbf{V}_k\}$ generated by AIRLS and AIRLS-MC decrease monotonically the respective objective functions i.e.,

$$f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) \leq f(\mathbf{U}_{k+1}, \mathbf{V}_k) \leq f(\mathbf{U}_k, \mathbf{V}_k). \tag{39}$$

Proof: See Appendix.

Corollary 1: The sequence $f(\mathbf{U}_k, \mathbf{V}_k)$ converges to $f^\infty \geq 0$, as $k \to \infty$, for both AIRLS and AIRLS-MC.

Proof: Since the objective functions for both algorithms are monotonically decreasing (Proposition 2) and bounded below by 0, the claim follows immediately.

A. Convergence to Stationary Points and Rate of Convergence

Having shown that the updates $(\mathbf{U}_k, \mathbf{V}_k)$ generated by AIRLS and AIRLS-MC monotonically decrease the corresponding objective functions, we herein derive the rates of convergence of the algorithms to stationary points of the functions. The subsequent analysis is along the lines of the one presented in [4].

Given any pair $(\mathbf{U}, \mathbf{V})$ we define matrices $\mathbf{U}_*, \mathbf{V}_*$ arising by the following minimization problems

$$\mathbf{U}_* = \underset{\mathbf{U}^+}{\text{argmin}} \quad l(\mathbf{U}^+|\mathbf{U}, \mathbf{V}) \tag{40}$$

$$\mathbf{V}_* = \underset{\mathbf{V}^+}{\text{argmin}} \quad g(\mathbf{V}^+|\mathbf{U}_*, \mathbf{V}). \tag{41}$$

Let us now denote as $\Delta((\mathbf{U}, \mathbf{V}), (\mathbf{U}_*, \mathbf{V}_*))$ the following measure of proximity between $(\mathbf{U}, \mathbf{V})$ and $(\mathbf{U}_*, \mathbf{V}_*)$,

$$\Delta((\mathbf{U}, \mathbf{V}), (\mathbf{U}_*, \mathbf{V}_*)) = \frac{1}{2}\Big(\|\mathbf{V}\,(\mathbf{U} - \mathbf{U}_*)^T\,\|_F^2$$
$$+ \|\mathbf{U}_*\,(\mathbf{V} - \mathbf{V}_*)^T\,\|_F^2\Big) + \frac{\lambda}{2}\Big(\|\mathbf{D}^{\frac{1}{2}}_{(\mathbf{U},\mathbf{V})}\,(\mathbf{U} - \mathbf{U}_*)^T\,\|_F^2$$
$$+ \|\mathbf{D}^{\frac{1}{2}}_{(\mathbf{U}_*,\mathbf{V})}\,(\mathbf{V} - \mathbf{V}_*)^T\,\|_F^2\Big). \tag{42}$$

Lemma 2: Successive differences in the objective values of cost functions $f(\mathbf{U}, \mathbf{V})$ corresponding to AIRLS and AIRLS-MC are bounded below as follows,

$$f(\mathbf{U}_k, \mathbf{V}_k) - f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}) \geq \Delta((\mathbf{U}_k, \mathbf{V}_k), (\mathbf{U}_{k+1}, \mathbf{V}_{k+1})). \tag{43}$$

Proof: See Appendix.

Lemma 3: $\Delta((\mathbf{U},\mathbf{V}),(\mathbf{U}_*,\mathbf{V}_*))=0$ if and only if $(\mathbf{U},\mathbf{V})$ generated by AIRLS (AIRLS-MC) algorithm is a fixed point of AIRLS (AIRLS-MC).

Proof: See Appendix.

As stated above, $\Delta((\mathbf{U}_k,\mathbf{V}_k),(\mathbf{U}_{k+1},\mathbf{V}_{k+1}))$ is actually used for quantifying the distance between $(\mathbf{U}_k,\mathbf{V}_k)$ and $(\mathbf{U}_{k+1},\mathbf{V}_{k+1})$ generated in successive iterations of the proposed algorithms. Thus, it is obvious that if the algorithms converge this measure will become equal to zero. For ease of notation, we will next denote this quantity as δ_k. That said, the main result of this section is summarized in the following proposition.

Proposition 3: a) Any limit point of the sequences $\{\mathbf{U}_k,\mathbf{V}_k\}$ generated by AIRLS and AIRLS-MC is a stationary point of the respective objective function $f(\mathbf{U},\mathbf{V})$, for $\lambda>0$. b) AIRLS and AIRLS-MC converge sublinearly to stationary points with their rates of convergence expressed as

$$\min_{1\le k\le K}\delta_k\le\frac{f(\mathbf{U}_1,\mathbf{V}_1)-f^{\infty}}{K}. \tag{44}$$

Proof: See Appendix.

Assumption 1: The eigenvalues of $\mathbf{U}_k^T\mathbf{U}_k$ and $\mathbf{V}_k^T\mathbf{V}_k$ for $k\ge 1$ are uniformly bounded below and above by l_L and l_U respectively, i.e.,

$$l_L\mathbf{I}_d\preceq\mathbf{U}_k^T\mathbf{U}_k\preceq l_U\mathbf{I}_d\quad\text{and}\quad l_L\mathbf{I}_d\preceq\mathbf{V}_k^T\mathbf{V}_k\preceq l_U\mathbf{I}_d. \tag{45}$$

Using Assumption 1 we can provide more refined information with regard to the rates of convergence, bringing into play the curvature characteristics of the cost functions as well as the regularization parameter λ.

Corollary 2: Under Assumption 1, we can derive the following convergence rate for Algortithms 1 and 2:

$$\min_{1\le k\le K}\|\mathbf{U}_{k+1}-\mathbf{U}_k\|_F^2+\|\mathbf{V}_{k+1}-\mathbf{V}_k\|_F^2 \le\frac{4\tau}{2l_L\tau+\lambda}\frac{f(\mathbf{U}_1,\mathbf{V}_1)-f^{\infty}}{K}, \tag{46}$$

where $\tau=\max_{1\le i\le d}(\|\boldsymbol{u}_i\|_2^2,\|\boldsymbol{v}_i\|_2^2)$.

Proof: It can be easily proved by suitably modifying δ_k using the inequalities $l_L\|\mathbf{U}_k-\mathbf{U}_{k+1}\|_F^2\le\|\mathbf{V}_k\left(\mathbf{U}_k-\mathbf{U}_{k+1}\right)^T\|_F^2\le l_U\|\mathbf{U}_k-\mathbf{U}_{k+1}\|_F^2$ and $l_L\|\mathbf{V}_k-\mathbf{V}_{k+1}\|_F^2\le\|\mathbf{U}_{k+1}\left(\mathbf{V}_k-\mathbf{V}_{k+1}\right)^T\|_F^2\le l_U\|\mathbf{V}_k-\mathbf{V}_{k+1}\|_F^2$.

VI. Experiments

Next simulated and real data experiments are provided for illustrating the key features of the proposed AIRLS and AIRLS-MC algorithms. It has been empirically observed that parameter p does not seem to play a crucial role in the performance of the algorithms. Therefore, in all experiments provided next, the parameter p is set to 1. For comparison purposes, an alternating regularized least squares (noted here as ALS) algorithm corresponding to the full-observation version of the softImpute-ALS proposed in [4] is utilized in the denoising type problems. In matrix completion experiments the softImpute-ALS algorithm, [4], and the iterative reweighted nuclear norm (IRNN) algorithm of [31] are employed. It should be noted that IRNN goes beyond the traditional nuclear norm minimization by imposing various sparsity imposing priors on the vector of singular values. This scheme gives rise to weighted nonconvex analogues of the traditional nuclear norm. In the sequel, we restrict our attention to IRNN which arises by applying the $\ell_{1/2}$ quasinorm on the vector of singular values. Note that IRNN, unlike AIRLS and softImpute-ALS, is not an MF-based approach and thus involves computationally demanding singular value decomposition (SVD) operations at each iteration. It should be noted that for the two proposed algorithms *a column pruning mechanism is applied*. That is, when a column of the matrix factors has been (approximately) zeroed, it is removed, thus reducing the column size of the factors. As a result, the per iteration complexity is being reduced during the execution of the algorithms. All experiments were conducted on an Intel Core i7-4790 CPU 3.60GHz x 8 CPU with 16GB RAM.

A. Simulated Data Experiments

Herein we highlight the benefits of the proposed AIRLS and AIRLS-MC algorithms on simulated data. To this end, the algorithms are tested on two different experimental setups i.e. a) for checking the performance of AIRLS in the presence of noise and b) for testing the capacity of AIRLS-MC in dealing with different percentages of missing data.

1) AIRLS: In order to validate the performance of AIRLS in the presence of noise a matrix $\mathbf{X}_0\in\mathbb{R}^{m\times n}$ with $m=1000$, $n=1000$ and varying rank $r\in\{5,10\}$ is randomly generated. Concretely, matrix $\mathbf{X}_0$ is produced by the product of two matrices i.e., $\mathbf{U}_0\in\mathbb{R}^{m\times r}$ and $\mathbf{V}_0^T\in\mathbb{R}^{r\times n}$ having zero-mean Gaussian entries of variance 1. Additive Gaussian i.i.d. noise of different signal to noise ratio (SNR) i.e., $\mathrm{SNR}\in\{10,20\}$ corrupts $\mathbf{X}_0$, thus resulting to the data matrix $\mathbf{Y}$, which is then provided as input to the tested algorithms. AIRLS is compared to the ALS algorithm, as mentioned before. As a quantitative metric we utilize the normalized reconstruction error (NRE) defined as $\mathrm{NRE}=\frac{\|\mathbf{X}_0-\hat{\mathbf{U}}\hat{\mathbf{V}}^T\|_F}{\|\mathbf{X}_0\|_F}$. Since we are interested in the recovery performance of the algorithms, the low-rank promoting parameter λ of the algorithms is selected from a set of values $\{0.1, 1, 5, 10, 50, 80, 100, 200\}$ via fine tuning in terms of the lowest achieved NRE. The algorithms stop when either the relative decrease of the reconstructed data between two successive iterations i.e., $\frac{\|\hat{\mathbf{U}}_k\hat{\mathbf{V}}_k^T-\hat{\mathbf{U}}_{k+1}\hat{\mathbf{V}}_{k+1}^T\|_F}{\|\hat{\mathbf{U}}_k\hat{\mathbf{V}}_k^T\|_F}$, becomes less than 10^{-5} or 500 iterations are reached. 100 independent runs are performed for each algorithm and the average values of the various quantities (elapsed time, NRE, iterations executed and estimated rank) are calculated. The initial rank is set to $d=25$.

In Table I, the results of AIRLS and ALS are given. Therein, it is shown that AIRLS offers better estimation performance than ALS in all experiments. Interestingly, in most cases, this happens in less time than that spent by ALS within fewer iterations.

Next we aim at illustrating the competence of AIRLS in estimating the actual rank of $\mathbf{X}_0$ as well as showing the gains obtained by using the column pruning mechanism, which is adopted for the proposed algorithms. To this end, the same experimental setting described above is used. Again, the initial rank d is set to 25, and the low-rank regularization parameter λ is fine-tuned with respect to the minimum achieved NRE.

TABLE I
RESULTS OBTAINED BY ALS AND AIRLS ON THE SIMULATED DENOISING EXPERIMENT

SNR	10						20					
rank	5			10			5			10		
Algorithm	# Iter	time(s)	NRE	# Iter	time(s)	NRE	# Iter	time(s)	NRE	# Iter	time(s)	NRE
ALS	34.82	1.59	0.058	248.80	11.69	0.066	159.42	7.52	0.015	239.92	11.30	0.020
AIRLS	143.96	5.91	0.031	112.97	4.70	0.044	40.92	1.68	0.010	98.56	4.16	0.014

TABLE II
RESULTS OBTAINED BY ALS, AIRLS AND AIRLS WITH NO COLUMN PRUNING MECHANISM (AIRLS-(NO CP)) ON A SIMULATED DENOISING EXPERIMENT FOR DIFFERENT VALUES OF THE RANK

rank	5			10			15			20		
Algorithm	est. rank	time(s)	NRE	est. rank	time(s)	NRE	est. rank	time(s)	NRE	est. rank	time(s)	NRE
ALS	25	15.69	0.034	25	15.45	0.040	25	15.45	0.041	25	14.46	0.041
AIRLS	5	1.86	0.017	10	5.66	0.025	15	5.82	0.030	20	5.97	0.035
AIRLS-(NO CP)	5	2.55	0.017	10	6.41	0.025	15	6.32	0.030	20	6.24	0.035

The SNR is kept fixed to 15 dB and 5 different values of the true rank, i.e., $r \in \{5, 10, 15, 20\}$ are considered. For comparison purposes, ALS and a variant of AIRLS having the column pruning mechanism deactivated are utilized. As it can be seen in Table II, contrary to ALS, AIRLS estimates the actual rank in less running time regardless of the use of the column pruning mechanism. Notably, in the case that the column pruning mechanism is deactivated, the number of the nonzero columns of the estimated matrix factors after convergence coincides with the true rank. This key observation verifies the column-sparsity promoting characteristic of the proposed low-rank regularization term. As it can be also seen, the two AIRLS versions converge to the same NRE. It is thus evident from Table II that the effect of the incorporation of the column pruning is merely to decrease the running time of the algorithms, as a consequence of the gradual reduction of the size of the matrix factors.

2) AIRLS-MC: To evaluate the performance of AIRLS-MC in different scenarios, we classify the experimental settings of this subsection according to the degrees of freedom ratio (FR), [18], defined as $\mathrm{FR} = r(2n - r)/\mathrm{card}(\Omega)$. Recovery becomes harsher as FR is close to 1, whereas easier problems arise when it takes values close to 0. AIRLS-MC is compared to softImpute-ALS and IRNN for FR equal to 0.4 and 0.6. In both cases a low-rank matrix $\mathbf{X}_0 \in \mathbb{R}^{m \times n}$ with $m = 1500$, $n = 1500$ and rank $r = 20$ is generated following the same setting as in the case of the denoising experiment described above. The NRE is used as the performance metric. For all algorithms, the parameter λ which is related to low-rank imposition, is fine tuned and the initial rank of the MF-based methods is set to 35. The algorithms run for 20 instances of each experiment and the mean values of iterations, NRE and time to converge are given in Table III. Moreover, the same stopping criteria mentioned previously are utilized. As is shown in Table III, AIRLS-MC offers significantly higher accuracy than softImpute-ALS in both experiments. On the other hand, IRNN performs similarly to AIRLS in terms of NRE, at the cost of a much higher runtime. This shortcoming of IRNN is due to the computationally demanding SVDs executed at each iteration.

TABLE III
RESULTS OF AIRLS-MC, SOFTIMPUTE-ALS AND IRNN ON THE SIMULATED MATRIX COMPLETION EXPERIMENT

FR	0.4			0.6		
Algorithm	# Iter	time(s)	NRE	# Iter	time(s)	NRE
softImpute-ALS	384	49.67	0.114	500	47.53	0.3902
AIRLS-MC	488	59.12	0.070	500	45.91	0.2882
IRNN	474	698.93	0.073	298	417.79	0.2812

B. Real Data Experiments

In this section we validate the performance of the proposed algorithms in two different real data experiments. The AIRL algorithm is evaluated in denoising a real hyperspectral image (HSI) and a collaborative filtering application is used for testing the performance of AIRLS-MC algorithm.

1) Hyperspectral Image Denoising: In this experiment we utilize the Washigton DC Mall AVIRIS HSI captured at $m = 210$ contiguous spectral bands in the 0.4 to 2.4 μm region of the visible and infrared spectrum. The HSI consists of $n = 22500$ (150×150) pixels. As is widely known, [32], hyperspectral data are highly coherent both in the spectral and the spatial domains. Therefore, by organizing the tested image in a matrix, whereby each column corresponds to the spectral bands and each row to the pixels, it turns out that this matrix can be well approximated by a low-rank one. This fact motivates us to exploit the low-rank structure of the HSI under study for efficiently denoising a highly corrupted version thereof by Gaussian i.i.d. noise of $\mathrm{SNR} = 6$ dB.

In Fig. 1, false RGB images of the recovered HSIs by the proposed AIRLS algorithm and ALS are provided. In both algorithms, the number of columns of the initial factors $\mathbf{U}_0$ and $\mathbf{V}_0$ is overstated to $d = 100$ and the algorithms terminate when the relative decrease of the reconstructed HSI between two successive iterations reaches a value less than 10^{-4}. Moreover, their low-rank promoting parameter λ is selected so as to lead to solution matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ of the same rank $r = 4$. As it can be noticed in Fig. 1, AIRLS reconstructs the HSI in a significantly improved accuracy as compared to ALS. This can be easily verified both by visually inspecting Figs. 1(a)–1(d) and quantitatively in terms of the estimated NRE (Fig. 1(e)). Notably, AIRLS converges in

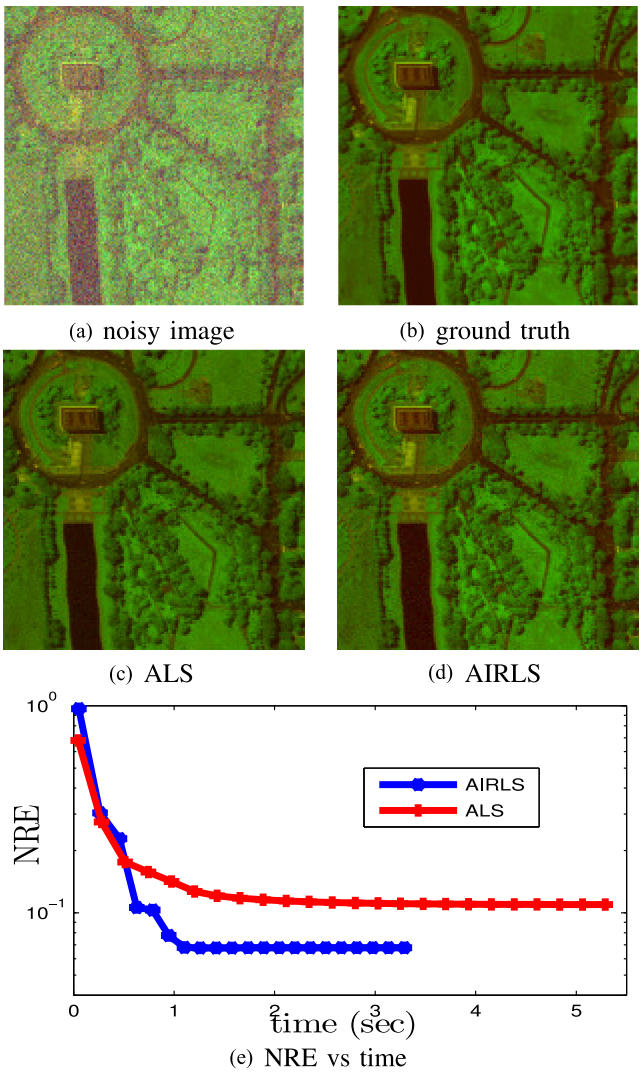


Fig. 1. Evaluation of AIRLS and ALS on the Washigton DC AVIRIS dataset.

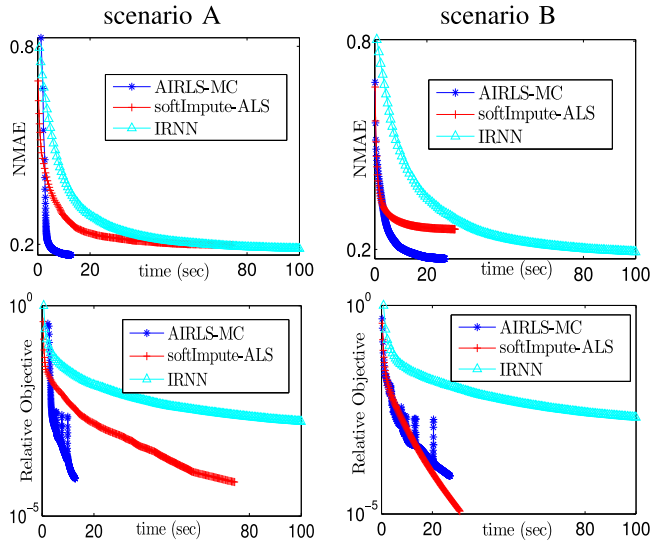


Fig. 2. NMAE and relative objective vs time evolution (up to 100 secs) of AIRLS-MC, softImpute-ALS and IRNN on the Movielens 100K validation dataset.

less iterations than those required by ALS (Fig. 1(e)), while at the same time less time per iteration is consumed, on average. The latter is achieved by virtue of the column pruning mechanism of AIRLS, which gradually reduces the size of matrix factors from $m \times 100$ and $n \times 100$ to $m \times 4$ and $n \times 4$, respectively. This way, after only a few initial iterations, when the rank starts to decrease, the per iteration time complexity of AIRLS becomes much smaller than that required in its early iterations, as well as the one of ALS.

2) MC on Movielens 100K and 1M datasets: Herein, we focus on testing the performance of AIRLS-MC algorithm on a popular collaborative filtering application i.e. a movie recommender system. To this end, we utilize two well-studied in literature large datasets: the Movielens 100K and the Movielens 1M datasets. Both datasets contain ratings by users collected over various periods of time, with integer values ranging from 1–5. Since most of the entries are missing, matrix completion algorithms can be utilized for predicting them. By assuming that there exists a high degree of correlation amongst the rating of different users, a low-rank structure can be meaningfully adopted for these datasets. For the case of the 100K dataset the "ub.base"[2] file which contains $\approx 90\%$ of the total ratings was splitted into two disjoint sets i.e., a training set (consisting of $\approx 65\%$ of the total per user ratings) and a validation set ($\approx 25\%$). The "ub.test" file which contains $\approx 10\%$ of the ratings was utilized as the test set. For the case of the 1M Movielens dataset, the "ratings.dat" file was splitted into 3 disjoint sets, that is, a training set consisting of $\approx 50\%$ of the total ratings per user, a validation set $\approx 25\%$ and a test set ($\approx 25\%$). Note that the 100K dataset contains 100000 ratings of 943 users on 1682 movies with each user having rated at least 20 movies. That said, we need to address a quite challenging matrix completion problem, since 93% of the elements are missing. The situation is even harsher for the 1M dataset, which includes 1 million ratings from 6040 users on 3900 movies and 96% missing data. Finally, the normalized mean absolute value error (NMAE) defined as $\text{NMAE} = \frac{\sum_{(i,j)\in\Omega} |[\mathbf{U}\mathbf{V}^T]_{ij} - [\mathbf{Y}]_{ij}|}{4\text{card}(\Omega)}$ is used as a performance metric.

First, we aim at illustrating the behavior of the proposed AIRLS-MC algorithm when it comes to the estimation performance and the speed of convergence. In this regard, for the case of the 100K dataset, the state-of-the-art IRNN and softImpute-ALS algorithms are utilized for comparison purposes. The low-rank promoting parameter λ of all competing algorithms is selected according to two different scenarios: A) we choose λ that achieves the minimum NMAE on the validation set after convergence and B) we select λ so that the estimated matrices by both the tested algorithms are of the same rank, equal to 10. It should be noted that the same stopping criterion used in the previous experiment is adopted also here. As it can be seen in Fig. 2 and Table IV, the softImpute-ALS algorithm requires in general

[2]Movielens 100K and 1M datasets can be downloaded from https://grouplens.org/datasets/movielens/

TABLE IV
RESULTS OBTAINED BY AIRLS-MC AND SOFTIMPUTE-ALS ON MOVIELENS 100K DATASET

			# Iter	msec/iter	total time (sec)	NMAE
scenario	A	softImpute-ALS	247	300.2	74.3	0.2362
		IRNN	500	598.5	299.2	0.2036
		AIRLS-MC	591	21.7	12.8	0.2005
	B	softImpute-ALS	156	197.6	30.8	0.2968
		IRNN	500	740.0	370.2	0.2029
		AIRLS-MC	969	28.5	27.6	0.2010

less iterations to converge than both AIRLS-MC and IRNN. However, the average per-iteration time complexity of AIRLS-MC is significantly less compared to its rivals. As is mentioned above, this is attributed to the column pruning scheme which decreases to a large degree the computational burden of the algorithm. This favorable property, results to a faster convergence of AIRLS-MC in both scenarios A and B as compared to both softImpute-ALS and IRNN in terms of time. Among the three algorithms tested, IRNN is clearly the most demanding one in terms of average per-iteration time complexity as it can be observed from Fig. 2. As mentioned above, this is ascribed to the fact that IRNN entails "expensive" SVD operations, in sharp contrast to the other two MF-based algorithms. It should be noted that in scenario A, the estimated by AIRLS-MC and IRNN matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ have rank equal to 6. On the other hand, in softImpute-ALS the solution matrices have rank equal to the one used at the initialization stage i.e., 100.

When it comes to the generalization performance of the proposed algorithm, from Table IV it can be observed that, in both scenarios A and B, AIRLS-MC achieved lower NMAE on the unseen test set than its MF counterpart softImpute-ALS and slightly lower NMAE than IRNN. This actually shows that the reduced computational complexity of AIRLS-MC does not come at a price of inferior performance in terms of the accuracy of the estimated matrices. Lastly, from Fig. 2 it can be noticed that the relative objective of AIRLS-MC presents abrupt increases at some iterations. It was experimentally verified that those changes (which imply large decreases of the successive values of the objective function) take place at iterations that coincide with zeroings of the columns of the matrix factors. This fact advocates that larger gains are obtained at iterations where the rank is reduced, as we are approaching at the low-rank solution matrices.

Fig. 3 and Table V show the performance of AIRLS-MC and softImpute-ALS on the 1M Movielens dataset.[3] The parameter λ of AIRLS-MC and softImpute-ALS is fined tuned, in the same way as in the 100K experiment, based on the best NMAE attained by the algorithms on the validation set. The rank is again initialized to $d = 100$ for both algorithms. Interestingly, AIRLS-MC reaches a more accurate solution in terms of the NMAE (as evaluated on the test set) in almost 40% of the time required by softImpute-ALS. Again, AIRLS-MC requires more iterations to converge as compared to its competitor. Nevertheless, as it can be also seen in Fig. 3, the column pruning mechanism which

[3]IRNN has not been included in this experiment owing to its higher computational requirements as compared to both the MF-based algorithms.

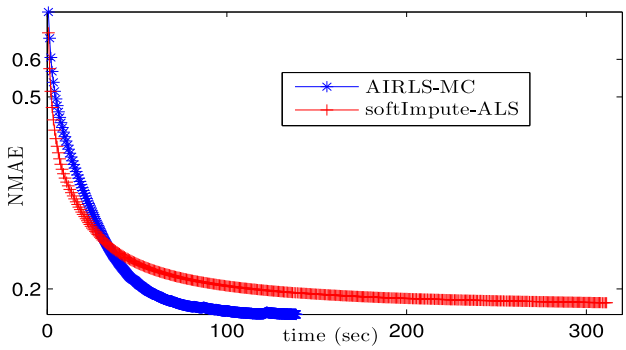


Fig. 3. Evaluation of AIRLS-MC and softImpute-ALS on 1M Movielens dataset.

TABLE V
RESULTS OBTAINED BY AIRLS-MC AND SOFTIMPUTE-ALS ON MOVIELENS 1M DATASET

	# Iter	msec/iter	total time (sec)	NMAE
softImpute-ALS	433	720	311.2	0.1862
AIRLS-MC	903	153	138.9	0.1760

is activated in the initial iterations of AIRLS-MC results to a significant reduction of the average time spent per iteration.

VII. CONCLUSION

This paper presents a novel generic formulation of the low-rank matrix factorization problem. Borrowing ideas from iteratively reweighted approaches for rank minimization, a reweighted version of the sum of the squared Frobenious norms of the matrix factors i.e., a non-convex variational characterization of the nuclear norm, is defined. The proposed framework encapsulates other state-of-the-art approaches for low-rank imposition on the matrix factorization setting. By focusing on a specific instance of this scheme we define a joint-column sparsity inducing regularizer that couples the columns of the matrix factors. As is mathematically shown, the resulting low-rank regularization term is a tight-upper bound of a specific version of the recently proposed weighted nuclear norm. The ubiquity of the proposed approach is demonstrated in the problems of denoising and matrix completion. To this end, under the block successive upper bound minimization framework, alternating IRLS type algorithms are devised for addressing the afore-mentioned problems. The efficiency of the proposed algorithms in handling big and high-dimensional data as compared to other state-of-the-art algorithms is illustrated in a wealth of simulated and real data experiments.

APPENDIX

A. Proof of Proposition 1

The upper bound expression given in (14) can be easily established starting from (10) with $\mathbf{W}_{\mathrm{U}} = \mathbf{W}_{\mathrm{V}} = \mathbf{W}$, and $\mathbf{W}$ given by (11), following the lines of [33, Lemma 7] and assuming, without loss of generality, that the columns of the concatenated

matrix $\left[\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right]$ are arranged in non-increasing order with respect to their ℓ_2 norms.

B. Proof of Lemma 1

The surrogate functions $l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)$ and $g(\mathbf{V}|\mathbf{U}_{k+1},\mathbf{V}_k)$ given in eqs. (24) and (29), are twice continuously differentiable and constitute approximations of the second order Taylor expansions of the initial cost functions around $(\mathbf{U}_k,\mathbf{V}_k)$ and $(\mathbf{U}_{k+1},\mathbf{V}_k)$ respectively. In (24), the true Hessian $\mathbf{H}_{\mathbf{U}_k}$ of $f(\mathbf{U},\mathbf{V}_k)$ at $\mathbf{U}_k$ has been approximated by the $md\times md$ positive-definite block diagonal matrix $\bar{\mathbf{H}}_{\mathbf{U}_k}$ defined in (25). $\bar{\mathbf{H}}_{\mathbf{V}_k}$ is similarly defined. Our analysis is next focused on $l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)$. It can be easily shown that similar derivations can be made for $g(\mathbf{V}|\mathbf{U}_{k+1},\mathbf{V}_k)$. As it can be seen by eq. (24), $l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)$ equals $f(\mathbf{U},\mathbf{V}_k)$ at $(\mathbf{U}_k,\mathbf{V}_k)$. In order to show that it majorizes $f(\mathbf{U},\mathbf{V}_k)$ for all other points closeby, it suffices to show that matrix $\mathbf{A}=\bar{\mathbf{H}}_{\mathbf{U}_k}-\mathbf{H}_{\mathbf{U}_k}$ is positive semi-definite [29]. Next we prove that for each of the two problems examined, the above-mentioned property holds for $\mathbf{A}$.

In denoising $\tilde{\mathbf{H}}_{\mathbf{U}_k}=\mathbf{V}_k^T\mathbf{V}_k+\lambda\mathbf{D}_{(\mathbf{U}_k,\mathbf{V}_k)}$, where $\mathbf{D}_{(\mathbf{U}_k,\mathbf{V}_k)}$ is defined in eq. (27). Moreover it can be shown that for the exact Hessian $\mathbf{H}_{\mathbf{U}_k}$ at $\mathbf{U}_k$ we get

$$\mathbf{H}_{\mathbf{U}_k}=\mathbf{I}_m\otimes(\mathbf{V}_k^T\mathbf{V}_k)+\lambda\mathbf{K}, \tag{47}$$

where $\mathbf{K}=[\mathbf{K}_{ij}], i,j=1,2,\ldots,m$ consists of $d\times d$ blocks $\mathbf{K}_{ij}$ defined in (48) shown at the bottom of this page. Hence matrix $\mathbf{A}=[\mathbf{A}_{ij}]$ is expressed as follows

$$\mathbf{A}=\mathbf{I}_m\otimes\mathbf{D}_{(\mathbf{U}_k,\mathbf{V}_k)}-\lambda\mathbf{K}. \tag{49}$$

Elaborating on $\mathbf{A}$ we get from (49), (48) and (27),

$$\mathbf{A}_{ij}=\lambda p(2-p)\mathrm{diag}\left(\frac{u_{i1}^k u_{j1}^k}{\left(\|\boldsymbol{u}_1^k\|_2^2+\|\boldsymbol{v}_1^k\|_2^2+\eta^2\right)^{2-p/2}},\ldots,\frac{u_{id}^k u_{jd}^k}{\left(\|\boldsymbol{u}_d^k\|_2^2+\|\boldsymbol{v}_d^k\|_2^2+\eta^2\right)^{2-p/2}}\right). \tag{50}$$

Notice that for

$$\mathbf{B}_i=\sqrt{\lambda p(2-p)}\mathrm{diag}\left(\frac{u_{i1}^k}{\left(\|\boldsymbol{u}_1^k\|_2^2+\|\boldsymbol{v}_1^k\|_2^2+\eta^2\right)^{1-p/4}},\ldots,\frac{u_{id}^k}{\left(\|\boldsymbol{u}_d^k\|_2^2+\|\boldsymbol{v}_d^k\|_2^2+\eta^2\right)^{1-p/4}}\right) \tag{51}$$

$\mathbf{A}_{ij}=\mathbf{B}_i^T\mathbf{B}_j$. So by defining $\mathbf{B}=[\mathbf{B}_1,\ldots,\mathbf{B}_m]$, it is straightforward that $\mathbf{A}=\mathbf{B}^T\mathbf{B}$, that is $\mathbf{A}$ is positive semi-definite.

In matrix completion, the exact Hessian $\mathbf{H}_{\mathbf{U}_k}$ differs from that given in (47) in the diagonal blocks only. More specifically, the ith diagonal block of $\mathbf{H}_{\mathbf{U}_k}$ takes now the form $\mathbf{V}^T\mathbf{\Phi}_i\mathbf{V}+\mathbf{K}_{ii}$, where $\mathbf{\Phi}_i$ is a $n\times n$ diagonal matrix containing ones on indexes included in the set Ω and related to the ith row of $\mathbf{Y}$ and zeros elsewhere. Since $\mathbf{V}^T\mathbf{V}-(\mathbf{V}^T\mathbf{\Phi}_i\mathbf{V})\succeq 0$, we can use the same arguments as above for proving the semi-definiteness of the respective matrix $\mathbf{A}$.

C. Proof of Proposition 2

From Lemma 1 we have,

$$l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)\geq f(\mathbf{U},\mathbf{V}_k) \tag{52}$$

Since $\mathbf{U}_{k+1}=\underset{\mathbf{U}}{\mathrm{argmin}}\ \ l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)$ we get

$$l(\mathbf{U}_{k+1}|\mathbf{U}_k,\mathbf{V}_k)\leq l(\mathbf{U}_k|\mathbf{U}_k,\mathbf{V}_k)\equiv f(\mathbf{U}_k,\mathbf{V}_k) \tag{53}$$

and $l(\mathbf{U}_{k+1}|\mathbf{U}_k,\mathbf{V}_k)\geq f(\mathbf{U}_{k+1},\mathbf{V}_k)$ which leads to

$$f(\mathbf{U}_{k+1},\mathbf{V}_k)\leq f(\mathbf{U}_k,\mathbf{V}_k). \tag{54}$$

Following the same reasoning, and since $\mathbf{V}_{k+1}=\mathrm{argmin}_{\mathbf{V}}\ \ g(\mathbf{V}|\mathbf{U}_{k+1},\mathbf{V}_k)$ we get

$$g(\mathbf{V}_k|\mathbf{U}_{k+1},\mathbf{V}_k)\equiv f(\mathbf{U}_{k+1},\mathbf{V}_k)\geq g(\mathbf{V}_{k+1}|\mathbf{U}_{k+1},\mathbf{V}_k)\geq f(\mathbf{U}_{k+1},\mathbf{V}_{k+1}) \tag{55}$$

Combining (54) and (55) we get (39).

D. Proof of Lemma 2

Using Lemma 1, we have,

$$f(\mathbf{U}_k,\mathbf{V}_k)-f(\mathbf{U}_{k+1},\mathbf{V}_k)\geq l(\mathbf{U}_k|\mathbf{U}_k,\mathbf{V}_k)-l(\mathbf{U}_{k+1}|\mathbf{U}_k,\mathbf{V}_k)\quad\text{and} \tag{56}$$

$$f(\mathbf{U}_{k+1},\mathbf{V}_k)-f(\mathbf{U}_{k+1},\mathbf{V}_{k+1})\geq g(\mathbf{V}_k|\mathbf{U}_{k+1},\mathbf{V}_k)-g(\mathbf{V}_{k+1}|\mathbf{U}_{k+1},\mathbf{V}_k) \tag{57}$$

Adding (56) and (57) we reach to the following inequality

$$f(\mathbf{U}_k,\mathbf{V}_k)-f(\mathbf{U}_{k+1},\mathbf{V}_{k+1})\geq l(\mathbf{U}_k|\mathbf{U}_k,\mathbf{V}_k)-l(\mathbf{U}_{k+1}|\mathbf{U}_k,\mathbf{V}_k)+g(\mathbf{V}_k|\mathbf{U}_{k+1},\mathbf{V}_k)-g(\mathbf{V}_{k+1}|\mathbf{U}_{k+1},\mathbf{V}_k) \tag{58}$$

Since $\mathbf{U}_{k+1}$ and $\mathbf{V}_{k+1}$ are stationary points of $l(\mathbf{U}|\mathbf{U}_k,\mathbf{V}_k)$ and $g(\mathbf{V}|\mathbf{U}_{k+1},\mathbf{V}_k)$ respectively

($\nabla_{\mathbf{U}}l(\mathbf{U}_{k+1}|\mathbf{U}_k,\mathbf{V}_k)=\mathbf{0}$ and $\nabla_{\mathbf{V}}g(\mathbf{V}_{k+1}|\mathbf{U}_{k+1},\mathbf{V}_k)=\mathbf{0}$) and by their second order Taylor expansions around

$$\mathbf{K}_{ij}=\begin{cases}p\mathrm{diag}\left(\frac{\|\boldsymbol{u}_1^k\|_2^2+\|\boldsymbol{v}_1^k\|_2^2-(2-p)(u_{i1}^k)^2+\eta^2}{\left(\|\boldsymbol{u}_1^k\|_2^2+\|\boldsymbol{v}_1^k\|_2^2+\eta^2\right)^{2-p/2}},\ldots,\frac{\|\boldsymbol{u}_d^k\|_2^2+\|\boldsymbol{v}_d^k\|_2^2-(2-p)(u_{id}^k)^2+\eta^2}{\left(\|\boldsymbol{u}_d^k\|_2^2+\|\boldsymbol{v}_d^k\|_2^2+\eta^2\right)^{2-p/2}}\right),\text{if } i=j\\ p(2-p)\mathrm{diag}\left(\frac{-u_{i1}^k u_{j1}^k}{\left(\|\boldsymbol{u}_1^k\|_2^2+\|\boldsymbol{v}_1^k\|_2^2+\eta^2\right)^{2-p/2}},\ldots,\frac{-u_{id}^k u_{jd}^k}{\left(\|\boldsymbol{u}_d^k\|_2^2+\|\boldsymbol{v}_d^k\|_2^2+\eta^2\right)^{2-p/2}}\right),\text{if } i\neq j\end{cases} \tag{48}$$

$(\mathbf{U}_{k+1}, \mathbf{V}_k)$ and $(\mathbf{U}_{k+1}, \mathbf{V}_{k+1})$ we have

$$\begin{aligned} & l(\mathbf{U}_k|\mathbf{U}_k, \mathbf{V}_k) - l(\mathbf{U}_{k+1}|\mathbf{U}_k, \mathbf{V}_k) \\ &= \frac{1}{2}\mathrm{tr}\{(\mathbf{U}_k - \mathbf{U}_{k+1})\left(\mathbf{V}_k^T\mathbf{V}_k \right. \\ &\quad \left. + \lambda\mathbf{D}_{(\mathbf{U}_k, \mathbf{V}_k)}\right)(\mathbf{U}_k - \mathbf{U}_{k+1})^T\} \qquad (59) \\ &= \frac{1}{2}\|\mathbf{V}_k\,(\mathbf{U}_k - \mathbf{U}_{k+1})^T\,\|_F^2 \\ &\quad + \frac{\lambda}{2}\|\mathbf{D}^{\frac{1}{2}}_{(\mathbf{U}_k, \mathbf{V}_k)}\,(\mathbf{U}_k - \mathbf{U}_{k+1})^T\,\|_F^2 \qquad (60) \end{aligned}$$

and

$$\begin{aligned} & g(\mathbf{V}_k|\mathbf{U}_{k+1}, \mathbf{V}_k) - g(\mathbf{V}_{k+1}|\mathbf{U}_{k+1}, \mathbf{V}_k) \\ &= \frac{1}{2}\mathrm{tr}\{(\mathbf{V}_k - \mathbf{V}_{k+1})\left(\mathbf{U}_{k+1}^T\mathbf{U}_{k+1} \right. \\ &\quad \left. + \lambda\mathbf{D}_{(\mathbf{U}_{k+1}, \mathbf{V}_k)}\right)(\mathbf{V}_{k+1} - \mathbf{V}_k)^T\} \qquad (61) \\ &= \frac{1}{2}\|\mathbf{U}_{k+1}\,(\mathbf{V}_k - \mathbf{V}_{k+1})^T\,\|_F^2 \\ &\quad + \frac{\lambda}{2}\|\mathbf{D}^{\frac{1}{2}}_{(\mathbf{U}_{k+1}, \mathbf{V}_k)}\,(\mathbf{V}_k - \mathbf{V}_{k+1})^T\,\|_F^2 \qquad (62) \end{aligned}$$

Combining (60), (62) and (58) we get inequality (43).

E. Proof of Lemma 3

If $(\mathbf{U}, \mathbf{V})$ is a fixed point, i.e. $\mathbf{U} = \mathbf{U}_*$ and $\mathbf{V} = \mathbf{V}_*$, then it is easily shown that $\Delta((\mathbf{U}, \mathbf{V}), (\mathbf{U}_*, \mathbf{V}_*)) = 0$. Conversely, using (60) and (62) and since all the summands of $\Delta((\mathbf{U}, \mathbf{V}), (\mathbf{U}_*, \mathbf{V}_*))$ are non-negative, we have that if $\Delta((\mathbf{U}, \mathbf{V}), (\mathbf{U}_*, \mathbf{V}_*)) = 0$ then

$$l(\mathbf{U}|\mathbf{U}, \mathbf{V}) - l(\mathbf{U}_*|\mathbf{U}, \mathbf{V}) = 0 \;\text{ and} \qquad (63)$$

$$g(\mathbf{V}|\mathbf{U}_*, \mathbf{V}) - g(\mathbf{V}_*|\mathbf{U}_*, \mathbf{V}) = 0. \qquad (64)$$

Since both $l(\mathbf{U}|\mathbf{U}, \mathbf{V})$ and $g(\mathbf{V}|\mathbf{U}_*, \mathbf{V})$ are strictly convex functions, $\mathbf{U}_*$ and $\mathbf{V}_*$ are uniquely acquired. Hence the above equalities hold only if $(\mathbf{U}, \mathbf{V}) = (\mathbf{U}_*, \mathbf{V}_*)$, that is $(\mathbf{U}, \mathbf{V})$ is a fixed point of AIRLS (AIRLS-MC).

F. Proof of Proposition 3

a) We say that $(\mathbf{U}_*, \mathbf{V}_*)$ is a first order stationary point of $f(\mathbf{U}, \mathbf{V})$ given either in (22) or (31) if the following holds

$$\nabla_{\mathbf{U}} f(\mathbf{U}_*, \mathbf{V}_*) = \mathbf{0}, \;\; \nabla_{\mathbf{V}} f(\mathbf{U}_*, \mathbf{V}_*) = \mathbf{0}. \qquad (65)$$

Due to the adopted upper bound minimization approach, it is easily shown that (65) can be equivalently restated as [4],

$$\mathbf{U}_* = \arg\min_{\mathbf{U}} l(\mathbf{U}|\mathbf{U}_*, \mathbf{V}_*), \;\; \mathbf{V}_* = \arg\min_{\mathbf{V}} g(\mathbf{V}|\mathbf{U}_*, \mathbf{V}_*) \qquad (66)$$

i.e. $(\mathbf{U}_*, \mathbf{V}_*)$ is a fixed-point of AIRLS (AIRLS-MC).

Now, for $\lambda > 0$, the sequence $\{\mathbf{U}_k, \mathbf{V}_k\}$ generated by AIRLS (AIRLS-MC) remains bounded since $f(\mathbf{U}_k, \mathbf{V}_k)$ is coercive (i.e., $f(\mathbf{U}_k, \mathbf{V}_k) \to +\infty$ iff $\|\mathbf{U}\|_F \to +\infty$ or $\|\mathbf{V}\|_F \to +\infty$) and thus contains convergent subsequences. Let $(\mathbf{U}_*, \mathbf{V}_*)$ be a limit point of AIRLS (AIRLS-MC). That said, there will be a subsequence $\{\mathbf{U}_k, \mathbf{V}_k\}$ that converges to $(\mathbf{U}_*, \mathbf{V}_*)$ hence $\Delta((\mathbf{U}_k, \mathbf{V}_k), (\mathbf{U}_*, \mathbf{V}_*)) \to 0$. From Lemma 3, we know that $\Delta((\mathbf{U}_k, \mathbf{V}_k), (\mathbf{U}_*, \mathbf{V}_*)) = 0$ iff $(\mathbf{U}_*, \mathbf{V}_*)$ is a fixed point of the algorithms. Hence, due to the equivalence of (65) and (66), it can be easily conjectured that $(\mathbf{U}_*, \mathbf{V}_*)$ will be also a stationary point of the minimized cost function.

b) Recall that $\delta_k = \Delta((\mathbf{U}_k, \mathbf{V}_k), (\mathbf{U}_{k+1}, \mathbf{V}_{k+1}))$. Then from (43) by adding K successive terms we get,

$$\sum_{k=1}^{K} \delta_k \le f(\mathbf{U}_1, \mathbf{V}_1) - f(\mathbf{U}_K, \mathbf{V}_K) \le f(\mathbf{U}_1, \mathbf{V}_1) - f^{\infty} < \infty. \qquad (67)$$

Note that all the terms of the sequence δ_k take nonnegative values. Let us now assume that there exists a subsequence of δ_k that converges to a positive number. In such a case the sum $\sum_{k=1}^{K} \delta_k$ would not be bounded as $K \to \infty$, which contradicts (67). Therefore, all subsequences of δ_k converge to zero, i.e. the sequence δ_k also converges to zero. From Lemma 3, the zero limit point of δ_k is in fact a fixed point of AIRLS (AIRLS-MC) which as said above, is a stationary point of the respective objective function $f(\mathbf{U}, \mathbf{V})$.

By substituting the first part of inequality (67) by $K\min_{1\le k\le K}\delta_k \le \sum_{k=1}^{K}\delta_k$ and solving for $\min_{1\le k\le K}\delta_k$ we get (44), which establishes a sublinear convergence rate for the proposed algorithms [4].

REFERENCES

[1] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. San Francisco, CA, USA: Academic Press, 2015.

[2] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Dept. Electr. Eng., Stanford Univ., Stanford, CA, USA, 2002.

[3] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[4] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank SVD via fast alternating least squares." *J. Mach. Learn. Res.*, vol. 16, pp. 3367–3402, 2015.

[5] R. Sun and Z. Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.

[6] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, Australia, Aug. 2017, pp. 1233–1242.

[7] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3614–3628, Jul. 2018.

[8] Q. Li, Z. Zhu, and G. Tang, "The non-convex geometry of low-rank matrix optimization," *Inf. Inference: A J. IMA*, 2018, Art. no. iay003. [Online]. Available: https://doi.org/10.1093/imaiai/iay003

[9] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur, "Two algorithms for orthogonal nonnegative matrix factorization with application to clustering," *Neurocomputing*, vol. 141, pp. 15–25, 2014.

[10] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Math. Program. Comput.*, vol. 4, pp. 333–361, 2012.

[11] N. Srebro and A. Shraibman, "Rank, trace-norm and max-norm," in *COLT*, vol. 5. New York, NY, USA: Springer, 2005, pp. 545–560.

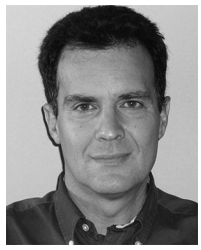
[12] F. Shang, Y. Liu, and J. Cheng, "Tractable and scalable schatten quasi-norm approximations for rank minimization," in *Proc. Artif. Intell. Statist.*, 2016, pp. 620–629.

[13] F. Shang, Y. Liu, and J. Cheng, "Scalable algorithms for tractable schatten quasi-norm minimization," in *Proc. Nat. Conf. Artif. Intell.*, 2016, pp. 2016–2022.

[14] F. Shang, J. Cheng, Y. Liu, Z.-Q. Luo, and Z. Lin, "Bilinear factor matrix norm minimization for robust PCA: Algorithms and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2066–2080, Sep. 2018.
[15] B. Haeffele, E. Young, and R. Vidal, "Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 2007–2015.
[16] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.
[17] M. Fornasier, H. Rauhut, and R. Ward, "Low-rank matrix recovery via iteratively reweighted least squares minimization," *SIAM J. Optim.*, vol. 21, no. 4, pp. 1614–1640, 2011.
[18] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Mach. Learn. Res.*, vol. 13, pp. 3441–3473, 2012.
[19] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 4130–4137.
[20] P. V. Giampouras, A. A. Rontogiannis, and K. D. Koutroumbas, "Online low-rank subspace learning from incomplete data using rank revealing ℓ_2/ℓ_1 regularization," in *Proc. IEEE Statist. Signal Process. Workshop*, Jun. 2016, pp. 1–5.
[21] P. V. Giampouras, A. A. Rontogiannis, and K. D. Koutroumbas, "ℓ_1/ℓ_2 regularized non-convex low-rank matrix factorization," in *Proc. Signal Process. With Adaptive Sparse Structured Representations*, Jun. 2017.
[22] 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 Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.
[23] F. Nie, H. Huang, and C. Ding, "Low-rank matrix recovery via efficient Schatten-p norm minimization," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 655–661.
[24] S. Squires, A. Prügel-Bennett, and M. Niranjan, "Rank selection in nonnegative matrix factorization using minimum description length," *Neural Comput.*, vol. 29, no. 8, pp. 2164–2176, 2017.
[25] F. Shang, Y. Liu, and J. Cheng, "Unified scalable equivalent formulations for Schatten quasi-norms," CUHK Tech. Rep. CSE-ShangLC20160307, Mar. 7, 2016.
[26] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. Signal Process. Adaptive Sparse Structured Representations*, 2009.
[27] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1592–1605, Jul. 2013.
[28] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.
[29] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.
[30] A. Beck, "On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes," *SIAM J. Optim.*, vol. 25, no. 1, pp. 185–209, 2015.
[31] C. Lu, J. Tang, S. Yan, and Z. Lin, "Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 829–839, Feb. 2016.
[32] P. V. Giampouras, K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Simultaneously sparse and low-rank abundance matrix estimation for hyperspectral image unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4775–4789, Aug. 2016.
[33] S. Nakajima and M. Sugiyama, "Theoretical analysis of Bayesian matrix factorization," *J. Mach. Learn. Res.*, vol. 12, no. Sep, pp. 2583–2648, 2011.

Paris V. Giampouras was born in Athens, Greece, in 1986. In 2011, he received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, and the M.Sc degree in information technologies in medicine and biology in 2014 from the Department of Informatics, National and Kapodistrian University of Athens, Athens, Greece, where he received the Ph.D. degree in 2018.

His main research interests include the areas of signal processing and machine learning focusing on sparse and low-rank representations of large-scale data and their application to collaborative filtering and hyperspectral image processing.



Athanasios A. Rontogiannis (SM9'2–M'97) received the (5-yr) Diploma degree in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1991, the M.A.Sc. degree in electrical and computer engineering from the University of Victoria, Victoria, BC, Canada, in 1993, and the Ph.D. degree in signal processing and communications from the National and Kapodistrian University of Athens, Athens, Greece, in 1997.

From 1998 to 2003, he was a Lecturer with the University of Ioannina, Ioannina, Greece. In 2003 he joined the National Observatory of Athens, where he is currently a Research Director with the Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing. His research interests include the general area of statistical signal and image processing with emphasis on adaptive signal processing, hyperspectral image processing, compressive sensing, sparse and low-rank signal representations, and fast signal processing algorithms. On these topics, he has co-authored more than 100 articles in refereed journals and conference proceedings.

Dr. Rontogiannis has been a program committee member in more than 25 conferences, in one of them as Co-Chair and in three of them as Area Chair. He has served at the Editorial Boards of the *EURASIP Journal on Advances in Signal Processing*, Springer (2008–2017), the *EURASIP Signal Processing Journal*, and Elsevier (since 2011). Since 2017, he has been serving as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is a member of the EURASIP and the Technical Chamber of Greece. Since January 2018, he has also been the Chair of the IEEE Signal Processing Society Greece Chapter.

Konstantinos D. Koutroumbas received the Diploma degree from the University of Patras, Patra, Greece, in 1989, the M.Sc. degree in advanced methods in computer science from the Queen Mary College, University of London, London, U.K., in 1990, and the Ph.D. degree from the University of Athens, Athens, Greece, in 1995.

Since 2001, he has been with the Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece, where currently he is a Research Director. His research interests include mainly pattern recognition, time series estimation, and their application (a) to remote sensing and (b) to the estimation of characteristic quantities of the upper atmosphere. He is a coauthor of the books *Pattern Recognition* (1st, 2nd, 3rd, 4th editions, Academic, 2008) and *Introduction to Pattern Recognition: A MATLAB Approach* (Academic, 2010). He has more 5000 citations in his work.