A PROJECTED NEWTON-TYPE ALGORITHM FOR NONNEGATIVE MATRIX FACTORIZATION WITH MODEL ORDER SELECTION

Paris V. Giampouras, Athanasios A. Rontogiannis, Konstantinos D. Koutroumbas

IAASARS, National Observatory of Athens, 15236, Penteli, Greece

ABSTRACT

Nonnegative matrix factorization (NMF) has attracted considerable attention over the past few years as is met in many modern machine learning applications. NMF presents some inherent challenges when it comes both to its theoretical understanding and the task of devising efficient algorithmic tools. In this paper, we deal with an issue that is inherent in NMF, i.e., the a priori unawareness of the true nonnegative rank. To this end, a novel constrained NMF formulation is proposed. The main premise of the new formulation is to first assume an overestimate of the rank and then reduce it by imposing column sparsity jointly on the nonnegative matrix factors using proper penalization. Borrowing ideas from the block successive upper bound minimization framework, an alternating minimization strategy is followed, while inexact projected Newton-type updates are used in order to guarantee the descent direction of the cost function at each iteration. The effectiveness of the proposed approach is verified on simulated data and a real music signal decomposition experiment.

Index Terms— NMF, nonnegative rank, column sparsity, projected Newton, BSUM

1. INTRODUCTION

Nonnegative matrix factorization (NMF) calls for analyzing a nonnegative matrix $\mathbf{X} \in \mathcal{R}_+^{m\times n}$ into a product of two smaller size matrices $\mathbf{W} \in \mathcal{R}_+^{m\times r}$ and $\mathbf{H} \in \mathcal{R}_+^{n\times r}$ with $r \leq \min(m,n)$, i.e., $\mathbf{X} = \mathbf{W}\mathbf{H}^T$. NMF holds a prominent position in the fields of machine learning and signal processing over the last two decades since it finds itself in a wide range of applications such as hyperspectral unmixing, topic modelling, face recognition, clustering, etc.

NMF is NP-hard, [1], and for this reason it has attracted considerable attention in terms of its theoretical understanding. Among the various aspects that have been recently addressed in the literature, a large number of works investigate the conditions that guarantee identifiability (see Definition 1 in [2]) of the matrix factors. Identifiability conditions are of significant importance in many applications of NMF such as blind source separation. This is so since the uniqueness (up to scaling and permutation) of the optimal matrix factors obtained by solving the relevant formulations of NMF problem, is a prerequisite for a reliable physical intepretation of these matrix factors. Moreover, a great amount of research has been devoted to the development of computationally efficient algorithms that approximately solve various NMF formulations, [3]. In the vast majority of these works, it is assumed that the inner dimension r, i.e., the number of rank one terms needed for the exact reconstruction of $\mathbf{X}$ (termed nonnegative rank of $\mathbf{X}$) is known beforehand. However, in real life applications a priori knowledge of the nonnegative rank r is not available. The recovery of the true nonnegative rank in NMF problems can be cast as a *model order selection problem*. However, since the number of the unknown parameters in NMF ($\mathcal{O}((m+n)r)$) is related to the size of the dataset, traditional information-theoretic based criteria, such as the Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC) cannot be applied, [4].

In this paper, we adopt the Euclidean distance as a data fitting term and propose a novel constrained NMF formulation which accounts for the unawareness of r. Concretely, we put forth a novel and generic regularization term consisting of the sum of the weighted squared Frobenius norms of the matrix factors. This term is then specified by selecting appropriate diagonal weight matrices. By doing so, column sparsity is promoted on both matrix factors $\mathbf{W}$ and $\mathbf{H}$ and thus the initially overstated order of the NMF is gradually being reduced. Following the block successive upper bound minimization (BSUM) philosophy, we address the resulting optimization problem by alternatingly using inexact projected Newton type updates for the matrix factors. Experimental results obtained in both a simulated data and a real music signal decomposition experiment empirically verify the merits of the proposed approach in simultaneously performing NMF and unveiling the true r.

We acknowledge support of this work 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 f 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).

2. RELATION TO PRIOR WORK

The problem of recovering the true model order (nonnegative rank) r in NMF has been addressed mainly within a Bayesian framework and in particular by applying ideas stemming from automatic relevance determination (ARD). In that respect, by assuming a safe overestimate of the true r, sparsity promoting prior distributions have been placed on $\mathbf{W}$ and $\mathbf{H}$, that unveil the true rank by zeroing columns of these matrix factors, [5]. Our work is related mostly with that in [4]. In [4], the generic beta-divergence is used as the data fitting term and NMF is addressed via a maximum a posteriori probability (MAP) type approach. The resulting ARD-NMF algorithm arises by solving a MAP optimization problem. Column sparsity is promoted by applying the logarithm function on either the ℓ_2 or ℓ_1 norms of the columns of the matrix, which is formed by concatenating the matrix factors. Contrary, in our work column sparsity is imposed via the use of the ℓ_1/ℓ_2 norm. In addition, the derived algorithm is based on inexact projected Newton-type updates instead of the multiplicative-type updates utilized in [4]. The proposed algorithm, is actually of similar logic to PNMF proposed in [6]. However in [6], the authors assume that r is known beforehand, while exact Hessians are used. As a consequence, full-rank matrix factors are required. Contrary, in our algorithm the full-rank condition is relaxed, while the use of approximate Hessians help us enjoy this merit without any increase of the computational complexity of the derived algorithm, as compared to PNMF.

3. PROPOSED FORMULATION

In this paper, nonnegative matrix factorization is formulated as follows

$$\min_{\{\mathbf{W}\geq\mathbf{0},\mathbf{H}\geq\mathbf{0}\}} \ \frac{1}{2}\|\mathbf{X}-\mathbf{W}\mathbf{H}^T\|_F^2 + \lambda\left(\|\mathbf{W}\mathbf{D}^{\frac{1}{2}}\|_F^2 + \|\mathbf{H}\mathbf{D}^{\frac{1}{2}}\|_F^2\right) \tag{1}$$

where $\mathbf{X}\in\mathcal{R}_+^{m\times n}$, $\mathbf{W}\in\mathcal{R}_+^{m\times d}$, $\mathbf{H}\in\mathcal{R}_+^{n\times d}$, $d\geq r$ and λ is the regularization parameter. For $\mathbf{D}=\mathbf{I}_d$, (1) boils down to the constrained NMF formulation proposed in [7] for enforcing smoothness on the matrix factors.

Next we set the weight matrix as

$$\mathbf{D} = \mathrm{diag}\Big(\left(\|\boldsymbol{w}_1\|_2^2 + \|\boldsymbol{h}_1\|_2^2\right)^{-\frac{1}{2}}, \left(\|\boldsymbol{w}_2\|_2^2 + \|\boldsymbol{h}_2\|_2^2\right)^{-\frac{1}{2}}, \ldots, \left(\|\boldsymbol{w}_d\|_2^2 + \|\boldsymbol{h}_d\|_2^2\right)^{-\frac{1}{2}} \Big), \tag{2}$$

where $\boldsymbol{w}_i, \boldsymbol{h}_i, i=1,2,\ldots,d$ denote the columns of $\mathbf{W}$ and $\mathbf{H}$, respectively. Problem (1) is hence transformed to the following optimization scheme

$$\min_{\{\mathbf{W}\geq\mathbf{0},\mathbf{H}\geq\mathbf{0}\}} \ \frac{1}{2}\|\mathbf{X}-\mathbf{W}\mathbf{H}^T\|_F^2 + \lambda\sum_{i=1}^{d}\sqrt{\|\boldsymbol{w}_i\|_2^2 + \|\boldsymbol{h}_i\|_2^2}, \tag{3}$$

in which the regularization term is tantamount to applying the joint-sparsity imposing ℓ_1/ℓ_2 norm on the columns of the concatenated matrix $\left[\begin{smallmatrix}\mathbf{W}\\ \mathbf{H}\end{smallmatrix}\right]$, [8].

Remark 1. *The proposed column sparsity promoting regularizer is a) non-smooth and b) non-separable w.r.t.* $\mathbf{W}$ *and* $\mathbf{H}$.

4. MINIMIZATION ALGORITHM

First we define the following cost function

$$f(\mathbf{W},\mathbf{H}) = \frac{1}{2}\|\mathbf{X}-\mathbf{W}\mathbf{H}^T\|_F^2 + \lambda\sum_{i=1}^{d}\sqrt{\|\boldsymbol{w}_i\|_2^2 + \|\boldsymbol{h}_i\|_2^2 + \eta^2} \tag{4}$$

where η is a small constant which is added for smoothing purposes. In what follows, we present a projected Newton-type method for efficiently addressing the NMF problem arising by minimizing (4) with respect to $\mathbf{W}$ and $\mathbf{H}$. Our goal is to benefit by exploiting the curvature information of the formed cost function. However the constrained nature of NMF induces some subtleties needed to be properly handled. To this end, we next develop a minimization algorithm that alternatingly updates matrices $\mathbf{W}$ and $\mathbf{H}$ so that they a) always belong to the feasibility set and b) guarantee the descent direction of the cost function at each iteration.

More specifically, following the block successive upper bound minimization rationale, [9], in each iteration k we adopt surrogate quadratic upper-bound approximations of the cost functions $f(\mathbf{W},\mathbf{H}_k)$ and $f(\mathbf{W}_{k+1},\mathbf{H})$ for updating matrices $\mathbf{W}$ and $\mathbf{H}$, respectively. Towards this, approximate Hessian matrices of $f(\mathbf{W},\mathbf{H}_k)$ and $f(\mathbf{W}_{k+1},\mathbf{H})$ are utilized. Let us focus on the approximate Hessian of $f(\mathbf{W},\mathbf{H}_k)$ denoted as $\bar{\mathbf{Q}}_{\mathbf{W}_k}$ (The approximate Hessian $\bar{\mathbf{Q}}_{\mathbf{H}_k}$ of $f(\mathbf{W}_{k+1},\mathbf{H})$ is defined likewise). $\bar{\mathbf{Q}}_{\mathbf{W}_k}$ is defined as

$$\bar{\mathbf{Q}}_{\mathbf{W}_k} = \mathbf{I}_m \otimes \tilde{\mathbf{Q}}_{\mathbf{W}_k}, \tag{5}$$

where $\otimes$ denotes the Kronecker product operation and $\tilde{\mathbf{Q}}_{\mathbf{W}_k}$ is a $d\times d$ matrix defined as

$$\tilde{\mathbf{Q}}_{\mathbf{W}_k} = \mathbf{H}_k^T\mathbf{H}_k + \lambda\mathbf{D} \tag{6}$$

where the latest known values of $\mathbf{W}$ and $\mathbf{H}$ are used in $\mathbf{D}$.

Let us now consider the so-called set of *active constraints*, [10], defined w.r.t. each row $\mathbf{w}_i$ of $\mathbf{W}$ at iteration k as

$$\mathcal{I}_{\mathbf{w}_i}^k = \{j|0\leq w_{ij}^k \leq \epsilon^k, [\nabla_{\mathbf{W}} f(\mathbf{W}_k,\mathbf{H}_k)]_{ij} > 0\}, \tag{7}$$

where $\epsilon^k = \min(\varepsilon, \|\mathbf{W}_k - \nabla_{\mathbf{W}} f(\mathbf{W}_k,\mathbf{H}_k)\|_F^2)$ (with ε a small positive constant). A similar set $\mathcal{I}_{\mathbf{h}_i}^k$ is defined based on the rows $\mathbf{h}_i$ of matrix $\mathbf{H}$. As is analytically explained in [6], these sets contain the coordinates of the row elements of matrices $\mathbf{W}$ and $\mathbf{H}$ that belong to the boundaries of the constrained sets, and at the same time are stationary at iteration k.

That said, the quadratic surrogate functions $l(\mathbf{W}|\mathbf{W}_k, \mathbf{H}_k)$ and $g(\mathbf{H}|\mathbf{W}_{k+1}, \mathbf{H}_k)$ are defined as,

$$l(\mathbf{W}|\mathbf{W}_k, \mathbf{H}_k) = f(\mathbf{W}_k, \mathbf{H}_k) + \mathrm{tr}\{(\mathbf{W} - \mathbf{W}_k)^T \nabla_{\mathbf{W}} f(\mathbf{W}_k, \mathbf{H}_k)\} + \frac{1}{2\alpha_{\mathbf{W}}^k} \mathrm{vec}\,(\mathbf{W} - \mathbf{W}_k)^T \, \bar{\mathbf{Q}}_{\mathbf{W}}^{\mathcal{I}_{\mathbf{W}}^k} \mathrm{vec}\,(\mathbf{W} - \mathbf{W}_k) \quad (8)$$

and

$$g(\mathbf{H}|\mathbf{W}_{k+1}, \mathbf{H}_k) = f(\mathbf{W}_{k+1}, \mathbf{H}_k) + \mathrm{tr}\{(\mathbf{H} - \mathbf{H}_k)^T \nabla_{\mathbf{H}} f(\mathbf{W}_{k+1}, \mathbf{H}_k)\} + \frac{1}{2\alpha_{\mathbf{H}}^k} \mathrm{vec}\,(\mathbf{H} - \mathbf{H}_k)^T \, \bar{\mathbf{Q}}_{\mathbf{H}}^{\mathcal{I}_{\mathbf{H}}^k} \mathrm{vec}\,(\mathbf{H} - \mathbf{H}_k)\,, \quad (9)$$

where $\mathrm{vec}(\cdot)$ stands for row vectorization and $\alpha_{\mathbf{W}}^k$ and $\alpha_{\mathbf{H}}^k$ denote step size parameters. Following the projected Newton strategy which guarantees the descent of the cost function at each iteration approximate Hessian matrices denoted as $\bar{\mathbf{Q}}_{\mathbf{W}}^{\mathcal{I}_{\mathbf{W}}}$ and $\bar{\mathbf{Q}}_{\mathbf{H}}^{\mathcal{I}_{\mathbf{H}}}$ are used in (8) and (9). Both matrices are block diagonal and consist of m and n, respectively, $d \times d$ *distinct* diagonal blocks. That is to say, the ith diagonal blocks of these matrices at iteration k, namely $\tilde{\mathbf{Q}}_{\mathbf{W}}^{\mathcal{I}_{\mathbf{w}_i}^k}$ and $\tilde{\mathbf{Q}}_{\mathbf{H}}^{\mathcal{I}_{\mathbf{h}_i}^k}$, are partially diagonalized versions of the $d \times d$ matrices $\tilde{\mathbf{Q}}_{\mathbf{W}_k}$ and $\tilde{\mathbf{Q}}_{\mathbf{H}_k}$ (see (6)). More specifically,

$$[\tilde{\mathbf{Q}}_{\mathbf{W}}^{\mathcal{I}_{\mathbf{w}_i}^k}]_{pl} = \begin{cases} 0, \text{if } p \neq l, \text{and either } p \in \mathcal{I}_{\mathbf{w}_i}^k \text{ or } l \in \mathcal{I}_{\mathbf{w}_i}^k \\ [\tilde{\mathbf{Q}}_{\mathbf{W}_k}]_{pl} \text{ otherwise} \end{cases}$$

and $\tilde{\mathbf{Q}}_{\mathbf{H}}^{\mathcal{I}_{\mathbf{h}_i}^k}$ is defined similarly. $\mathbf{W}$ and $\mathbf{H}$ are updated by *inexactly* solving the following constrained minimization problems,

$$\mathbf{W}_{k+1} = \underset{\mathbf{W} \geq \mathbf{0}}{\mathrm{argmin}}\; l(\mathbf{W}|\mathbf{W}_k, \mathbf{H}_k) \quad (10)$$

$$\text{and} \quad \mathbf{H}_{k+1} = \underset{\mathbf{H} \geq \mathbf{0}}{\mathrm{argmin}}\; g(\mathbf{H}|\mathbf{W}_{k+1}, \mathbf{H}_k) \quad (11)$$

giving rise to feasible updates in the form

$$\mathrm{vec}(\mathbf{W}_{k+1}) = [\mathrm{vec}(\mathbf{W}_k) - \alpha_{\mathbf{W}}^k \left(\bar{\mathbf{Q}}_{\mathbf{W}}^{\mathcal{I}_{\mathbf{W}}^k}\right)^{-1} \mathrm{vec}(\nabla_{\mathbf{W}} f(\mathbf{W}_k, \mathbf{H}_k))]_+ \quad (12)$$

$$\mathrm{vec}(\mathbf{H}_{k+1}) = [\mathrm{vec}(\mathbf{H}_k) - \alpha_{\mathbf{H}}^k \left(\bar{\mathbf{Q}}_{\mathbf{H}}^{\mathcal{I}_{\mathbf{H}}^k}\right)^{-1} \mathrm{vec}(\nabla_{\mathbf{H}} f(\mathbf{W}_{k+1}, \mathbf{H}_k))]_+, \quad (13)$$

where $[x]_+ = \max(x, 0)$. The step size parameters $\alpha_{\mathbf{W}}^k$ and $\alpha_{\mathbf{H}}^k$ are calculated based on the Armijo rule on the projection arc, [10], with the goal of achieving sufficient decrease of the initial cost function per iteration. The resulting alternating projected Newton-type NMF algorithm is given in Algorithm 1.

```
Algorithm 1: Proposed NMF algorithm
Input: Y, λ, β_W, β_H, σ, ε = 10^-6
Initialize: k = 0, W^0, H^0, D
repeat
    Estimate the set of active constraints I_W^k
    m_k = 0
    while sufficient decrease
          based on Armijo rule is not satisfied do
        m_k = m_k + 1, α_W^k = β_W^{m_k}
    end
    Update W_{k+1} using (12)
    Estimate the set of active constraints I_H^k
    m_k = 0
    while sufficient decrease
          based on Armijo rule is not satisfied do
        m_k = m_k + 1, α_H^k = β_H^{m_k}
    end
    Update H_{k+1} using (13)
    k = k + 1
until convergence
Output: Ŵ = W_{k+1}, Ĥ = H_{k+1}
```

Remark 2. *Contrary to the projected Newton NMF method of [6], in our case the adopted approximate Hessian matrices are always positive definite for $\lambda > 0$ and hence invertible, offering stability to the derived algorithm. Moreover, since these matrices are also partially diagonal, efficient implementations can be followed for reducing the computational cost as in [6].*

5. EXPERIMENTAL RESULTS

Herein we aim at empirically verifying the merits of the proposed NMF algorithm. Towards this, we next provide a simulated data and a real music decomposition experiment. For comparison purposes the ARD-NMF algorithm of [4] and the projected Newton NMF (PNMF) algorithm of [6] are utilized. To make fair comparisons, the beta function of ARD-NMF of [4] is reduced to the squared Frobenious norm.

5.1. Simulated data experiment

In this experiment, the performance of the proposed NMF algorithm is evaluated in a simulated NMF experiment. To this end, the nonnegative entries of matrix factors $\mathbf{W}_* \in \mathcal{R}_+^{500\times 10}$ and $\mathbf{H}_* \in \mathcal{R}_+^{500\times 10}$ are sampled from a half-normal distribution of zero mean and unit variance. A nonnegative matrix $\mathbf{X}_*$ is generated as $\mathbf{X}_* = \mathbf{W}_* \mathbf{H}_*^T$ and is corrupted by additive i.i.d. noise also sampled from a half-normal distribution resulting to $\mathrm{SNR} = 20dB$. The proposed NMF and the ARD-NMF algorithm of [4] are oblivious to the true nonnegative rank $r = 10$ of $\mathbf{X}_*$ and are initialized with an overestimate of it, i.e., $d = 25$. PNMF is run for both $d = r$ and $d = 25$. As

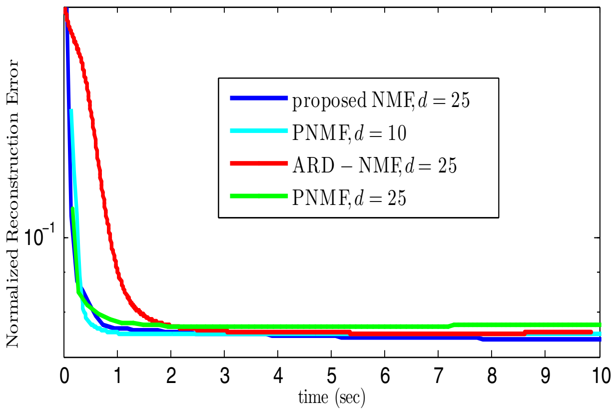

Fig. 1. Performance comparison in terms of NRE $= \frac{\|\mathbf{X}_* - \mathbf{W}\mathbf{H}^T\|_F}{\|\mathbf{X}_*\|_F}$ (average of 10 independent runs of the experiment) among the proposed NMF, the ARD-NMF and PNMF algorithms.

is shown in Fig. 1, the convergence rate of the proposed NMF algorithm is similar to PNMF even though, contrary to PNMF, it uses inexact updates for $\mathbf{W}$ and $\mathbf{H}$ per iteration. Moreover, ARD-NMF and the proposed algorithm achieve similar performance to PNMF (with $d = r$), albeit they ignore r and are intialized with an overestimate of it.

5.2. Real data experiment

Herein, we test the competence of the proposed NMF algorithm in decomposing a real music signal. For this reason, the most relevant state-of-the-art algorithm i.e., ARD-NMF is used for comparison purposes. The music signal analyzed, is a short piano sequence i.e., a monophonic 15 seconds-long signal recorded in real conditions, as described in [4]. As it can be noticed in Fig. 2, it is composed of four piano notes that overlap in all the duration thereof. Following the same process as in [4], the original signal is tranformed into the frequency domain via the short-time Fourier transform (STFT). To this end, a Hamming window of size $L = 1024$ is utilized. By appropriately setting up the overlapping between the adjacent frames we are led to a spectrogram whereby the signal is represented by 673 frames in 513 frequency bins. The power of this spectrogram is then provided as input to the tested algorithms. The initial rank is set to 20.

In Fig. 3, the first 7 components obtained by the two algorithms are ordered in decreasing values of the standard deviations of the time domain waveforms. As it can be noticed, the proposed NMF estimated the correct number of components, that is 6. Notably, the first four components of the proposed NMF correspond to the four notes while the rest two ones come from the sound of a hammer hitting the strings and the sound produced by the sustain pedal when it is released. On the contrary, ARD-NMF estimated 20 components, meaning that no rank minimization took place thus implying a data overfitting behavior. It should be emphasized that the favorable performance of the proposed NMF algorithm occurs

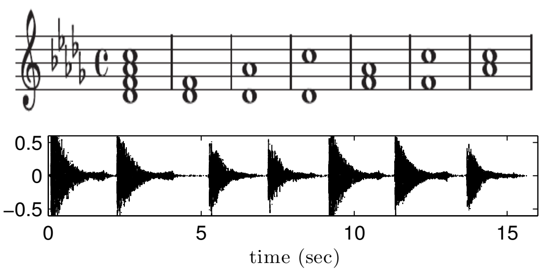

Fig. 2. Music score (top) and original audio signal (bottom)

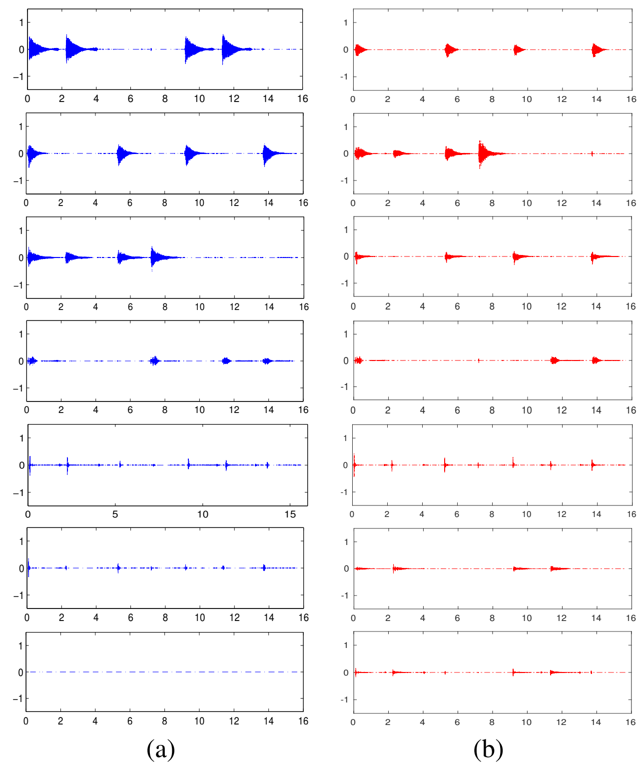

Fig. 3. Music components obtained by (a) the proposed NMF and (b) ARD-NMF on the short piano sequence.

though the noise is implicitly modeled as Gaussian i.i.d. Interestingly, as it can be seen in [4], the proposed NMF performed similarly to ARD IS-NMF, i.e., the version of ARD-NMF which makes more appropriate assumptions as to the noise statistics, by modeling it as Itakura-Saito.

6. CONCLUSIONS

In this paper a novel NMF algorithm was introduced. The main premise of the proposed approach is to simultaneously perform NMF and extract the true order of NMF. To this end, a new regularization term which imposes jointly column sparsity is utilized. The formulated optimization problem was addressed following a block successive upper bound minimization strategy coupled with inexact projected Newton-type updates. Empirical results obtained in simulated and a real music decomposition experiment illustrate the merits of the proposed approach.

7. REFERENCES

[1] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2009.

[2] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 328–332, March 2018.

[3] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, no. 257, 2014.

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

[5] J. L. Hinrich and M. Mørup, "Probabilistic sparse nonnegative matrix factorization," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 488–498.

[6] P. Gong and C. Zhang, "Efficient nonnegative matrix factorization via projected Newton method," *Pattern Recognition*, vol. 45, no. 9, pp. 3557–3565, 2012.

[7] V.P. Pauca, J. Piper, and R.J Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear algebra and its applications*, vol. 416, no. 1, pp. 29–47, 2006.

[8] P. V. Giampouras, A. A Rontogiannis, and K. D. Koutroumbas, "Alternating iteratively reweighted least squares minimization for low-rank matrix factorization," *(submitted) IEEE Transactions on Signal Processing*, 2018.

[9] 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, 2016.

[10] D. Bertsekas, "Projected Newton methods for optimization problems with simple constraints," *SIAM Journal on control and Optimization*, vol. 20, no. 2, pp. 221–246, 1982.