

# COMPRESSED NONNEGATIVE MATRIX FACTORIZATION IS FAST AND ACCURATE

Mariano Tepper, Guillermo Sapiro

**Abstract**—Nonnegative matrix factorization (NMF) has an established reputation as a useful data analysis technique in numerous applications. However, its usage in practical situations is undergoing challenges in recent years. The fundamental factor to this is the increasingly growing size of the datasets available and needed in the information sciences. To address this, in this work we propose to use structured random compression, that is, random projections that exploit the data structure, for two NMF variants: classical and separable. In separable NMF (SNMF) the left factors are a subset of the columns of the input matrix. We present suitable formulations for each problem, dealing with different representative algorithms within each one. We show that the resulting compressed techniques are faster than their uncompressed variants, vastly reduce memory demands, and do not encompass any significant deterioration in performance. The proposed structured random projections for SNMF allow to deal with arbitrarily shaped large matrices, beyond the standard limit of tall-and-skinny matrices, granting access to very efficient computations in this general setting. We accompany the algorithmic presentation with theoretical foundations and numerous and diverse examples, showing the suitability of the proposed approaches.

**Index Terms**—Nonnegative matrix factorization, separable nonnegative matrix factorization, structured random projections, big data.

## I. INTRODUCTION

The number and diversity of the fields that make use of data analysis is rapidly increasing, from economics and marketing to medicine and neuroscience. In all of them, data is being collected at an astounding speed: databases are now measured in gigabytes and terabytes, including trillions of point-of-sale transactions, worldwide social networks, and gigapixel images. Organizations need to rapidly turn these terabytes of raw data into significant insights for their users to guide their research, marketing, investment, and/or management strategies.

Matrix factorization is a fundamental data analysis technique. Whereas its usefulness as a theoretical tool is beyond doubt now, its usage in practical situations has undergone a few challenges in recent years. Among other factors contributing to this are new developments in computer hardware architecture and new applications in the information sciences.

Perhaps the key aspect is that the matrices to analyze are becoming astonishingly big. Classical algorithms are not designed to cope with the amount of information present in these large-scale problems. We may even hypothesize that, if proper tools for these problems were widely available for

This work was partially supported by NSF, ONR, NGA, ARO, and NSSEFF. The authors are with the Department of Electrical and Computer Engineering, Duke University, NC 27708 USA (e-mail: {mariano.tepper,guillermo.sapiro}@duke.edu)

commercial computer power, such rich datasets would be created at an increasing speed.

In this big data scenario, data communication is one of the main performance bottlenecks for numerical algorithms (here, we mean communication in a broad sense, including for example, network transfers and secondary memory access). Since the data cannot be easily stored in main memory, performing fewer passes over the original data, even at the cost of more floating-point operations, may result in substantially faster techniques.

Lastly, the architecture of computing units is evolving towards massive parallelism (consider, for example, general purpose GPUs and MapReduce models [1]). Numerical algorithms should adapt to these environments and exploit their benefits for boosting their performance.

In recent years, Nonnegative Matrix Factorization (NMF) [2] has been frequently used since it provides a good way for modeling many real-life applications (e.g., recommender systems [3] and audio processing [4]). NMF seeks to represent a nonnegative matrix (i.e., a matrix with nonnegative entries) as the product of two nonnegative matrices. One of the reasons for the method's popularity is that the use of non-subtractive linear combinations renders the factorization, in many cases, easily interpretable. The goal of this work is to develop algorithms, based on structured random projections, for computing NMF for big data matrices.

### A. Two flavors of nonnegative matrix factorization

Given an  $m \times n$  nonnegative matrix  $\mathbf{A}$ , NMF is formally defined as

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times r}, \mathbf{Y} \in \mathbb{R}^{r \times n}} \|\mathbf{A} - \mathbf{XY}\|_F^2 \quad \text{s.t.} \quad \mathbf{X}, \mathbf{Y} \geq 0, \quad (1)$$

where  $r$  is a parameter that controls the size of factors  $\mathbf{X}$  and  $\mathbf{Y}$  and, hence, the factorization's accuracy. For simplicity, we use  $\mathbf{B} \geq 0$  to denote a matrix  $\mathbf{B}$  with nonnegative entries.

Despite its appealing advantages, NMF does present some theoretical and practical challenges. In the general case, NMF is known to be NP-Hard [5] and highly ill-posed [6, and references therein]. However, there are matrices that exhibit a particular structure such that NMF can be solved efficiently (i.e., in polynomial time) [7].

**Definition 1.** A nonnegative matrix  $\mathbf{A}$  is  $r$ -separable if there exists an index set  $\mathcal{K}$  of cardinality  $r$  over the columns of  $\mathbf{A}$  and a nonnegative matrix  $\mathbf{Y} \in \mathbb{R}^{r \times n}$ , such that

$$\mathbf{A} = (\mathbf{A})_{:\mathcal{K}} \mathbf{Y}, \quad (2)$$

where  $(\mathbf{A})_{:\mathcal{K}}$  represents the matrix obtained by horizontally stacking the columns of  $\mathbf{A}$  indexed by  $\mathcal{K}$ . Consequently, a nonnegative matrix  $\mathbf{A}$  is near  $r$ -separable if it can be represented as

$$\mathbf{A} = (\mathbf{A})_{:\mathcal{K}} \mathbf{Y} + \mathbf{N}, \quad (3)$$

where  $\mathbf{N}$  is a noise matrix.

When  $\mathbf{A}$  presents this type of special structure, the NMF problem (now denoted as separable NMF, SNMF) can be simply modeled as

$$\min_{\substack{\mathcal{K} \subset \{1, \dots, n\} \\ \mathbf{Y} \in \mathbb{R}^{r \times n}}} \|\mathbf{A} - (\mathbf{A})_{:\mathcal{K}} \mathbf{Y}\|_F^2 \quad \text{s.t.} \quad \begin{array}{l} \#\mathcal{K} = r, \\ \mathbf{Y} \geq 0, \end{array} \quad (4)$$

where the choice of the Frobenius norm corresponds to a Gaussian noise matrix  $\mathbf{N}$ . Having a more constrained structure for the left factor (i.e.,  $\mathbf{X} = (\mathbf{A})_{:\mathcal{K}}$ ) makes the problem significantly easier to solve, improving the stability and the speed of the involved algorithms.

### B. Structured random projections

In recent years, we have seen an increase in the popularity of randomized algorithms for computing partial matrix decompositions. These partial decompositions assume that most of the action of a matrix occurs in a subspace. The key observation here is that such a subspace can be identified through random sampling. After projecting the input matrix into this subspace (i.e., compressing it), the desired low-rank factorization can be obtained by manipulating deterministically this compressed matrix. In many cases, this approach outperforms its classical competitors in terms of accuracy, speed, and robustness. See [8] for a thorough review of these techniques.

### C. Contributions and organization

We propose an algorithmic solution for computing structured random projections of extremely large matrices (i.e., matrices so large that even after compression they do not fit in main memory). This is useful as a general tool for computing many different matrix decompositions (beyond NMF, which is the particular focus of this work). Our approach leads to the implementation of compression algorithms that perform out-of-core computations (i.e., loading information in main memory only as needed).

We propose to use structured random projections for NMF and show that, in practice, their use implies a substantial increase in speed. This performance boost does not come at the price of significant errors with respect to the uncompressed solutions. We show this for representative algorithms of different NMF approaches, namely, multiplicative updates [9], active set method for nonnegative least squares [10], and ADMM [11].

We present a general SNMF algorithm based on structured random projections, reaching to similar conclusions as in the general NMF case. While there are in the literature very efficient SNMF algorithms for tall-and-skinny matrices [12], we show that, when the rank of the desired decomposition is lower than the number of columns of the input matrix, the proposed algorithm is substantially faster than its competitors.

Interestingly, the use of structured random projections allows to compute SNMF for arbitrarily large matrices, eliminating the tall-and-skinny requirement while preserving efficiency. Our code is available at <http://www.marianotepper.com.ar/research/cnmf>.

The remainder of the paper is organized as follows. In Section II we provide an overview of random projection methods for matrix factorization and provide some theoretical results relevant to this work. In sections III and IV we propose a set of techniques for using random projections for NMF and SNMF, respectively. Extensive experimental results on diverse problems are presented in Section V, studying the performance of the proposed techniques on both medium and large-scale problems. Finally, we provide some concluding remarks in Section VI.

## II. ON RANDOMIZATION AND MATRIX DECOMPOSITIONS

In this section we begin by describing the random projection algorithm used throughout this work. We also present theory that provides some guarantees for the use of random projections in matrix decomposition (in this work we use interchangeably projection or compression). Finally, we discuss the performance limits of the algorithm when dealing with big data and introduce a way to overcome such limitations.

In problems (1) and (4), the rank of the desired matrix factorization is prespecified. In the following, we will thus assume that we are given a matrix  $\mathbf{A}$ , a target rank  $r$ , and an oversampling parameter  $r_{\text{ov}}$  (its role will become clear next).

We define a Gaussian random matrix  $\Omega$  as a matrix whose entries are drawn independently from a standard Gaussian distribution, i.e., each entry  $(\Omega)_{ij}$  is a realization of an independent and identically distributed random variable with distribution  $\mathcal{N}(0, 1)$ .

The overall approach to matrix factorization presented in [8] consists of the following three steps:

- 1) Compute an approximate basis for the range of the input matrix  $\mathbf{A}$ : we construct a matrix  $\mathbf{Q}$ , with  $r + r_{\text{ov}}$  orthonormal columns (i.e.,  $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ , where  $\mathbf{I}$  is the  $(r + r_{\text{ov}}) \times (r + r_{\text{ov}})$  identity matrix), for which

$$\|\mathbf{A} - \mathbf{Q} \mathbf{Q}^T \mathbf{A}\|_2 \approx \min_{\text{rank}(\mathbf{Z}) \leq r} \|\mathbf{A} - \mathbf{Z}\|_2 = \sigma_{r+1}, \quad (5)$$

where  $\sigma_j$  denotes the  $j$ -th largest singular value of  $\mathbf{A}$ . In other words,  $\mathbf{Q} \mathbf{Q}^T \mathbf{A}$  is a good rank- $r$  approximation of  $\mathbf{A}$ .

- 2) Compute a factorization of  $\mathbf{Q}^T \mathbf{A}$ .
- 3) Multiply the leftmost factor of the decomposition by  $\mathbf{Q}$ , all other factors remain unchanged.

Throughout this paper, we will use the algorithm in Fig. 1 for performing Step (5). For more details about this algorithm, we refer the reader to [8]. Since the algorithm exploits the structure in  $\mathbf{A}$ , trying to find a subspace were the majority of its action happens, we will refer to this technique as *structured random compression*.

In the following, we present some results from [8] that demonstrate the nice theoretical characteristics of the compression matrix  $\mathbf{Q}$ , obtained with the algorithm in Fig. 1. Let  $\mathbb{E}$  denote the expectation with respect to the random matrix.

---

**input** : a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , a target rank  $r \in \mathbb{N}^+$ , an oversampling parameter  $r_{\text{ov}} \in \mathbb{N}^+$  ( $r + r_{\text{ov}} \leq m$ ), an exponent  $w \in \mathbb{N}$ .  
**output**: a compression matrix  $\mathbf{Q} \in \mathbb{R}^{m \times (r+r_{\text{ov}})}$  for  $\mathbf{A}$ .

- 1 Draw a Gaussian random matrix  $\Omega_L \in \mathbb{R}^{n \times (r+r_{\text{ov}})}$ ;
- 2 Form the matrix product  

$$\mathbf{B} = (\mathbf{A}\mathbf{A}^T)^w \mathbf{A}\Omega_L;$$
- 3 Let  $\mathbf{Q}$  be an orthogonal basis for  $\mathbf{B}$ , obtained using the QR decomposition;

---

Fig. 1. Structured random compression algorithm.

**Theorem 1** ([8]). *Given a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , a target rank  $r \in \mathbb{N}^+$ , and an oversampling parameter  $r_{\text{ov}} \in \mathbb{N}^+$  ( $r + r_{\text{ov}} \leq m$ ), execute the algorithm in Fig. 1 with  $w = 0$  (no power iterations). We obtain a matrix  $\mathbf{Q} \in \mathbb{R}^{m \times (r+r_{\text{ov}})}$ . Let  $\mathbf{P} = \mathbf{Q}\mathbf{Q}^T$ . Then,*

$$\mathbb{E} \|\mathbf{A} - \mathbf{PA}\|_F \leq \left(1 + \frac{r}{r_{\text{ov}}-1}\right)^{1/2} \left(\sum_{j>r} \sigma_j^2\right)^{1/2}, \quad (6)$$

$$\mathbb{E} \|\mathbf{A} - \mathbf{PA}\|_2 \leq \left[1 + \frac{4\sqrt{r+r_{\text{ov}}}}{r_{\text{ov}}-1} \sqrt{\min\{m, n\}}\right] \sigma_{r+1}. \quad (7)$$

Note that  $\left(\sum_{j>r} \sigma_j^2\right)^{1/2}$  and  $\sigma_{r+1}$  are the smallest possible errors, see Equation (5).

**Theorem 2** ([8]). *Frame the same hypotheses of Theorem 1. Assume  $r_{\text{ov}} \geq 4$ . Then,  $\forall u, t \geq 1$ ,*

$$\begin{aligned} \|\mathbf{A} - \mathbf{PA}\|_F &\leq \left(1 + t\sqrt{12r/r_{\text{ov}}}\right)^{1/2} \left(\sum_{j>r} \sigma_j^2\right)^{1/2} + \\ &\quad ut \frac{e\sqrt{r+r_{\text{ov}}}}{r_{\text{ov}}+1} \sigma_{r+1}, \end{aligned} \quad (8)$$

with failure probability at most  $5t^{-r_{\text{ov}}} + 2e^{-u^2/2}$ . We also have

$$\begin{aligned} \|\mathbf{A} - \mathbf{PA}\|_2 &\leq 3\sqrt{r+r_{\text{ov}}} \left(\sum_{j>r} \sigma_j^2\right)^{1/2} + \\ &\quad \left(1 + t\sqrt{8(r+r_{\text{ov}})r_{\text{ov}} \log r_{\text{ov}}}\right) \sigma_{r+1}, \end{aligned} \quad (9)$$

with failure probability at most  $6(r_{\text{ov}})^{-r_{\text{ov}}}$ .

Beyond proving that the achieved error is very close to the optimal error, the above theorems provide a theoretical justification for the oversampling parameter  $r_{\text{ov}}$ . It grants more freedom in the choice of  $\mathbf{Q}$ , crucial in the effectiveness of Step (2) [8]. This freedom allows the probability of failure to decrease exponentially fast as  $r_{\text{ov}}$  grows.

**Theorem 3** ([8]). *Given a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , a target rank  $r \in \mathbb{N}^+$ , an oversampling parameter  $r_{\text{ov}} \in \mathbb{N}^+$  ( $r + r_{\text{ov}} \leq m$ ), and an exponent  $w \in \mathbb{N}$ , execute the algorithm in Fig. 1. We obtain a matrix  $\mathbf{Q} \in \mathbb{R}^{m \times (r+r_{\text{ov}})}$ . Let  $\mathbf{P} = \mathbf{Q}\mathbf{Q}^T$ . Then,*

$$\mathbb{E} \|\mathbf{A} - \mathbf{PA}\|_2 \leq c^{1/(2w+1)} \sigma_{r+1}, \quad (10)$$

where

$$c = 1 + \sqrt{\frac{r}{(r_{\text{ov}}+1)}} + \frac{e\sqrt{r+r_{\text{ov}}}}{r_{\text{ov}}} \sqrt{\min\{m, n\} - k}. \quad (11)$$

As we increase the exponent  $w$ , the power scheme drives the extra factor in the error to one exponentially fast. As noted in [8], finding an analogous bound for the Frobenius norm is still an open problem.

Throughout this work we use a Gaussian test matrix  $\Omega$ . Other alternative test matrices can be used in its place, such as the subsampled randomized Hadamard and Fourier transforms [8, 13]. The product  $\mathbf{A}\Omega$  can be significantly faster when using a test matrix obtained with these transforms, giving an automatic speedup. From this perspective, all the experimental results in this paper present a *worst case* scenario with respect to running times.

*Note.* An alternative to structured random compression would be to just left-multiply  $\mathbf{A}$  by a Gaussian random matrix  $\Omega$ . Let us define the compression matrix  $\mathbf{Q}_\Omega \in \mathbb{R}^{m \times s}$  as

$$\mathbf{Q}_\Omega = s^{-1/2} \Omega, \quad (12)$$

where  $\Omega$  is a Gaussian random matrix. Then, instead of computing measures with the data matrix  $\mathbf{A}$  on the  $m$ -dimensional space, the much smaller matrix  $\mathbf{Q}_\Omega^T \mathbf{A}$  can be used to compute approximations in the  $s$ -dimensional space. It is well studied that Gaussian projection preserves the  $\ell_2$  norm [e.g., 14, and references therein]. However, our extensive experiments show that structured random compression achieves better performance than Gaussian compression. Intuitively, Gaussian compression is a general data-agnostic tool, whereas structured compression uses information from the matrix (an analogous of training). Theoretical research is needed to fully justify this performance gap.

#### A. Big data algorithmic solutions

By design, the product in line 2 of the algorithm in Fig. 1 forms a tall and skinny matrix  $\mathbf{B} \in \mathbb{R}^{m \times (r+r_{\text{ov}})}$ , where  $m \gg (r + r_{\text{ov}})$ . We have thus successfully reduced the number of columns in  $\mathbf{B}$  from  $n$  to  $r + r_{\text{ov}}$ . While matrix  $\mathbf{A}$  may not fit in main memory, we can still perform the necessary computations using  $\mathbf{B}$  without significant loss of precision.

An interesting question arises when working with large matrices: what happens if the number of rows  $m$  is so large that even  $\mathbf{B}$  does not fit in main memory? Assuming that we need to store  $\mathbf{B}$  in secondary memory (i.e., the hard drive), how do we compute its QR decomposition (line 3 of the algorithm in Fig. 1)?

A suitable and efficient algorithm to address the latter question is the direct TSQR (tall-and-skinny QR) [12]. For completeness, we give its outline in Appendix A. The highlight of TSQR is that it is designed for being parallelizable while minimizing the dependencies between parallel computations (i.e., communication costs). Thus, it adheres perfectly to the main mantra of this work.

An interesting byproduct of using TSQR is that there is no need to form the entire matrix  $\mathbf{B}$  in main memory. See Appendix A for further details. This allows to implement an *out-of-core* version of the compression algorithm, that is, where the involved matrices do not reside in main memory.

Let us note that the use of TSQR for computing random compression is introduced in this paper for the first time,

providing a *true* scalable solution for computing many types of matrix decompositions (i.e., beyond NMF) when both the number of rows and columns of the input matrix are large.

### B. Matrix decompositions with alternative norms

The algorithm in Fig. 4 works under the Frobenius and nuclear norms, as detailed in the theorems presented above. These two cases already cover a significant range of matrix decompositions that are commonly used in practice.

However, other norms are becoming increasingly popular in recent years. For example, NMF is widely used in audio processing with the Itakura-Saito distance instead of the Frobenius norm in Problem (1). The entrywise  $\ell_1$  norm is also very popular when the input matrix  $\mathbf{A}$  is contaminated with impulsive noise. In these cases, proper structured random projection algorithms need to be used, adapted to the right type of measure for the application at hand.

In particular, we are currently investigating the use of the framework here developed for NMF under an  $\ell_1$  norm. In such a case, the fast Cauchy transform appears as a suitable alternative for the task [15].

## III. RANDOMLY COMPRESSED NMF

The goal of this section is to efficiently solve Problem (1) for large input matrices. We do not aim at developing a new NMF algorithm, but rather to illustrate how structured random projections can be used to enhance the speed of existing algorithms and make them usable for big data. As detailed in Section V, this speedup does not come at the price of significantly higher reconstruction errors.

Most NMF algorithms work by iterating the following two steps:

- Find  $\mathbf{X}_{k+1} \in \mathbb{R}^{m \times r}$ ,  $\mathbf{X}_{k+1} \geq 0$ , such that

$$\|\mathbf{A} - \mathbf{X}_{k+1}\mathbf{Y}_k\|_F^2 \leq \|\mathbf{A} - \mathbf{X}_k\mathbf{Y}_k\|_F^2. \quad (13a)$$

- Find  $\mathbf{Y}_{k+1} \in \mathbb{R}^{r \times n}$ ,  $\mathbf{Y}_{k+1} \geq 0$ , such that

$$\|\mathbf{A} - \mathbf{X}_{k+1}\mathbf{Y}_{k+1}\|_F^2 \leq \|\mathbf{A} - \mathbf{X}_{k+1}\mathbf{Y}_k\|_F^2. \quad (13b)$$

This general formulation encompasses different particular algorithms such as multiplicative updates [9] and several variants of alternating nonnegative least squares [10, 16, 17]. The latter consists of a particular case of Algorithm (13b) in which its right-hand sides are minimized to the end. We thus obtain the following algorithm:

$$\mathbf{X}_{k+1} = \underset{\mathbf{X} \in \mathbb{R}^{m \times r}}{\operatorname{argmin}} \|\mathbf{A} - \mathbf{X}\mathbf{Y}_k\|_F^2 \quad \text{s.t. } \mathbf{X} \geq 0, \quad (14a)$$

$$\mathbf{Y}_{k+1} = \underset{\mathbf{Y} \in \mathbb{R}^{r \times n}}{\operatorname{argmin}} \|\mathbf{A} - \mathbf{X}_{k+1}\mathbf{Y}\|_F^2 \quad \text{s.t. } \mathbf{Y} \geq 0. \quad (14b)$$

Let us assume that we apply the algorithm in Fig. 1 to  $\mathbf{A}$  and  $\mathbf{A}^T$  and obtain two matrices  $\mathbf{L} \in \mathbb{R}^{m \times (r+r_{ov})}$ ,  $\mathbf{R} \in \mathbb{R}^{(r+r_{ov}) \times n}$ , respectively. By construction,  $\mathbf{L}$  and  $\mathbf{R}$  have orthonormal columns and rows, respectively. Also let  $\check{\mathbf{A}} = \mathbf{A}\mathbf{R}^T$ ,  $\hat{\mathbf{A}} = \mathbf{L}^T\mathbf{A}$ .

Using matrices  $\mathbf{L}$  and  $\mathbf{R}$ , we propose to approximate Algorithm (13b) with the iterations

---

```

input : a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , a target rank  $r \in \mathbb{N}^+$ , an oversampling parameter  $r_{ov} \in \mathbb{N}^+$  ( $r + r_{ov} \leq \min\{m, n\}$ ), an exponent  $w \in \mathbb{N}$ .
output: nonnegative matrices  $\mathbf{X}_k \in \mathbb{R}^{m \times r}$ ,  $\mathbf{Y}_k \in \mathbb{R}^{r \times n}$ .
1 Compute compression matrices  $\mathbf{L} \in \mathbb{R}^{m \times (r+r_{ov})}$ ,  $\mathbf{R} \in \mathbb{R}^{(r+r_{ov}) \times n}$ ;
2  $k \leftarrow 1$ ;
3 Initialize  $\mathbf{Y}_k$ ;
4 repeat
5    $\check{\mathbf{Y}}_k \leftarrow \mathbf{Y}_k\mathbf{R}^T$ ;
6   Find  $\mathbf{X}_{k+1} \in \mathbb{R}^{m \times r}$ ,  $\mathbf{X}_{k+1} \geq 0$ , such that
      
$$\|\check{\mathbf{A}} - \mathbf{X}_{k+1}\check{\mathbf{Y}}_k\|_F^2 \leq \|\check{\mathbf{A}} - \mathbf{X}_k\check{\mathbf{Y}}_k\|_F^2;$$

7    $\hat{\mathbf{X}}_{k+1} \leftarrow \mathbf{L}^T\mathbf{X}_{k+1}$ ;
8   Find  $\mathbf{Y}_{k+1} \in \mathbb{R}^{r \times n}$ ,  $\mathbf{Y}_{k+1} \geq 0$ , such that
      
$$\|\hat{\mathbf{A}} - \hat{\mathbf{X}}_{k+1}\mathbf{Y}_{k+1}\|_F^2 \leq \|\hat{\mathbf{A}} - \hat{\mathbf{X}}_k\mathbf{Y}_k\|_F^2;$$

// The optimizations in lines 6 and 8 can be performed using any variant of multiplicative updates or any nonnegative least squares method.
9    $k \leftarrow k + 1$ ;
10 until convergence;

```

---

Fig. 2. NMF using structured random compression.

- Find  $\mathbf{X}_{k+1} \in \mathbb{R}^{m \times r}$ ,  $\mathbf{X}_{k+1} \geq 0$ , such that

$$\|\check{\mathbf{A}} - \mathbf{X}_{k+1}\mathbf{Y}_k\mathbf{R}^T\|_F^2 \leq \|\check{\mathbf{A}} - \mathbf{X}_k\mathbf{Y}_k\mathbf{R}^T\|_F^2. \quad (15a)$$

- Find  $\mathbf{Y}_{k+1} \in \mathbb{R}^{r \times n}$ ,  $\mathbf{Y}_{k+1} \geq 0$ , such that

$$\|\hat{\mathbf{A}} - \mathbf{L}^T\mathbf{X}_{k+1}\mathbf{Y}_{k+1}\|_F^2 \leq \|\hat{\mathbf{A}} - \mathbf{L}^T\mathbf{X}_{k+1}\mathbf{Y}_k\|_F^2. \quad (15b)$$

Equivalently, using  $\mathbf{L}$  and  $\mathbf{R}$ , we propose to approximate Algorithm (14b) with the iterations

$$\mathbf{X}_{k+1} = \underset{\mathbf{X} \in \mathbb{R}^{m \times r}}{\operatorname{argmin}} \|\check{\mathbf{A}} - \mathbf{X}\mathbf{Y}_k\mathbf{R}^T\|_F^2 \quad \text{s.t. } \mathbf{X} \geq 0, \quad (16a)$$

$$\mathbf{Y}_{k+1} = \underset{\mathbf{Y} \in \mathbb{R}^{r \times n}}{\operatorname{argmin}} \|\hat{\mathbf{A}} - \mathbf{L}^T\mathbf{X}_{k+1}\mathbf{Y}\|_F^2 \quad \text{s.t. } \mathbf{Y} \geq 0. \quad (16b)$$

The algorithm in Fig. 2 contains an overview of the proposed NMF algorithm using structured random compression. For our experiments regarding the techniques described in Section III, as representative examples of Algorithm (13b) and Algorithm (14b), we respectively use the active set method [10] and the multiplicative updates in [18, Eq. (8)].

We achieve a significant size reduction of the matrices in algorithms (15b) and (16b). For each of these algorithms, we reduced the number of columns from  $n$  to  $r + r_{ov}$  in equations (15a) and (16a) and the number of rows from  $m$  to  $r + r_{ov}$  in equations (15b) and (16b). This makes the system much faster to solve, but more importantly in our context, it greatly reduces the cost of data communication in parallel frameworks. For example, after compression, large matrices might fit in GPU memory.

Alternatively, Problem (1) can be equivalently re-formulated

as

$$\min_{\substack{\mathbf{X}, \mathbf{U} \in \mathbb{R}^{m \times r} \\ \mathbf{Y}, \mathbf{V} \in \mathbb{R}^{r \times n}}} \|\mathbf{A} - \mathbf{XY}\|_F^2 \quad \text{s.t.} \quad \begin{aligned} \mathbf{U} &= \mathbf{X}, \mathbf{V} = \mathbf{Y}, \\ \mathbf{U}, \mathbf{V} &\geq 0. \end{aligned} \quad (17)$$

Again, using the matrices  $\mathbf{L}$  and  $\mathbf{R}$  defined above, we propose to approximate Problem (17) with

$$\min_{\substack{\mathbf{X}, \mathbf{U} \in \mathbb{R}^{m \times r} \\ \mathbf{Y}, \mathbf{V} \in \mathbb{R}^{r \times n}}} \|\mathbf{LL}^T (\mathbf{A} - \mathbf{XY}) \mathbf{R}^T \mathbf{R}\|_F^2 \quad \text{s.t.} \quad \begin{aligned} \mathbf{U} &= \mathbf{X}, \\ \mathbf{V} &= \mathbf{Y}, \\ \mathbf{U}, \mathbf{V} &\geq 0. \end{aligned} \quad (18)$$

Let  $\tilde{\mathbf{A}} = \mathbf{L}^T \mathbf{AR}^T$ ,  $\tilde{\mathbf{X}} = \mathbf{L}^T \mathbf{X}$ , and  $\tilde{\mathbf{Y}} = \mathbf{YR}^T$ . We propose to further approximate Problem (17) with

$$\min_{\substack{\mathbf{U} \in \mathbb{R}^{m \times r}, \mathbf{V} \in \mathbb{R}^{r \times n} \\ \tilde{\mathbf{X}} \in \mathbb{R}^{(r+rov) \times r} \\ \tilde{\mathbf{Y}} \in \mathbb{R}^{r \times (r+rov)}}} \|\tilde{\mathbf{A}} - \tilde{\mathbf{X}} \tilde{\mathbf{Y}}\|_F^2 \quad \text{s.t.} \quad \begin{aligned} \mathbf{U} &= \mathbf{L} \tilde{\mathbf{X}}, \\ \mathbf{V} &= \tilde{\mathbf{Y}} \mathbf{R}, \\ \mathbf{U}, \mathbf{V} &\geq 0. \end{aligned} \quad (19)$$

The alternating direction method of multipliers (ADMM) can be used for solving Problem (17) [11]. Thus, a similar technique can solve Problem (19). The details of the proposed algorithm are presented in Appendix B.

The level of compression in Problem (19) is significantly higher than in algorithms (15b) and (16b). The latter formulations only employ (alternated) single-sided compression, whereas the former uses a (simultaneous) double-sided compression. One may be inclined to think that such an aggressive compression might lead to greater errors; however, in practice, this is not the case. Studying this behavior from a theoretical standpoint might shed light into this interesting characteristic.

#### A. Limits of NMF for big data

When matrix  $\mathbf{A}$  gets sufficiently large, solving Problem (1) becomes challenging. The compression techniques here presented significantly alleviate the problem for in-core computations and are easily extensible for out-of-core computations. For example, each iteration of the multiplicative updates algorithm can be implemented on a MapReduce framework [19]; its structured compressed version can be easily adapted in this framework, greatly reducing communication costs thanks to the use of smaller matrices. Implementing our compressed ADMM algorithm on a MapReduce framework is just as straightforward.

However, when dealing with large volumes of data, the practical problem actually resides in the iterative nature of the algorithms. As an example, consider that the execution time of a single iteration of the multiplicative algorithm on a MapReduce framework is measured in hours for *sparse* matrices with millions of columns and rows [19, 20]. As expected, the issue is hugely exacerbated for dense matrices.

## IV. RANDOMLY COMPRESSED SEPARABLE NMF

Following Definition 3, let us now assume that matrix  $\mathbf{A}$  is (near)  $r$ -separable. Most state-of-the-art techniques for computing SNMF, see Problem (4), are based on the following two-step approach:

- 1) Extract  $r$  columns of  $\mathbf{A}$ , indexed by  $\mathcal{K}$ . The literature usually refers to them as extreme columns.
- 2) Solve

$$\mathbf{Y} = \operatorname{argmin}_{\mathbf{H} \in \mathbb{R}^{r \times n}} \|\mathbf{A} - (\mathbf{A})_{:\mathcal{K}} \mathbf{H}\|_F^2 \quad \text{s.t.} \quad \mathbf{H} \geq 0. \quad (20)$$

The literature on SNMF has mainly focused on Step (1) of the above algorithm. There are several types of algorithms for performing this task [21–24]. As for Step (20), Problem (20) involves solving  $n$  nonnegative least squares problems separately, i.e.,

$$(\mathbf{Y})_{:i} = \operatorname{argmin}_{\mathbf{h} \in \mathbb{R}^m} \|\mathbf{A}_{:i} - (\mathbf{A})_{:\mathcal{K}} \mathbf{h}\|_F^2 \quad \text{s.t.} \quad \mathbf{h} \geq 0. \quad (21)$$

This makes Step (20) trivially parallelizable.

Let  $\mathbf{Q} \in \mathbb{R}^{m \times m}$  be an orthonormal basis for  $\mathbf{A} \in \mathbb{R}^{m \times n}$

$$\mathbf{Q}^T \mathbf{A} = \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix}, \quad \mathbf{Q}^T (\mathbf{A})_{:\mathcal{K}} = \begin{bmatrix} (\mathbf{R})_{:\mathcal{K}} \\ 0 \end{bmatrix}, \quad (22)$$

where  $\mathbf{R} \in \mathbb{R}^{n \times n}$ . A key observation here is that the zero rows do not provide information for finding extreme columns of  $\mathbf{A}$  [12]. We also trivially have that, for any orthonormal matrix  $\mathbf{Q} \in \mathbb{R}^{m \times m}$ ,

$$\|\mathbf{Q}^T (\mathbf{A} - \mathbf{XY})\|_F \propto \|\mathbf{A} - \mathbf{XY}\|_F. \quad (23)$$

Then,

$$\mathbf{Y} = \operatorname{argmin}_{\mathbf{H} \geq 0} \|\mathbf{A} - (\mathbf{A})_{:\mathcal{K}} \mathbf{H}\|_F^2 \quad (24a)$$

$$= \operatorname{argmin}_{\mathbf{H} \geq 0} \|\mathbf{Q}^T (\mathbf{A} - (\mathbf{A})_{:\mathcal{K}} \mathbf{H})\|_F^2 \quad (24b)$$

$$= \operatorname{argmin}_{\mathbf{H} \geq 0} \|\mathbf{R} - (\mathbf{R})_{:\mathcal{K}} \mathbf{H}\|_F^2. \quad (24c)$$

Notice that Problem (24c) has succeeded to reduce the problem size to  $n \times n$  from the original  $m \times n$  Problem (20). We then obtain the following three-step algorithm [12]:

- 1) Compute  $\mathbf{Q}$  using, e.g., a QR decomposition of  $\mathbf{A}$ .
- 2) Find  $r$  extreme columns of  $\mathbf{R} = \mathbf{Q}^T \mathbf{A}$ , indexed by  $\mathcal{K}$ .
- 3) Solve

$$\mathbf{Y} = \operatorname{argmin}_{\mathbf{H} \in \mathbb{R}^{r \times n}} \|\mathbf{R} - (\mathbf{R})_{:\mathcal{K}} \mathbf{H}\|_F^2 \quad \text{s.t.} \quad \mathbf{H} \geq 0. \quad (25)$$

As the main assumption in NMF and SNMF is that  $\mathbf{A}$  has (or can be approximated by) a low-rank structure, by all practical means we expect that  $r \ll \min(m, n)$ ; otherwise, it would not even make sense to try these type of decompositions. We claim that little to no information is lost by replacing the full orthonormal basis with a rank-preserving basis that projects the data into a lower-dimensional space.

As the reader might be already suspecting, we propose to obtain such a basis via the use structured random projections. This involves a small but conceptually important change in the above SNMF algorithm. Replace Step (1) by

- 1) Compute a structured random compression matrix  $\mathbf{Q}$  for  $\mathbf{A}$ .

The proposed algorithm is depicted in Fig. 3. Let us now detail the main differences with the QR-based algorithm.

---

**input** : a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , a target rank  $r \in \mathbb{N}^+$ , an oversampling parameter  $r_{ov} \in \mathbb{N}^+$  ( $r + r_{ov} \leq \min\{m, n\}$ ), an exponent  $w \in \mathbb{N}$ .  
**output**: index  $\mathcal{K}$  of the extreme columns of  $\mathbf{A}$ , nonnegative matrix  $\mathbf{Y} \in \mathbb{R}^{r \times n}$ .

- 1 Compute  $\mathbf{Q} \in \mathbb{R}^{m \times (r+r_{ov})}$  using the algorithm in Fig. 1;
- 2 Find  $r$  extreme columns of  $\mathbf{R} = \mathbf{Q}^T \mathbf{A}$ , indexed by  $\mathcal{K}$ ;
- 3 Solve  $\forall i \in [0, n)$

$$(\mathbf{Y})_{:i} \leftarrow \underset{\mathbf{h} \in \mathbb{R}^{r+r_{ov}}}{\operatorname{argmin}} \| \mathbf{R}_{:i} - (\mathbf{R})_{:\mathcal{K}} \mathbf{h} \|_F^2 \quad \text{s.t. } \mathbf{h} \geq 0; \quad (26)$$


---

Fig. 3. SNMF using structured random compression.

First, let us note that  $\mathbf{R} = \mathbf{Q}^T \mathbf{A}$  is now an  $(r + r_{ov}) \times n$  matrix instead of an  $n \times n$  matrix. This allows to process matrices that have many more columns, as storing  $\mathbf{R}$  has become orders of magnitude easier/cheaper. Also note that each nonnegative least squares problem in Problem (21) has also become orders of magnitude smaller and thus faster to solve. Again, the huge decrease in communication costs for parallel implementations is even more important in our context than the gain in computational speed.

Second, the computation of the basis itself has become much faster. This is easy to understand when we compare the algorithm in Fig. 1, which only computes the QR decomposition of an  $m \times (r + r_{ov})$  matrix, with the QR decomposition of the full  $m \times n$  matrix. Of course, as the ratio  $r/n$  decreases, the proposed algorithm becomes faster.

Let us assume for a moment that  $n$  is sufficiently small such that we can use the TSQR algorithm directly on the input matrix  $\mathbf{A}$ , but not trivially small. As detailed in Appendix A, the QR decomposition in Equation (30) in the appendix is the only centralized step in TSQR; the amount of information that needs to be transmitted to carry this step is, again, orders of magnitude smaller when using structured random compressions.

*Note.* The separable NMF model is similar to the model presented in [25] (and in [26] without non-negativity constraints)

$$\min_{\mathbf{T} \in \mathbb{R}^{n \times n}} \| \mathbf{A} - \mathbf{AT} \|_F^2 + \lambda \| \mathbf{T} \|_{\text{row-0}} \quad \text{s.t. } \mathbf{T} \geq 0, \quad (27)$$

where  $\| \mathbf{T} \|_{\text{row-0}}$  denotes the number of non-zero rows. The similarity resides in that selecting a subset of rows from  $\mathbf{T}$  is equivalent to selecting a subset of columns from  $\mathbf{A}$ . This problem can be relaxed into a convex problem by replacing the  $\ell_0$  pseudo-norm by a (possibly weighted)  $\ell_1$  norm. However, whichever optimization technique we choose for solving this problem, it will involve an iterative algorithm, where an  $n \times n$  system is solved in every iteration. In [25], the problem is shrunk by clustering the columns of  $\mathbf{A}$  and feeding a new matrix, only containing the cluster centers, into Equation (27). For this reasons, in our view, the SNMF model, as presented here, presents a cleaner and faster alternative to Equation (27).

## V. EXPERIMENTAL RESULTS

We will now present numerous examples supporting the use of structured random projections for NMF and SNMF, both in terms of speed and accuracy.

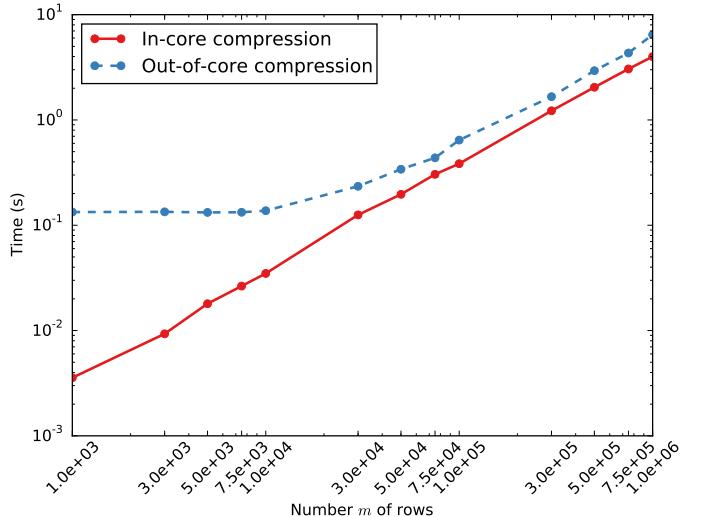


Fig. 4. **Performance of out-of-core compression.** We tested different values for  $m$ , while fixing  $n = 500$ . The out-of-core algorithm for structured random compression presented in Section II is slower for matrices with approximately less than  $2 \cdot 10^4$  rows. For larger matrices, it exhibits the same complexity as the in-core one (linear in  $m$ ). Out-of-core computations do not come at the price of a significantly slower compression algorithm, though permit to work with significantly larger matrices.

Before jumping to these problems, in Fig. 4 we show a simulation of the nice properties of the out-of-core compression algorithm presented in Section II. We performed our tests on  $m \times n$  matrices with Gaussian entries, where different values for  $m$  were tested, ranging from  $10^3$  to  $10^6$ , and  $n = 500$  in all cases. This ensures that all matrices fit in main memory, allowing (1) to compress them with the in-core algorithm, and (2) to disregard disk access times, making the comparisons fair. The out-of-core algorithm for structured random compression is slower for matrices with approximately less than  $2 \cdot 10^4$  rows; for these small matrices, the overhead of processing the matrix per blocks becomes evident (notice though that both computing times are well under 1 second). For larger matrices, the overhead's impact becomes less significant, and both algorithms exhibit the same overall performance (linear in  $m$ ). In summary, we observe the expected behavior: the greater flexibility of the proposed out-of-core compression algorithm for processing large matrices does not cause performance to degrade with respect to the in-core one.

## A. NMF

For our experiments regarding the techniques presented in Section III, as representative examples of Algorithm (13b) and Algorithm (14b), we respectively use the active set method [10] and the multiplicative updates in [18, Eq. (8)]. For these two algorithms, we compared with a vanilla version and a variant using Gaussian projection, as presented in [27] (also see Section II). We also implemented the ADMM algorithm in [11] and the proposed ADMM algorithm with structured random compression. All the methods were implemented in Matlab. In all tests, we set  $w = 4$  and  $r_{ov} = 10$  in the compression algorithm in Fig. 1; we further adjust the value of  $r_{ov}$  so that  $r + r_{ov} = \min(\max(20, r + r_{ov}), n)$ .

TABLE I. Performance when compressing a PET image. See Fig. 8 for a detailed explanation of the setup. As we can observe, the use of Gaussian compression (GC) is detrimental to the reconstruction error (higher values indicate a lower error), while the proposed structured random compression (SC) has no significant impact on it. As a counterpart, the use of SC significantly decreases the computing time with respect to the original method. The best values for each column are highlighted in green.

	Error ( $-\log_{10}$ )			Time (s)
	Mean	STD	Median	
Multiplicative	3.628	3.435	3.980	177.813
Multiplicative - GC	3.496	3.304	3.834	19.742
Multiplicative - SC	3.626	3.433	3.979	71.437
ADMM	3.638	3.436	4.027	32.195
ADMM - SC	3.636	3.433	4.028	23.168
Active set	3.628	3.436	3.976	18.251
Active set - GC	3.536	3.350	3.882	14.277
Active set - SC	3.638	3.436	4.024	11.371
ALS <sup>1</sup>	3.638	3.436	4.023	18.162
ALS with proj. grad. <sup>1</sup>	3.634	3.436	4.004	58.208

<sup>1</sup> Obtained from <http://cogsys.imm.dtu.dk/toolbox/nmf/>.

We begin by showing in Fig. 7 simulations results of the different NMF variants on synthetic examples. The first interesting observation from these examples is that, although the computation of the compression matrix is more costly for structured than for Gaussian compression, this might not end up reflected in the overall computing time; this is because, in general, the NMF variant with Gaussian compression requires more iterations to converge. The second observation is that the NMF variants that use structured compression yield very similar relative reconstruction errors than their uncompressed counterparts (higher in one example, lower in three). For multiplicative updates and ADMM, the gain in speed of using structured compression is huge; for active set, the speedup is not as dramatic. Lastly, Gaussian compression seems to come at the cost of higher reconstruction errors.

We also run different NMF algorithms on a hyperspectral positron emission tomography (PET) image, see Fig. 8. This example allows to visually compare the errors produced by the different methods. The NMF methods with Gaussian compression create “clusters” of errors (particular areas in which the errors seem to concentrate). In Table I we show several error statistics and the computing time for the different methods. The statistics also reflect the same behavior as our visual previous inspection. Structured compression has a positive effect on the computing time (it decreases), and no significant effect on the error statistics.

Climate datasets are very interesting to analyze using NMF. We believe that the evidence of a low rank model within climate data is of interest by itself. Nonnegativity is a useful addition since, under this model, the effects of different factors cannot cancel each other. The technical details and results of an experiment using climate data are shown in Fig. 9. In this case, we only use the active set method for our comparisons. We found that two factors explain the data with enough accuracy. Both factors seem to correspond to two very different seasons across the globe, and they exhibit inversely correlated periodic patterns. While the left and right factors obtained using structured compression are very similar to their uncompressed counterparts, Gaussian compression introduces

visible artifacts in the resulting factorization. Structured random compression also is the fastest of the three methods.

Our last classical NMF example consists of a popular application: biclustering. In this case, we bicluster a bipartite social network, i.e., that contains two different types of nodes. In our particular example, these two types correspond to characters from Marvel comic books and to the comic books in which they appear. We performed NMF with  $r = 10$  (recall that  $r$  is the number of factors). We then thresholded each column of  $\mathbf{X}$  and each row of  $\mathbf{Y}$  to obtain sparse components that we define as a bicluster (we could have also added a sparsity term to the formulation, but opted for a simpler approach that does not introduce additional complexity). For each column (row) of  $\mathbf{X}$  ( $\mathbf{Y}$ ), we set to zero the entries smaller than the column (row) mean plus three standard deviations. Then, for display purposes, we only keep the largest 25 entries in each column of  $\mathbf{X}$  if there are more than that number of nonzero entries. In Fig. 10 we show two of the biclusters obtained in such a way. It becomes quickly apparent that structured compression does not introduce significant artifacts in the biclusters, whereas the clusters found with Gaussian compression are heavily intertwined (all ten factors seem to be mixed together). For example, Mary Jane Parker-Watson, Spider-Man’s wife, is not a recurring character of the Fantastic Four comic books.

To summarize, the overall observation is that structured compression brings additional speed to NMF methods without introducing significant errors. On the other hand, Gaussian compression seems to come at the cost of higher reconstruction errors and is not consistently faster than structured compression.

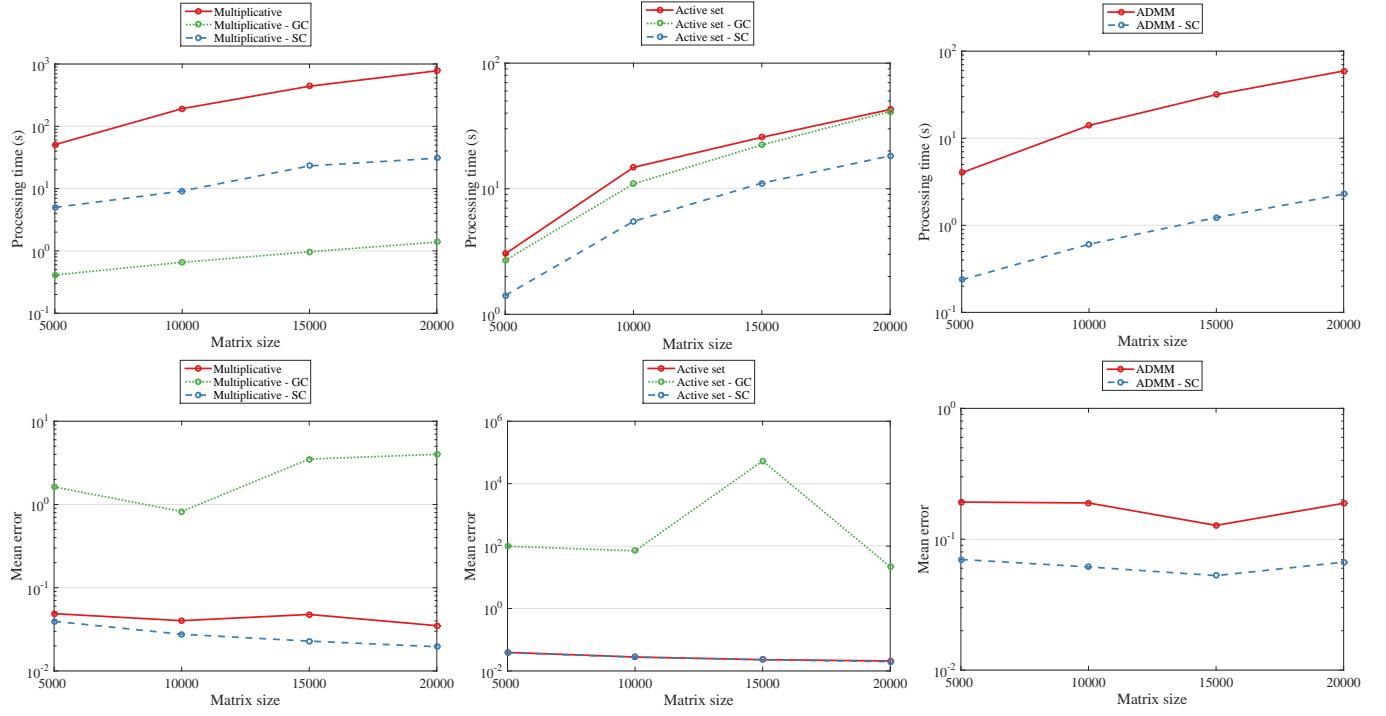
### B. Separable NMF

We implemented our SNMF algorithms in Python, using the dask and into libraries<sup>1</sup> to perform out-of-core matrix computations (i.e., without fully loading the involved matrices in main memory). A byproduct of this implementation choice is that we can compute SNMF on very large matrices on a regular laptop, without having to resort to a cluster. To the best of our knowledge, our TSQR implementation is the first publicly available one that runs on any regular laptop using out-of-core computations.

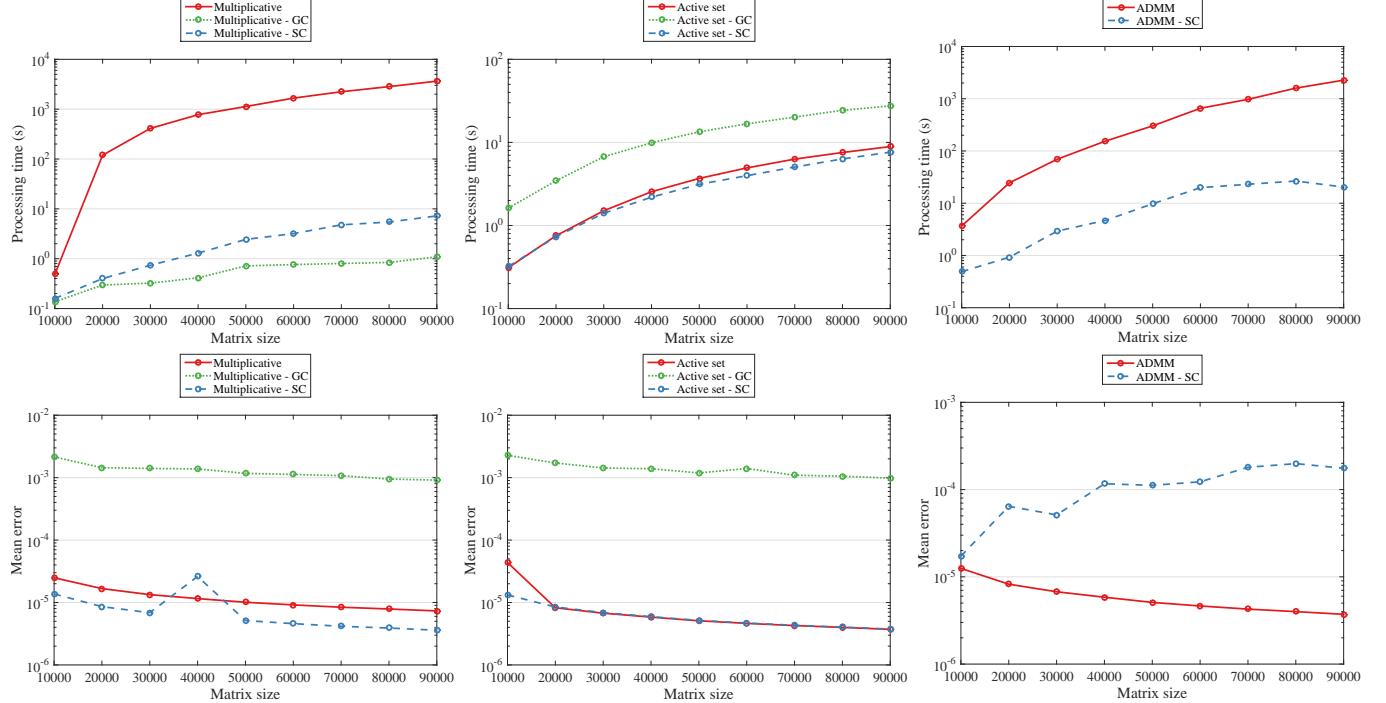
We perform all of our comparisons with the SNMF algorithm using the QR decomposition [12], analyzed in Section IV. We use SPA [21, 24], and XRAY [23] as the column selection algorithms. Throughout this section, we simply use compression to refer to structured compression. In all tests, we set  $w = 0$  and  $r_{ov} = 10$  in the compression algorithm in Fig. 1; we further adjust the value of  $r_{ov}$  so that  $r + r_{ov} = \min(\max(20, r + r_{ov}), n)$ .

We first present results on synthetic matrices in Fig. 11. We produced different matrices of fixed size by varying their rank, see Fig. 11(a). In general, we aim at explaining the data matrix with a small fraction of its columns. The proposed compression method for SNMF is faster when fewer factors are needed to explain the data. On the other hand, QR-based methods have always the same (high) computing time,

<sup>1</sup><http://dask.readthedocs.org/>, <http://into.readthedocs.org/>



(a) **Synthetic dense matrices.** The matrix size indicates the number of rows  $m$ ; the number of columns  $n$  is fixed to  $n = 0.75m$  in all cases. Since the matrices are dense,  $\delta = 1$ .



(a) **Synthetic sparse matrices.** The matrix size indicates the number of rows  $m$ ; the number of columns  $n$  and the sparsity level  $\delta$  are fixed to  $n = 0.75m$  and  $\delta = 10^{-2}$  in all cases.

Fig. 7. **Performance comparison on synthetic matrices.** We first generate two matrices  $\mathbf{X}_{\text{GT}} \in \mathbb{R}^{m \times r}$ ,  $\mathbf{Y}_{\text{GT}} \in \mathbb{R}^{r \times n}$ , where their entries are uniformly distributed in  $[0, 1]$  with probability  $\delta$ , or zero with probability  $1 - \delta$ . We then build  $\mathbf{A} = \mathbf{X}_{\text{GT}} \mathbf{Y}_{\text{GT}} + \mathbf{N}$ , where the entries of  $\mathbf{N}$  are normally distributed with probability  $\delta^2$ , or zero with probability  $1 - \delta^2$ . GC and SC stand for Gaussian and structured compression, respectively. The reconstruction error is reported as the mean over 10 different runs. While both GC and SC are generally faster than the original uncompressed methods (top row), the accuracy levels of the latter are only matched (and sometimes even outmatched) by SC (bottom row).

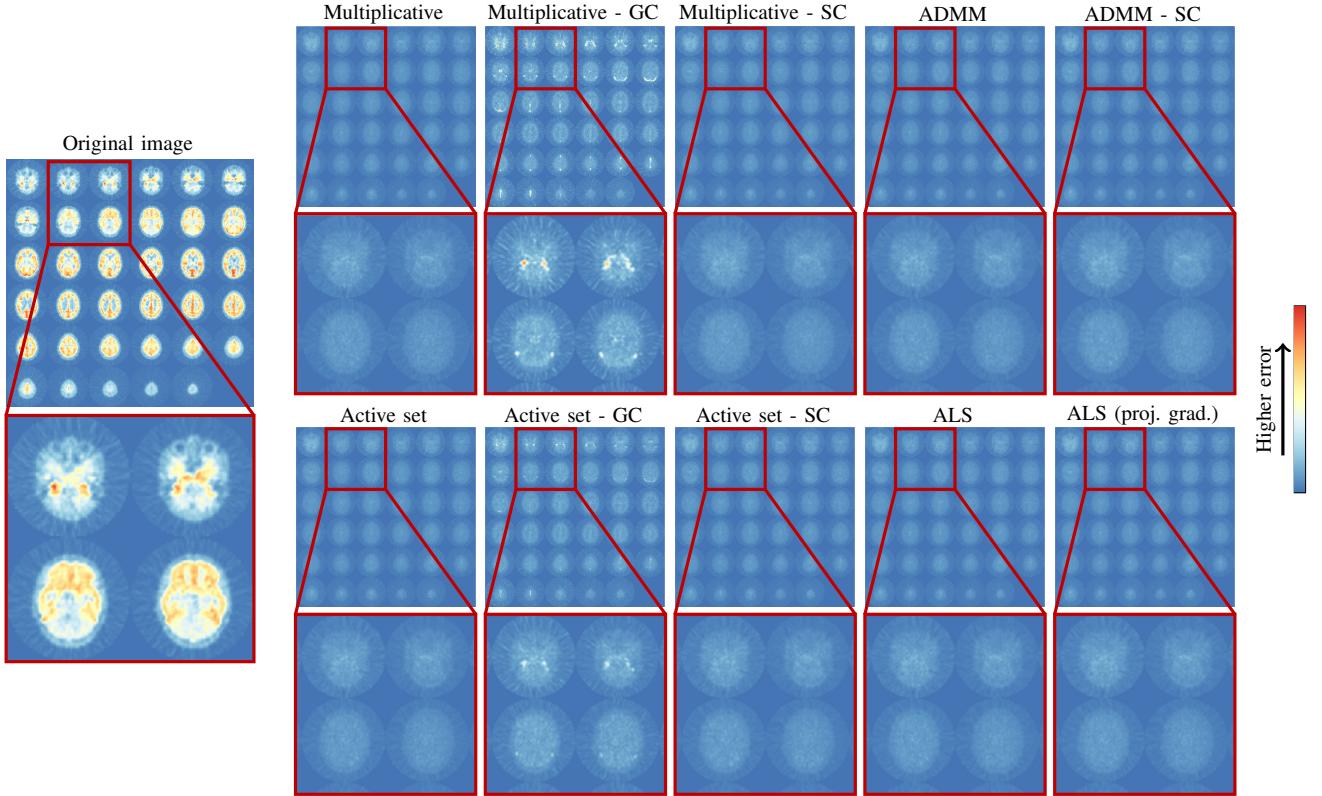


Fig. 8. **Reconstruction errors when compressing a positron emission tomography (PET) image** (<http://cogsys.imm.dtu.dk/toolbox/nmf/>). The image is composed of 40 temporal frames, where each frame is a  $128 \times 128 \times 35$  3D image (35 is the number of  $128 \times 128$  slices). The matrix size is then  $573440 \times 40$  and we perform NMF with  $r = 5$ . As we can observe in each slice (a few of them are highlighted with zoom-ins), the use of Gaussian compression (GC) increases the reconstruction errors, while structured random compression (SC) has no identifiable effect. See Table I for additional numerical results.

no matter how simple is the structure of the data. We also investigated how much faster is the proposed method with respect to QR-based approaches. We generated  $m \times n$  input matrices, where  $m$  is fixed and  $n$  varies; we then extract  $n/10$  columns. Remember that QR-based approaches solve an  $n \times n$  version of Problem (24c), while the proposed compressed approach solves an  $(r + r_{\text{ov}}) \times n$  version. This difference is reflected almost exactly in the speedup that we observe in Fig. 11(b): about an order of magnitude is gained with the proposed scheme.

In Fig. 12 we analyze the same dataset as in Fig. 9. Interestingly, a similar conclusion is reached using SNMF and NMF. The data is well explained by the same two factors (in this case, two extreme columns). Notice that the analyzed matrix is fat and the QR-based approach provides no speedup, i.e.,  $\mathbf{R} \in \mathbb{R}^{m \times n}$  in Problem (24c). On the other hand, the proposed approach produces a smaller problem independently of the input matrix's shape. Quantitatively, in this example, compressed SNMF is two orders of magnitude faster than the QR-based SNMF.

Our last example consists on an application for selecting representative frames from videos. We first examine a short clip (5 seconds long, 120 frames) of the open-source movie “Elephants Dream” at a resolution of 360p ( $640 \times 360$ ). In Table II we show a summary of the comparisons performed with this video. An example of the frames extracted by SPA with compression is shown in Fig. 13.

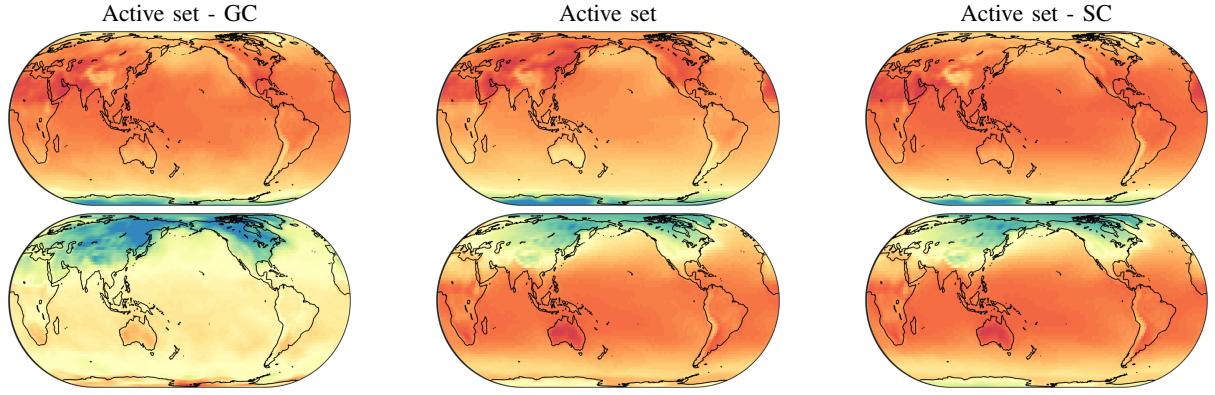
Our first observation is that the proposed compressed SNMF

is at least an order of magnitude faster than the QR-based variant. Second, since the matrix built from video is not truly low-rank, projecting the matrix into a low-rank subspace by means of compression seems to yield better results than when using the QR decomposition. Intuitively, compression eliminates some variability in the data in such a way that it can be better approximated by SNMF.

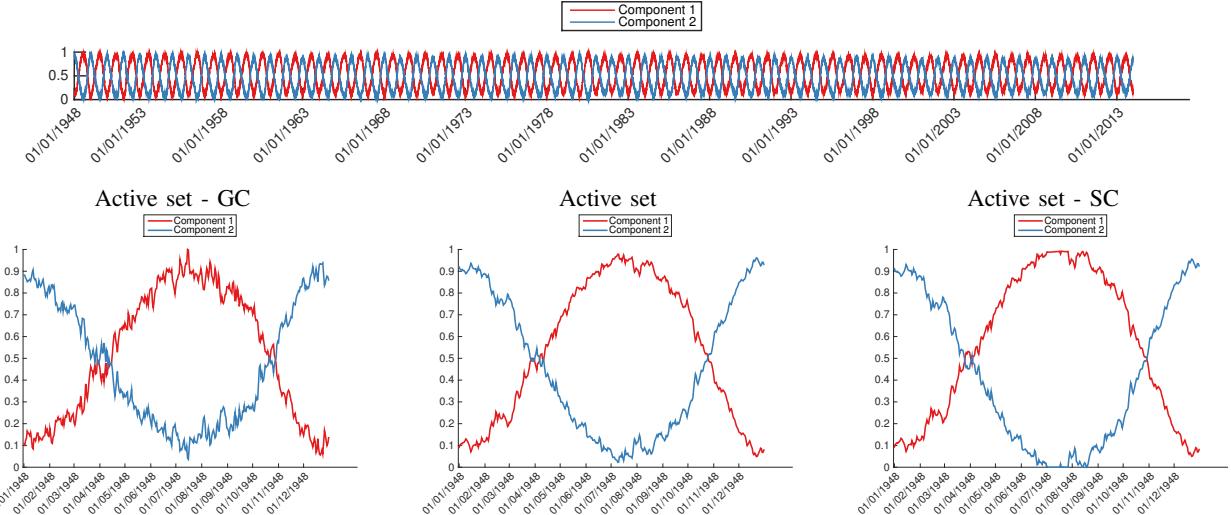
Although not strictly comparable, because it does not impose nonnegativity constraints, we included in our comparisons the method for extracting representative elements from [26]. As discussed in Section IV, this method's formulation does not scale gracefully with large input matrices. A fact that is easily reflected in the slow running time, even for a relatively small example.

*Scaling to Big Data:* We also run tests on the complete open-source movie “Elephants Dream.”<sup>2</sup> The movie is approximately 11 minutes long (15691 frames). We processed the video at two resolutions, 360p ( $640 \times 360$ ), and 1080p ( $1920 \times 1080$ ), resulting in  $691200 \times 15691$  and  $6220800 \times 15691$  matrices, respectively. The HDF5 files occupy 43.55 GB and 391.13 GB, respectively, not fitting in main memory. Using compressed SPA, we extract 130 representatives (extreme columns) from the video, one every 120 frames (5 seconds). At 360p we obtained a relative error of 0.2941 in 1891 seconds (about 32 minutes). At 1080p, we obtained a relative error of 0.2676 in 20776 seconds (about 5:46 hours) processing both

<sup>2</sup><http://www.elephantdream.org>



(a) **Left factors analysis.** Interestingly, the first factor (top row) corresponds to summer and winter in the north and south hemispheres, respectively, while the second factor (bottom row) corresponds to summer in the south and winter in the north. Visually, it is very clear that SC introduces much less artifacts than GC compared to the vanilla method (center).



(b) **Right factors analysis.** In the top row, we can easily observe that the two factors are periodic and inversely correlated, corroborating the winter/summer duality between both components. In the bottom row, we observe that the GC factors are much more noisy (about an order of magnitude larger), compared to the original and SC methods.

Fig. 9. **NMF on gridded climate data** (<http://www.esrl.noaa.gov/psd/repository/>). The data contains daily mean surface temperatures arranged in a  $144 \times 73$  grid since 1948 (23742 days in total), forming a  $10512 \times 23742$  matrix. We perform NMF using the active set method with  $r = 2$ . The computing times for the method in its vanilla version (center), with Gaussian compression (GC) and, with structured random compression (SC) were 50, 70, and 20 seconds, respectively; the respective relative reconstruction errors were 0.0459, 0.0537, and 0.0458, confirming the conclusions reached through visual inspection.

matrices on a laptop with 16GB of memory.

## VI. CONCLUSIONS

In this work we proposed to use structured random projections for NMF and SNMF. For NMF, we presented formulations for three popular techniques, namely, multiplicative updates [9], active set method for nonnegative least squares [10], and ADMM [11]. For SNMF, we presented a general technique that can be used with any algorithm. In all cases, we showed that the resulting compressed techniques are faster than their uncompressed variants and, at the same time, do not introduce significant errors in the final result.

There are in the literature very efficient SNMF algorithms for tall-and-skinny matrices. Interestingly, the use of structured random projections allows to compute SNMF for arbitrarily large matrices, granting access to very efficient computations in the general setting.

As a byproduct, we also propose an algorithmic solution for computing structured random projections of extremely large

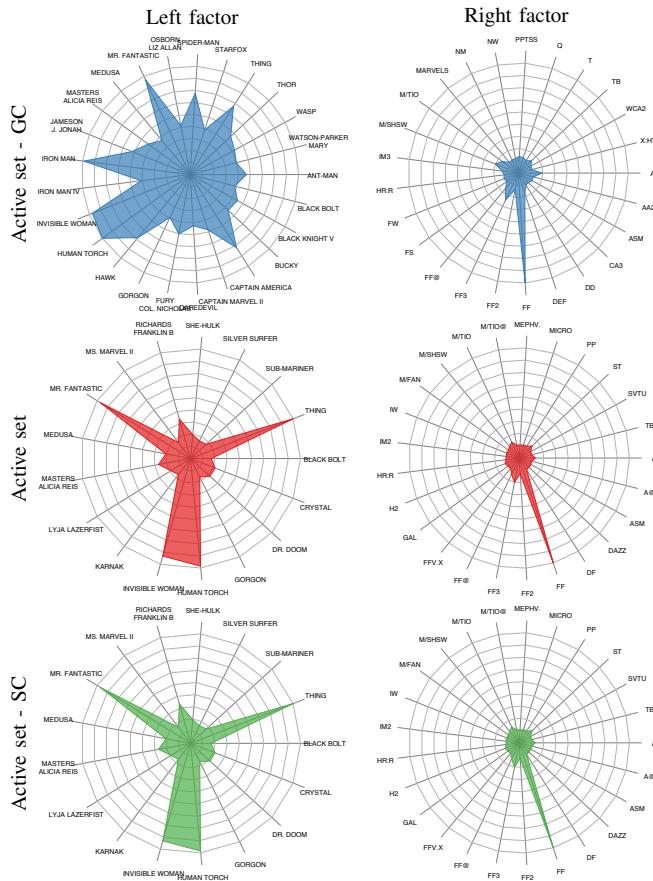
matrices (i.e., matrices so large that even after compression they do not fit in main memory). This is useful as a general tool for computing many different matrix decompositions, such as the singular value decomposition, for example.

We are currently investigating the problem of replacing the Frobenius norm with an  $\ell_p$  norm in our compressed variants of NMF and SNMF. In this setting, the fast Cauchy transform [15] is a suitable alternative to structured random projections. Compression consists of sampling and rescaling rows of  $\mathbf{A}$ , thus identifying the so-called *coreset* of the problem. This formulation is of particular interest for network analysis, where we need to deal with sparse structures.

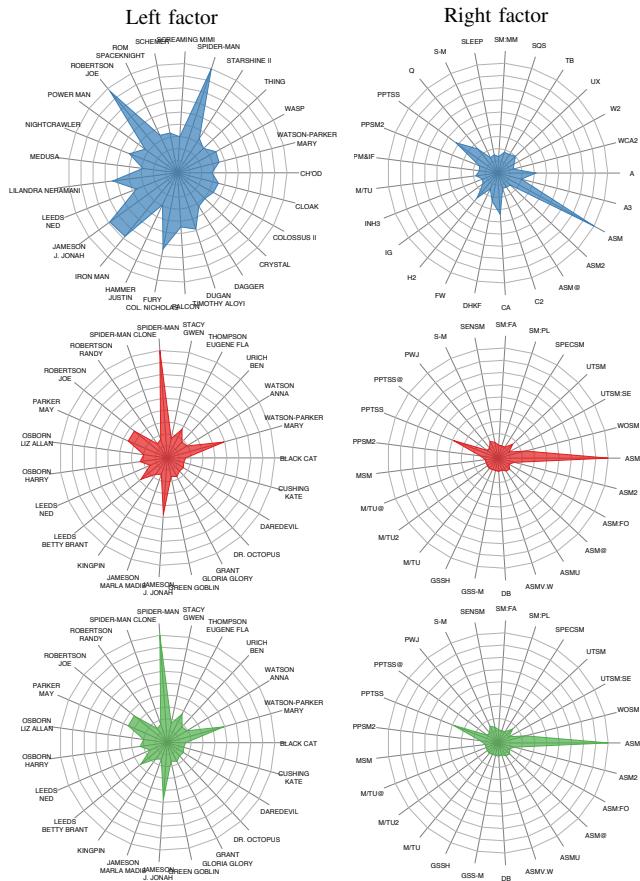
## ACKNOWLEDGMENTS

The authors would like to thank Mauricio Delbracio for many useful scientific discussions and Matthew Rocklin for his help and technical support with the dask and into libraries.

Fantastic Four



## Spider-Man



**Fig. 10. Biclustering the Marvel Universe collaboration network** (<http://www.chronologyproject.com/>). The network links Marvel characters and the Marvel comic books in which they appear, and exhibits most characteristics of real-life collaboration networks [28]. It can be represented as an  $m \times n$  matrix, where  $m = 6445$  and  $n = 12850$  are the number of characters and comics, respectively. We bicluster this matrix using NMF with  $r = 10$ , aiming at obtaining 10 very representative groups of characters appearing jointly in different comic books. The  $i$ th bicluster ( $i = 1 \dots 10$ ) is formed by the  $i$ th column of  $\mathbf{X}$  and the  $i$ th row of  $\mathbf{Y}$  (small entries were set to zero, as explained in Section V-A). The radar plots represent the coefficients of these vectors. We show two biclusters that we identify with characters from the Fantastic Four (first two columns of the figure) and the Spider-Man (last two columns of the figure) comics. Active set NMF correctly identifies that Mr Fantastic, The Thing, the Invisible Woman, and the Human Torch are the four most recurring characters in the “Fantastic Four” (FF) series. Similarly, active set NMF correctly identifies that Spider-Man/Peter Parker, Mary Jane Watson-Parker (Peter Parker’s wife), and Jonah Jameson (Peter Parker’s boss) are the most recurring characters in the “Amazing Spider-Man” (ASM) and “Peter Parker, The Spectacular Spider-Man” (PPTSS) series. It is clear that the biclusters recovered using structured random compression (SC) are very close to the biclusters found with no compression; contrarily, Gaussian compression (GC) significantly affects the biclustering result. All 10 biclusters can be found at <http://www.marianoteppe.com.ar/research/cnmf>.

## APPENDIX

### A. QR decompositions for tall-and-skinny matrices

The direct TSQR algorithm uses a simple but highly efficient approach for computing that QR decomposition of a tall and skinny matrix. Let  $\mathbf{A}$  be the  $m \times n$  matrix to decompose ( $m \gg n$ ). The direct TSQR algorithm starts by splitting  $\mathbf{A}$  into a stack of  $b$  blocks

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\mathcal{K}_1} \\ \vdots \\ \mathbf{A}_{\mathcal{K}_b} \end{bmatrix}, \quad (28)$$

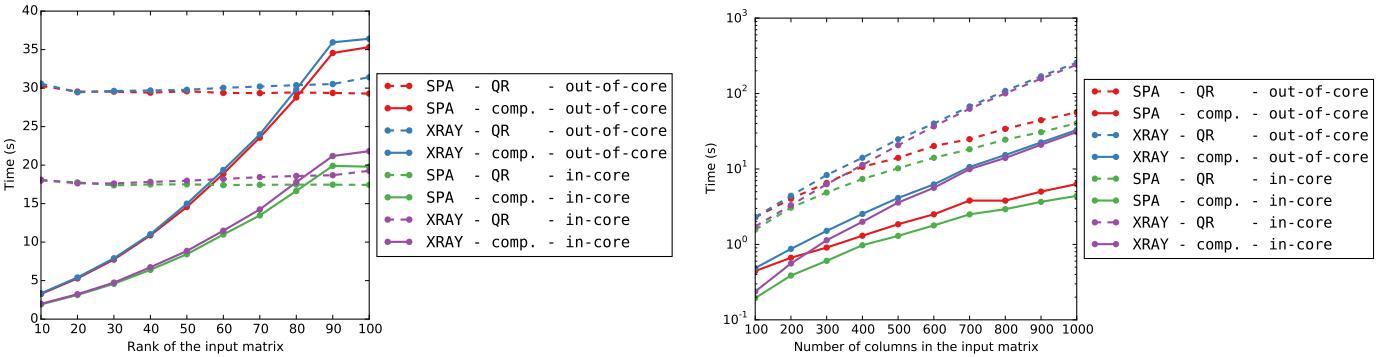
where  $\mathcal{K}_i$  denotes the set of rows selected in the  $i$ th block. Each block  $\mathbf{A}_{\mathcal{K}_i}$  is factorized into its components  $\mathbf{Q}_{\mathcal{K}_i}^{(1)}, \mathbf{R}_{\mathcal{K}_i}$ , using any standard QR decomposition algorithm. This can be

written in matrix form as

$$\begin{bmatrix} \mathbf{A}_{\mathcal{K}_1:} \\ \vdots \\ \mathbf{A}_{\mathcal{K}_b:} \end{bmatrix}_{m \times n} = \underbrace{\begin{bmatrix} \mathbf{Q}_{\mathcal{K}_1:}^{(1)} & & \\ & \ddots & \\ & & \mathbf{Q}_{\mathcal{K}_b:}^{(1)} \end{bmatrix}_{m \times bn}}_{\text{Block-diagonal matrix}} \begin{bmatrix} \mathbf{R}_{\mathcal{K}_1:} \\ \vdots \\ \mathbf{R}_{\mathcal{K}_b:} \end{bmatrix}_{bn \times n}. \quad (29)$$

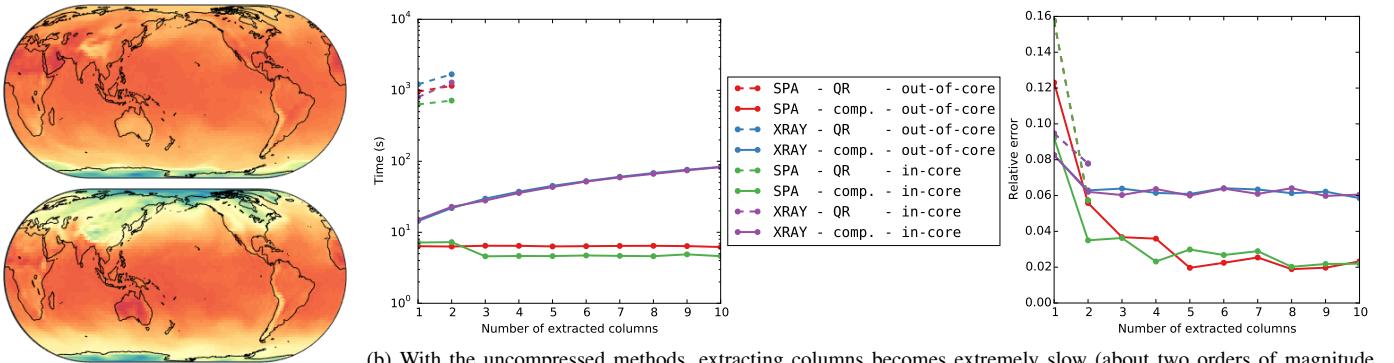
The second step is to gather the matrix composed by vertically stacking the factors  $\mathbf{R}_{\mathcal{K}_i}$ , and computing an additional QR decomposition, i.e.,

$$\begin{bmatrix} \mathbf{R}_{\mathcal{K}_1:} \\ \vdots \\ \mathbf{R}_{\mathcal{K}_b:} \end{bmatrix}_{bn \times n} = \underbrace{\begin{bmatrix} \mathbf{Q}_{\mathcal{K}_1:}^{(2)} \\ \vdots \\ \mathbf{Q}_{\mathcal{K}_b:}^{(2)} \end{bmatrix}_{b^m \times n}}_{\text{b}^m \times n} \underbrace{\mathbf{R}_{n \times n}}_{n \times n}. \quad (30)$$



(a) We extract  $r$  columns, where  $r$  is the rank of the  $10^6 \times 100$  input matrix. As expected, the computing time of the QR-based methods does not change with the number of extracted columns. On the other hand, compressed methods are faster when the rank of the input matrix is low compared to its size. In this case, out-of-core methods appear slower than in-core ones (slightly above 2 $\times$ ). We use an oversampling factor  $r_{ov} = 10$  for compression, which explains the flattening of the compressed curves towards their end.

Fig. 11. **Performance of different SNMF algorithms on synthetic matrices.** We generate the input matrix  $\mathbf{A} = \mathbf{X}_{\text{GT}} \mathbf{Y}_{\text{GT}}$ , where  $\mathbf{X}_{\text{GT}} \in \mathbb{R}^{m \times r}$  and  $\mathbf{Y}_{\text{GT}} \in \mathbb{R}^{r \times n}$  have normally distributed entries ( $r$  and  $n$  take different values in subfigures (a) and (b)). All algorithms select the same set of columns, thus producing equal errors.



(a) Columns extracted with SPA-comp. When  $r = 2$ , the extreme columns look similar to the ones found with traditional NMF, see Fig. 9.

(b) With the uncompressed methods, extracting columns becomes extremely slow (about two orders of magnitude slower) than with the compressed methods. Since compressed SNMF is faster with 10 columns than QR-based SNMF with two columns, we simply stopped the computation of the latter after 2 columns. Notice that these QR-based methods are explicitly designed to be faster for tall-and-skinny matrices, but end-up being extremely slow for fat matrices. The proposed compressed SNMF is also very fast for fat matrices.

Fig. 12. **SNMF on gridded climate data.** Same dataset as in Fig. 9. The data form a fat  $10512 \times 23742$  matrix. We study the performance of SNMF in terms of computing speed and relative error as the number  $r$  of columns changes. As with NMF, the data is well explained with only two factors by observing the decay in the reconstruction error. SPA with compression seems not to increase its computing time as the number of extracted columns increases; this is due to forcing the compression algorithm to produce at least 20 rows, the subsequent column extraction in SPA is extremely efficient. Notice that SPA with compression is about four times faster than NMF using the active set method with compression, see Fig. 9.

This is the only centralized step in TSQR. We then multiply the intermediate Q factors to get the matrix

$$\mathbf{Q} = \underbrace{\begin{bmatrix} \mathbf{Q}_{\mathcal{K}_1,:}^{(1)} & & \\ & \ddots & \\ & & \mathbf{Q}_{\mathcal{K}_b,:}^{(1)} \end{bmatrix}}_{m \times bn} \underbrace{\begin{bmatrix} \mathbf{Q}_{\mathcal{K}_1,:}^{(2)} \\ \vdots \\ \mathbf{Q}_{\mathcal{K}_b,:}^{(2)} \end{bmatrix}}_{bn \times n} = \begin{bmatrix} \mathbf{Q}_{\mathcal{K}_1,:}^{(1)} \mathbf{Q}_{\mathcal{K}_1,:}^{(2)} \\ \vdots \\ \mathbf{Q}_{\mathcal{K}_b,:}^{(1)} \mathbf{Q}_{\mathcal{K}_b,:}^{(2)} \end{bmatrix}. \quad (31)$$

Finally note that  $\mathbf{A} = \mathbf{Q}\mathbf{R}$ , where  $\mathbf{Q}$  is an orthonormal matrix (obtained from the multiplication of two orthonormal matrices) and  $\mathbf{R}$  is by algorithmic design, upper triangular. Thus, these matrices form a QR decomposition of  $\mathbf{A}$ .

1) *TSQR for structured random compression:* When using TSQR for compressing a matrix  $\mathbf{A}$ , Fig. 1, the input matrix to decompose is

$$\mathbf{B} = (\mathbf{A}\mathbf{A}^T)^w \mathbf{A}\Omega, \quad (32)$$

where  $w \in \mathbb{N}$ . Let us assume, for simplicity, that  $w = 0$ . The input of TSQR is not the matrix  $\mathbf{B}$  as a whole, but blocks extracted from it. We can thus avoid storing the entire matrix  $\mathbf{B}$  in main memory, and compute its blocks as needed, i.e.,

$$\mathbf{B}_{\mathcal{K}_i,:} = \mathbf{A}_{\mathcal{K}_i,:} \Omega. \quad (33)$$

A similar (but more complex) indexing holds for  $w > 0$ .

### B. An ADMM algorithm for solving Problem (19)

We consider the augmented Lagrangian of Problem (19),

$$\begin{aligned} \mathcal{L}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \mathbf{U}, \mathbf{V}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}) = & \left\| \tilde{\mathbf{A}} - \tilde{\mathbf{X}}\tilde{\mathbf{Y}} \right\|_F^2 + \\ & + \boldsymbol{\Lambda} \bullet (\mathbf{L}\tilde{\mathbf{X}} - \mathbf{U}) + \frac{\lambda}{2} \left\| \mathbf{L}\tilde{\mathbf{X}} - \mathbf{U} \right\|_F^2 + \\ & + \boldsymbol{\Phi} \bullet (\tilde{\mathbf{Y}}\mathbf{R} - \mathbf{V}) + \frac{\phi}{2} \left\| \tilde{\mathbf{Y}}\mathbf{R} - \mathbf{V} \right\|_F^2, \end{aligned} \quad (34)$$

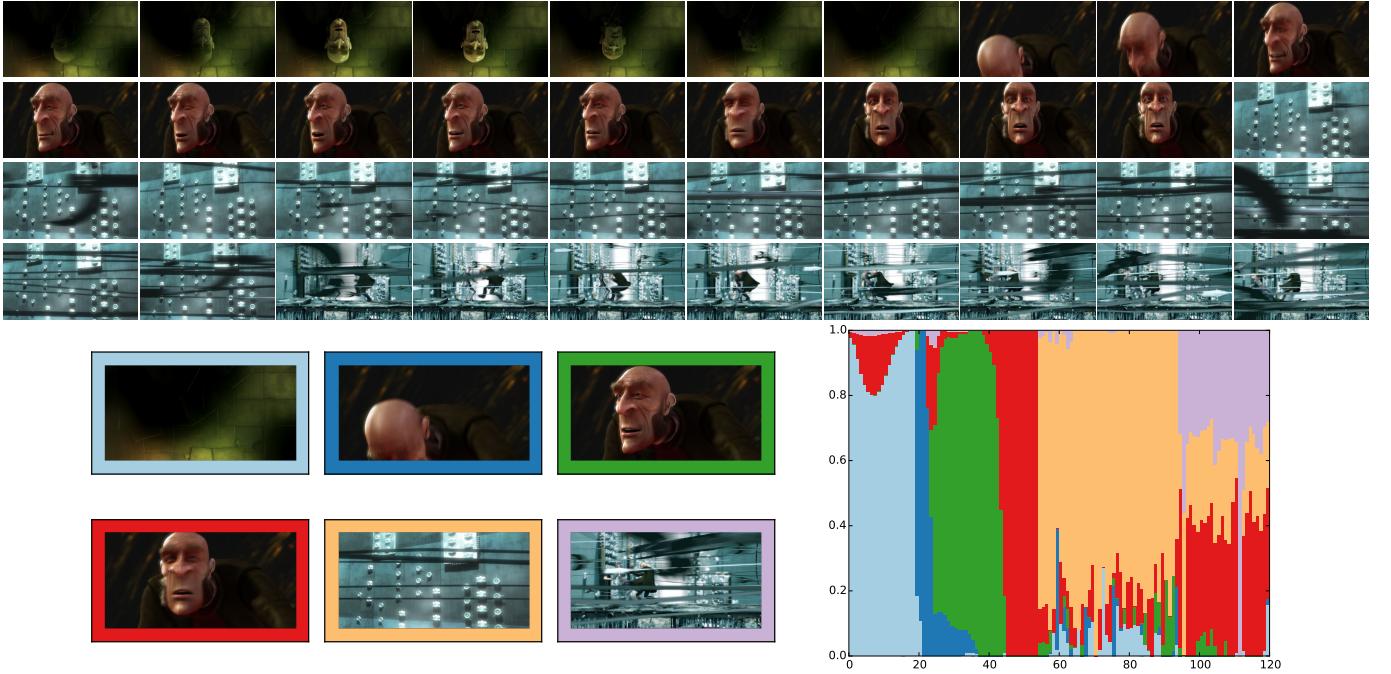


Fig. 13. **Extracting representative frames from a video** (<http://www.elephantsdream.org>). The video resolution is  $640 \times 360$  pixels and contains 120 frames (5 seconds). On the top block, we display 40 uniformly sampled frames. We build a  $691,200 \times 120$  matrix by vectorizing one frame per column (each frame has 3 color channels), and then use SNMF with compression to extract six representative frames (bottom left). On the bottom right we show the (normalized) columns of the matrix  $\mathbf{H}$  in Step (25), i.e., the reconstruction coefficients. It took 2.18 seconds to compute the result with relative errors of 0.2714 and of 0.4240 with respect to the compressed and the original matrices, respectively.

TABLE II. **Extracting representative frames from a video** For details about the experiment setup, see Fig. 13. We are considering a (relatively small)  $691,200 \times 120$  matrix to be able to compare the performance of in-core and out-of-core methods and with ESV [26], which is not fit for large scale matrices. The proposed compression scheme for SNMF (SPA-COMP) greatly improves speed with no detriment for the reconstruction error. Notice that since the matrix is not actually low-rank (it is a video), enforcing the projection onto a subspace helps in finding a better solution (SPA-COMP versus SPA-QR).

Methods	Comp. model	$r$	Time (s)	Rel. error
SPA-COMP	in-core	6	2.28	0.4240
SPA-COMP	out-of-core	6	4.75	0.4293
SPA-QR	in-core	6	18.76	0.5446
SPA-COMP	in-core	9	2.31	0.3626
SPA-COMP	out-of-core	9	4.59	0.3610
SPA-QR	in-core	9	19.08	0.4453
ESV [26] ( $\alpha = 2$ ) <sup>1</sup>	in-core	9	57.38	0.3751 <sup>2</sup>
SPA-COMP	in-core	15	2.65	0.3068
SPA-COMP	out-of-core	15	5.50	0.3047
SPA-QR	in-core	15	19.93	0.4011
ESV [26] ( $\alpha = 50$ ) <sup>1</sup>	in-core	15	68.05	0.1358 <sup>2</sup>

<sup>1</sup>  $\alpha$  is a regularization parameter that (indirectly) controls the number of representatives  $r$ .

<sup>2</sup> The errors are not directly comparable since this formulation does not impose nonnegativity.

where  $\Lambda \in \mathbb{R}^{m \times r}$ ,  $\Phi \in \mathbb{R}^{r \times n}$  are Lagrange multipliers,  $\lambda, \phi \in \mathbb{R}^+$  are penalty parameters, and  $\mathbf{B} \bullet \mathbf{C} = \sum_{i,j} (\mathbf{B})_{ij} (\mathbf{C})_{ij}$  for matrices  $\mathbf{B}, \mathbf{C}$  of the same size.

We use the Alternating Direction Method of Multipliers (ADMM) for solving Problem (19). The algorithm works in a coordinate descent fashion, successively minimizing  $\mathcal{L}$  with respect to  $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \mathbf{U}, \mathbf{V}$ , one at a time while fixing the others

at their most recent values, i.e.,

$$\tilde{\mathbf{X}}_{k+1} = \underset{\tilde{\mathbf{X}}}{\operatorname{argmin}} \mathcal{L}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_k, \mathbf{U}_k, \mathbf{V}_k, \Lambda_k, \Phi_k), \quad (35a)$$

$$\tilde{\mathbf{Y}}_{k+1} = \underset{\tilde{\mathbf{Y}}}{\operatorname{argmin}} \mathcal{L}(\tilde{\mathbf{X}}_{k+1}, \tilde{\mathbf{Y}}, \mathbf{U}_k, \mathbf{V}_k, \Lambda_k, \Phi_k), \quad (35b)$$

$$\mathbf{U}_{k+1} = \underset{\mathbf{U} \geq 0}{\operatorname{argmin}} \mathcal{L}(\tilde{\mathbf{X}}_{k+1}, \tilde{\mathbf{Y}}_{k+1}, \mathbf{U}, \mathbf{V}_k, \Lambda_k, \Phi_k), \quad (35c)$$

$$\mathbf{V}_{k+1} = \underset{\mathbf{V} \geq 0}{\operatorname{argmin}} \mathcal{L}(\tilde{\mathbf{X}}_{k+1}, \tilde{\mathbf{Y}}_{k+1}, \mathbf{U}_{k+1}, \mathbf{V}, \Lambda_k, \Phi_k), \quad (35d)$$

and then updating the multipliers  $\Lambda, \Phi$ . Each of these steps can be written in closed form and define our algorithm, see Fig. 14. In practice, we set  $\alpha, \beta, \gamma, \xi$  to 1.

We now provide a preliminary convergence property of the proposed ADMM algorithm. Our analysis follows closely the one in [11, Section 2.3].

To simplify notation, we consolidate all the variables as

$$Z = (\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \mathbf{U}, \mathbf{V}, \Lambda, \Phi).$$

A point  $Z$  is a Karush-Kuhn-Tucker (KKT) condition of Problem (19) if

$$(\tilde{\mathbf{X}}\tilde{\mathbf{Y}} - \tilde{\mathbf{A}})\tilde{\mathbf{Y}}^T + \Lambda = 0, \quad (36a)$$

$$\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}}\tilde{\mathbf{Y}} - \tilde{\mathbf{A}}) + \Phi = 0, \quad (36b)$$

$$\mathbf{L}\tilde{\mathbf{X}} - \mathbf{U} = 0, \quad (36c)$$

$$\tilde{\mathbf{Y}}\mathbf{R} - \mathbf{V} = 0, \quad (36d)$$

$$\Lambda \leq 0 \leq \mathbf{U}, \Lambda \circ \mathbf{U} = 0, \quad (36e)$$

$$\Phi \leq 0 \leq \mathbf{V}, \Phi \circ \mathbf{V} = 0, \quad (36f)$$

---

**input** : a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , a target rank  $r \in \mathbb{N}^+$ , an oversampling parameter  $r_{ov} \in \mathbb{N}^+$  ( $r + r_{ov} \leq \min\{m, n\}$ ), an exponent  $w \in \mathbb{N}$ .

**output**: nonnegative matrices  $\mathbf{U}_k \in \mathbb{R}^{m \times r}$ ,  $\mathbf{V}_k \in \mathbb{R}^{r \times n}$ .

- 1 Compute compression matrices  $\mathbf{L} \in \mathbb{R}^{m \times (r+r_{ov})}$ ,  $\mathbf{R} \in \mathbb{R}^{(r+r_{ov}) \times n}$ ;
- 2  $k \leftarrow 1$ ;
- 3 Initialize  $\mathbf{U}_k, \mathbf{V}_k$ ;
- 4  $\tilde{\mathbf{A}} \leftarrow \mathbf{L}^T \mathbf{A} \mathbf{R}^T$ ;  $\tilde{\mathbf{Y}} \leftarrow \mathbf{V}_k \mathbf{R}^T$ ;
- 5  $\Lambda_k \leftarrow \mathbf{0}$ ;  $\Phi_k \leftarrow \mathbf{0}$ ;
- 6  $\mathbf{I} \leftarrow$  the  $r \times r$  identity matrix
- 7 **repeat**
- 8    $\tilde{\mathbf{X}}_{k+1} \leftarrow (\tilde{\mathbf{A}} \tilde{\mathbf{Y}}_k^T + \lambda \mathbf{L}^T \mathbf{U}_k - \mathbf{L}^T \Lambda_k) (\tilde{\mathbf{Y}}_k \tilde{\mathbf{Y}}_k^T + \lambda \mathbf{I})^{-1}$ ;
- 9    $\tilde{\mathbf{Y}}_{k+1} \leftarrow (\tilde{\mathbf{X}}_{k+1}^T \tilde{\mathbf{X}}_{k+1} + \phi \mathbf{I})^{-1} (\tilde{\mathbf{X}}_{k+1}^T \tilde{\mathbf{A}} + \phi \mathbf{V}_k \mathbf{R}^T - \Phi_k \mathbf{R}^T)$ ;  
//  $(\mathcal{P}_+(\mathbf{B}))_{ij} = \max\{(\mathbf{B})_{ij}, 0\}$
- 10    $\mathbf{U}_{k+1} \leftarrow \mathcal{P}_+(\tilde{\mathbf{L}} \tilde{\mathbf{X}}_{k+1} + \lambda^{-1} \Lambda_k)$ ;
- 11    $\mathbf{V}_{k+1} \leftarrow \mathcal{P}_+(\tilde{\mathbf{Y}}_{k+1} \mathbf{R} + \phi^{-1} \Phi_k)$ ;
- 12    $\Lambda_{k+1} \leftarrow \Lambda_k + \xi \lambda (\mathbf{L} \tilde{\mathbf{X}}_{k+1} - \mathbf{U}_{k+1})$ ;
- 13    $\Phi_{k+1} \leftarrow \Phi_k + \xi \phi (\tilde{\mathbf{Y}}_{k+1} \mathbf{R} - \mathbf{V}_{k+1})$ ;
- 14    $k \leftarrow k + 1$ ;
- 15 **until** convergence;

---

Fig. 14. ADMM algorithm for NMF with structured random compression.

where  $\circ$  denotes the Hadamard (entrywise) matrix product.

**Proposition 1.** Let  $\{Z_k\}_{k=1}^\infty$  be a sequence generated by the algorithm in Fig. 14 that satisfies the condition

$$\lim_{k \rightarrow \infty} (Z_{k+1} - Z_k) = 0. \quad (37)$$

Then any accumulation point of  $\{Z_k\}_{k=1}^\infty$  is a KKT point of Problem (19).

*Proof:* From Assumption (37), we have

$$\tilde{\mathbf{X}}_{k+1} - \tilde{\mathbf{X}}_k \rightarrow 0, \quad (38a)$$

$$\tilde{\mathbf{Y}}_{k+1} - \tilde{\mathbf{Y}}_k \rightarrow 0, \quad (38b)$$

$$\Lambda_{k+1} - \Lambda_k \rightarrow 0, \quad (38c)$$

$$\Phi_{k+1} - \Phi_k \rightarrow 0, \quad (38d)$$

$$\mathbf{U}_{k+1} - \mathbf{U}_k \rightarrow 0, \quad (38e)$$

$$\mathbf{V}_{k+1} - \mathbf{V}_k \rightarrow 0. \quad (38f)$$

Plugging these subtractions in the variable updates in Fig. 14, we get

$$(\tilde{\mathbf{A}} - \tilde{\mathbf{X}}_k \tilde{\mathbf{Y}}_k) \tilde{\mathbf{Y}}_k^T - \mathbf{L} \Lambda_k \rightarrow 0, \quad (39a)$$

$$\tilde{\mathbf{X}}_{k+1}^T (\tilde{\mathbf{A}} - \tilde{\mathbf{X}}_{k+1} \tilde{\mathbf{Y}}_k) - \Phi_k \mathbf{R} \rightarrow 0, \quad (39b)$$

$$\mathbf{L} \tilde{\mathbf{X}}_{k+1} - \mathbf{U}_{k+1} \rightarrow 0, \quad (39c)$$

$$\tilde{\mathbf{Y}}_{k+1} \mathbf{R} - \mathbf{V}_{k+1} \rightarrow 0, \quad (39d)$$

$$\mathcal{P}_+(\mathbf{L} \tilde{\mathbf{X}}_{k+1} + \lambda^{-1} \Lambda_k) - \mathbf{U}_k \rightarrow 0, \quad (39e)$$

$$\mathcal{P}_+(\tilde{\mathbf{Y}}_{k+1} \mathbf{R} + \phi^{-1} \Phi_k) - \mathbf{V}_k \rightarrow 0. \quad (39f)$$

Notice that the terms  $\lambda (\mathbf{L}^T \mathbf{U}_k - \tilde{\mathbf{X}}_k)$  and  $\phi (\mathbf{V}_k \mathbf{R}^T - \tilde{\mathbf{Y}}_k)$  have been eliminated from equations (39a) and (39b) by invoking equations (39c) and (39d), respectively. Equations (36a–36d) are clearly satisfied by equations (39a–39d) at any limit

point

$$Z_\infty = (\tilde{\mathbf{X}}_\infty, \tilde{\mathbf{Y}}_\infty, \mathbf{U}_\infty, \mathbf{V}_\infty, \Lambda_\infty, \Phi_\infty).$$

We are then left to prove that equations (36e) and (36f) hold. Algorithm (35d) guarantees the non-negativity of  $\mathbf{U}_\infty, \mathbf{V}_\infty$ . Let us focus on Equation (36e) first. Equation (39e), when combined with Equation (39c), yields

$$\mathbf{U}_\infty = \mathcal{P}_+(\mathbf{U}_\infty + \lambda^{-1} \Lambda_\infty), \quad (40)$$

If  $(\mathbf{U}_\infty)_{ij} = 0$ , we get  $(\mathcal{P}_+(\lambda^{-1} \Lambda_\infty))_{ij} = 0$  and then  $(\Lambda_\infty)_{ij} \leq 0$ . If  $(\mathbf{U}_\infty)_{ij} > 0$ , we get  $(\mathbf{U}_\infty)_{ij} = \mathcal{P}_+(\mathbf{U}_\infty)_{ij}$  and  $(\Lambda_\infty)_{ij} = 0$ . From this, we obtain that Equation (36e) holds. An identical argument applies for equations (36f) and (39f).

With this, we have proven that any accumulation point of  $\{Z_k\}_{k=1}^\infty$  is a KKT point of Problem (17). From the equivalence of problems (1) and (17), any accumulation point of  $\{(\mathbf{X}_k, \mathbf{Y}_k)\}_{k=1}^\infty$  is a KKT point of Problem (1). ■

**Corollary 1.** Whenever  $\{Z_k\}_{k=1}^\infty$  converges, it converges to a KKT point of Problem (17).

Ideally, we would like to guarantee that Algorithm (35d) will always converge to a KKT point of Problem (19). The above simple result is an initial step in this direction, providing some assurance on the behavior of Algorithm (35d).

## REFERENCES

- [1] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” *Commun ACM*, vol. 51, no. 1, pp. 107–113, 2008.
- [2] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.
- [3] P. Melville and V. Sindhwani, “Recommender systems,” in *Encyclopedia of Machine Learning*. Springer, 2010, pp. 829–838.
- [4] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Comput*, vol. 21, no. 3, pp. 793–830, 2009.
- [5] S. A. Vavasis, “On the complexity of nonnegative matrix factorization,” *SIAM J Optim*, vol. 20, no. 3, pp. 1364–1377, 2010.
- [6] N. Gillis, “Sparse and unique nonnegative matrix factorization through data preprocessing,” *J Mach Learn Res*, vol. 13, no. 1, pp. 3349–3386, 2012.
- [7] S. Arora, R. Ge, R. Kannan, and A. Moitra, “Computing a nonnegative matrix factorization – provably,” in *STOC*, 2012.
- [8] N. Halko, P.-G. Martinsson, and J. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.
- [9] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2000.

- [10] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J Matrix Anal Appl*, vol. 30, pp. 713–730, 2008.
- [11] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Front Math China*, vol. 7, no. 2, pp. 365–384, 2012.
- [12] A. R. Benson, J. D. Lee, and D. F. Gleich, "Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrices," in *NIPS*, 2014.
- [13] J. Tropp, "Improved analysis of the subsampled randomized Hadamard transform," *Adv Adapt Data Anal*, vol. 3, no. 01n02, pp. 115–126, 2011.
- [14] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr Approx*, vol. 28, no. 3, pp. 253–263, 2008.
- [15] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff, "The fast Cauchy transform and faster robust linear regression," in *SODA*, 2013, pp. 466–477.
- [16] M. Chu, F. Diele, R. Plemmons, and S. Ragni, "Optimality, computation, and interpretation of nonnegative matrix factorizations," Tech. Rep., 2004.
- [17] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput*, vol. 19, no. 10, pp. 2756–2779, 2007.
- [18] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans Pattern Anal Mach Intell*, vol. 32, no. 1, pp. 45–55, 2010.
- [19] C. Liu, H.-C. Yang, J. Fan, L.-W. He, and Y.-M. Wang, "Distributed nonnegative matrix factorization for web-scale dyadic data analysis on MapReduce," in *WWW*, 2010.
- [20] R. Liao, Y. Zhang, J. Guan, and S. Zhou, "CloudNMF: A MapReduce implementation of nonnegative matrix factorization for large-scale biological datasets," *Genomics, Proteomics Bioinformatics*, vol. 12, no. 1, pp. 48–51, 2014.
- [21] M. Araújo, T. Saldanha, R. Galvão, T. Yoneyama, H. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multi-component analysis," *Chemometr Intell Lab Syst*, vol. 57, pp. 65–73, 2001.
- [22] V. Bittorf, B. Recht, R. Christopher, and J. Tropp, "Factoring nonnegative matrices with linear programs," in *NIPS*, 2012.
- [23] A. Kumar, V. Sindhwani, and P. Kambadur, "Fast conical hull algorithms for near-separable non-negative matrix factorization," in *ICML*, 2013.
- [24] N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans Pattern Anal Mach Intell*, vol. 36, no. 4, pp. 698–714, 2014.
- [25] E. Esser, M. Möller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Trans Image Process*, vol. 21, no. 7, pp. 3239–3252, 2012.
- [26] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," *CVPR*, 2012.
- [27] F. Wang and P. Li, "Efficient nonnegative matrix factorization with random projections," in *SDM*, 2010.
- [28] R. Alberich, J. Miro-Julia, and F. Rossello, "Marvel universe looks almost like a real social network," 2002, arXiv:cond-mat/0202174.