Nonnegative matrix factorization with side information for time series recovery and prediction

Jiali MEI

Abstract

Motivated by the reconstruction and prediction of electricity consumption, we extend Nonnegative Matrix Factorization (NMF) to take into account outside features. We consider Nonnegative Matrix Factorization in general linear measurement schemes, and propose a general framework which models non-linear relationship between features and the response variables. We extend previous theoretical results in NMF to obtain a sufficient condition on the identifability of matrix factorization. Based the classical Hierarchical Alternating Least Squares (HALS) algorithm, we propose a new algorithm (HALSX, or Hierarchical Alternating Least Squares with eXogeneous variables) which estimates the factorization model. The algorithm is validated on both simulated and real electricity consumption data, to show its performance in reconstruction and prediction.

1 Introduction

1.1 Motivation

Motivation for general linear operators

Motivation for using exogeneous variables

Motivation for NMF

1.2 General model definition

We are interested in reconstructing a nonnegative matrix $\mathbf{V}^* \in \mathbb{R}_+^{n_1 \times n_2}$, from N linear measurements,

$$\boldsymbol{\alpha} = \mathcal{A}(\mathbf{V}^*) \in \mathbb{R}^N, \tag{1}$$

where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^N$ is a linear operator. Formally, $\mathcal{A}$ can be represented by $\mathbf{A}_1$, ..., $\mathbf{A}_N$, N design matrices of dimension $n_1 \times n_2$, and each linear measurement can be represented by

$$\alpha_i = \mathrm{Tr}(\mathbf{V}^* \mathbf{A}_i^T) = \langle \mathbf{V}^*, \mathbf{A}_i \rangle. \tag{2}$$

The linear operator $\mathcal{A}$ is also called a mask.

Moreover, we suppose that the matrix of interest, $\mathbf{V}^*$, stems from a generative low-rank nonnegative model, in the following sense:

1. The nonnegative rank of $\mathbf{V}^*$ is k with $k \ll n_1, n_2$. Therefore, we can find two nonnegative factor matrices $\mathbf{F}_r \in \mathbb{R}_+^{n_1 \times k}$ and $\mathbf{F}_c \in \mathbb{R}_+^{n_2 \times k}$ so that

$$\mathbf{V}^* = \mathbf{F}_r \mathbf{F}_c^T.$$

2. There are some row features $\mathbf{X}_r \in \mathbb{R}^{n_1 \times d_1}$ and column features $\mathbf{X}_c \in \mathbb{R}^{n_2 \times d_2}$ connected to each row and column of $\mathbf{V}^*$. We note by $\mathbf{x}_r^i$ the i-th row of $\mathbf{X}_r$, and by $\mathbf{x}_c^i$ the i-th row of $\mathbf{X}_c$. There are also two link functions $f_r : \mathbb{R}^{d_1} \to \mathbb{R}^k$ and $f_c : \mathbb{R}^{d_1} \to \mathbb{R}^k$, so that

$$\mathbf{F}_r = f_r(\mathbf{X}_r),$$
$$\mathbf{F}_c = f_c(\mathbf{X}_c),$$

   where $f_r(\mathbf{X}_r) \in \mathbb{R}^{n_1 \times k}$ is the matrix obtained by stacking row vectors $f_r(\mathbf{x}_r^i)^T$, for $1 \leq i \leq n_1$ (*idem* for $f_c(\mathbf{X}_c) \in \mathbb{R}^{n_n \times k}$)

In this general setting, the features $\mathbf{X}_r$ and $\mathbf{X}_c$, the measurement operator $\mathcal{A}$, and the measurements $\boldsymbol{\alpha}$ are availble. The objective is to estimate the true matrix $\mathbf{V}^*$ as well as the factor matrices $\mathbf{F}_r$ and $\mathbf{F}_c$, by estimating the link functions f_r and f_c.

By specializing $\mathbf{X}_r$, $\mathbf{X}_c$, and $\mathcal{A}$, or restricting the search space of f_r and f_c, this general model includes a number of interesting applications, old and new.

The mask $\mathcal{A}$

- Identitiy: $N = n_1 n_2, \mathbf{A}_{i_1,i_2} = \mathbf{e}_{i_1} \mathbf{e}_{i_2}^T$, where $\mathbf{e}_i$ is the i-th canonical vector. This means the data $\boldsymbol{\alpha}$ directly includes every entry of $\mathbf{V}^*$.
- Matrix completion mask: $N < n_1 n_2$, the set of design matrix is a subset of identity mask.
- Random sensoring mask: the design matrices $\mathbf{A}_i$ dense matrices, sampled from a certain probability distribution. Typically, the probability distribution needs to verify some condtions, so that with a large probability, $\mathcal{A}$ verifies the Restricted Isometry Property (Recht et al. [2010]).
- Column or row mask: the design matrices are either a subset of the identity mask, but observations are entries of whole rows and columns of $\mathbf{V}^*$, or the design matrices are sparse matrices with non-zero entries only on a single row or a single column: the observations are therefore linear combinations of rows or columns of $\mathbf{V}^*$.
- Temporal aggregate mask: in this case, the matrix is composed of n_1 time series concerning n_2 individuals, and each measure is a temporal aggregate of the time serie of an individual. The design matrices are defined as $\mathbf{A}_i = \sum_{t=t_0(i)+1}^{t_0(i)+h(i)} \mathbf{e}_t \mathbf{e}_{s_i}^T$, where s_i is the individual concerned by the i-th measure, $t_0(i)+1$ the beginning the period covered by the measure, and $h(i)$ the number of periods covered by the measure.

The features $\mathbf{X}_r$ and $\mathbf{X}_c$

- Individual features: $\mathbf{X}_r = \mathbf{I}_{n_1}, \mathbf{X}_c = \mathbf{I}_{n_2}$. Basically, no additional information is supplied through these features. The row individuals and column individuals are each different.

Table 1 – Classification of matrix factorization with side information: by the mask, the link function, and the features included as side information.

Mask	Link function	Linear			Other regresion methods
	Features	Identity	Kernel features	General numeric features	General numeric features
Identity					
Matrix completion					
Matrix sensing					
Row and column mask					
Temporal aggregates					

- General numeric features: $\mathbf{X}_r \in \mathbb{R}^{n_1 \times d_1}$ and $\mathbf{X}_c \in \mathbb{R}^{n_2 \times d_2}$. This includes all numeric features.
- Features transformed by a kernel: certain information about the row and column individuals may not be in the form of a numeric vector. For example, if the row individuals are vertices of a graph, their connection to each other is interesting information for the problem, but it is difficult to encode as real vectors. In this case, features can be generated through a transformation, or by defining a kernel function.

The link functions f_r and f_c

- Linear link function: $\mathbf{F}_r = f_r(\mathbf{X}_r) = \mathbf{X}_r\mathbf{B}_r$, and $\mathbf{F}_c = f_c(\mathbf{X}_c) = \mathbf{X}_c\mathbf{B}_c$. In this case, we need to estimate $\mathbf{B}_r$ and $\mathbf{B}_c$ to fit the model. With identity matrices as row and column features, this case is reduced to the traditional matrix factorization model with

$$\mathbf{F}_r = \mathbf{B}_r, \quad \mathbf{F}_c = \mathbf{B}_c, \quad \mathbf{V}^* = \mathbf{F}_r\mathbf{F}_c^T = \mathbf{B}_r\mathbf{B}_c^T.$$

  When the features are transformed features from a kernel function, even a linear link function permits non-linear relationship between the features and the factor matrices.
- General regression models: when the relationship between the features and the variable of interest is not linear, any off-the-shelf regression methods can be plugged in to search for a non-linear link function.

In this work, we further restrict our scope on Nonnegative Matrix Factorization (NMF). That is, the matrix to be estimated, as well as the low-rank factorization, is nonnegative. Namely, $\mathbf{V}^* \geq \mathbf{0}$, $(f_r(\mathbf{X}_r), f_c(\mathbf{X}_c)) \geq \mathbf{0}$.

The optimization problem As in classical matrix factorization, we will minimize the quadratic error of the matrix approximation. In Section 3, we will propose

a general algorithm for the following optimization problem:

$$\begin{aligned} \min_{\mathbf{V}, f_r \in F_r^k, f_c \in F_c^k} \quad & \|\mathbf{V} - f_r(\mathbf{X}_r) f_c(\mathbf{X}_c)^T\|_F \\ \text{s.t.} \quad & f_r(\mathbf{X}_r) \geq \mathbf{0}, f_c(\mathbf{X}_c) \geq \mathbf{0}, \\ & \mathcal{A}(\mathbf{V}) = \mathbf{b} = \mathcal{A}(\mathbf{V}^*), \end{aligned} \tag{3}$$

where $F_r \subseteq (\mathbb{R})^{\mathbb{R}^{d_1}}$ and $F_c \subseteq (\mathbb{R})^{\mathbb{R}^{d_2}}$ are the functional spaces in which the row and column link functions are to be searched.

We will also compare (3) to another candidate estimator, defined as follows:

$$\begin{aligned} \min_{f_r \in F_r^k, f_c \in F_c^k} \quad & \|\mathbf{b} - \mathcal{A}(f_r(\mathbf{X}_r) f_c(\mathbf{X}_c)^T)\|_F \\ \text{s.t.} \quad & f_r(\mathbf{X}_r) \geq \mathbf{0}, f_c(\mathbf{X}_c) \geq \mathbf{0}. \end{aligned} \tag{4}$$

1.3 Prior works

Table 1 shows a taxonomy of matrix factorization models with side information, by the mask, the link function and the features used as side information. A number of problems in this taxonomy has been adressed in established literature.

Matrix factorization An abundant literature studies the general matrix factorization (without nonnegativity) problem under various measurement operators, when no additional information is provided (both row and column features are identity matrices). Notably, General operator Rohde and Tsybakov [2011] Identity mask (SVD) Completion mask Candès and Recht [2009] RIP measurement operator Recht et al. [2010], Bhojanapalli et al. [2016] Other operators are less considered Zuk and Wagner [2015]

Reduced-rank regression On the other hand, under the identity mask, the problem is known as reduced rank regression: the objective is to exploit the low-rank property in multivariate regression. This approach is was first developped very early (see Velu and Reinsel [2013] for a review). Recent developments on rank selection (Bunea et al. [2012]), adaptive estimation procedure (Chen et al. [2013], using non-parametric link function (Foygel et al. [2012]), etc. draw the parallel between reduced-rank regression and the matrix completion problem. However, measurement operators other than identity or matrix completion are rarely considered in this case.

Inductive matrix completion Building on theoretical boundaries on matrix completion, the authors of Jain and Dhillon [2013], Xu et al. [2013], Chiang et al. [2015] showed that by providing side information (the matrix $\mathbf{X}$), the number of measurements needed for exact matrix completion can be reduced. Moreover, the number of measurements necessary for successful matrix completion can be quantified by measuring the quality of the side information Chiang et al. [2015].

Parametric, nonparametric, kernels One of the first methods for including side information in collaborative filtering systems was proposed by Abernethy et al. [2009]. The authors generalized collaborative filtering into a operator estimation

problem. This method allow more general feature spaces than a numerical matrix, by applying a kernel function to side information. Si et al. [2016] proposed choosing the kernel function based on the goal of the application. Kekatos et al. [2014] applied the kernel-based collaborative filtering framework to electricity price forecasting. Their kernel choice is determined by multi-kernel leanrning methods.

2 Identification of nonnegative matrix factorization with side information

General matrix factorization is not a well-identified problem: for one pair of factors $(\mathbf{F}_r, \mathbf{F}_c)$, with $\mathbf{V}^* = \mathbf{F}_r\mathbf{F}_c^T$, any inversible matrix $\mathbf{R}$ produces another pair of factors, $(\mathbf{F}_r\mathbf{R}, \mathbf{F}_c(\mathbf{R}^{-1})^T)$, with $(\mathbf{F}_r\mathbf{R})(\mathbf{F}_c(\mathbf{R}^{-1})^T)^T = \mathbf{V}^*$. In order to adress this identification problem, one has to introduce extra constraints on the factors.

When the nonnegativity constraint is imposed on $\mathbf{F}_r$ and $\mathbf{F}_c$, however, it has been shown that sometimes the only inversible matrices that verify $\mathbf{F}_r\mathbf{R} \geq 0$ and $\mathbf{R}^{-1}\mathbf{F}_c \geq 0$ are the composition of a permutation matrix and a diagonal matrix with strictly positive diagonal elements. A nonnegative matrix factorization is said to be identified if the factors are unique up to permutation and scaling. In this section, we review some known necessary and sufficient conditions for NMF identification in the literature, and develop a sufficient condition for NMF identification in the context of linear numerical features.

In order to simplify our theoretical analysis, we focus on the identity mask in this section (the data matrix $\mathbf{V}^*$ is completely observed). Also, we derive the sufficient condition for row features. That is, we will derive conditions on $\mathbf{V}^*\mathbb{R}_0^{n_1\times n_2}$ and $\mathbf{X}_r \in \mathbb{R}^{n_1\times d_1}$, so that the nonnegative matrix factorization $\mathbf{V}^* = \mathbf{X}_r\mathbf{B}_r\mathbf{F}_c^T$, with $\mathbf{X}_r\mathbf{B}_r \geq 0, \mathbf{F}_c \geq 0$, is unique. A generalization to column features can be easily obtained.

2.1 Identification of NMF

Donoho and Stodden [2003] and Laurberg et al. [2008] proposed two necessary and sufficient conditions for the factorization to be unique. Both conditions use the following geometric interpretation of NMF introduced by Donoho and Stodden [2003].

Suppose that $\mathbf{V}^*$ is a n_1-by-n_2 nonnegative matrix of rank k, which has a rank-k nonnegative factorization. That is $\mathbf{V}^* = \mathbf{F}_r\mathbf{F}_c^T$, with $(\mathbf{F}_r, \mathbf{F}_c) \in \mathbb{R}_+^{n_1\times k} \times \mathbb{R}_+^{n_2\times k}$. Then, the columns of $\mathbf{V}^*$ are conical combinations of the columns in $\mathbf{F}_r$. Formally, $\text{cone}(\mathbf{F}_r)$, the conical hull of the columns of $\mathbf{F}_r$, is a polyhedral cone contained in the first orthant of $\mathbb{R}^{n_1}$. As $\mathbf{V}^*$ is of rank k, the rank of $\mathbf{F}_r$ is also k. This consequently implies that the extreme rays (also called *generators*) of $\text{cone}(\mathbf{F}_r)$ are exactly the columns of $\mathbf{F}_r$, which are linearly independant. $\text{cone}(\mathbf{F}_r)$ is therefore

- a *simplicial* cone of k generators,
- contained in $\mathbb{R}_+^{n_1}$,
- containing all columns of $\mathbf{V}^*$.

Inversely, if we take any cone $\mathcal{F} \subseteq \mathbb{R}^{n_1}$ verifying these three conditions, and define a matrix $\mathbf{F}$ whose columns are the k generators of $\mathcal{F}$, there will be a nonnegative matrix $\mathbf{G}$, so that $\mathbf{V}^* = \mathbf{F}\mathbf{G}$. The uniqueness of NMF is therefore equivalent to the

uniqueness of simplicial cones of k generators contained in the first orthant of $\mathbb{R}^{n_1}$ and containing all columns $\mathbf{V}^*$.

Laurberg et al. [2008] gave an equivalent geometric interpretaion in $\mathbb{R}^k$ which is formalized by the following theorem:

Theorem 1 (Laurberg et al. [2008]). *A k-dimensional NMF $\mathbf{V}^* = \mathbf{F}_r\mathbf{F}_c$ of a rank-k nonnegative matrix $\mathbf{V}^*$ is unique if and only if the nonnegative orthant $\mathbb{R}^k_+$ is the only simplicial cone $\mathcal{A}$ with k extreme rays satisfying*

$$\text{cone}(\mathbf{F}_r^T) \subseteq \mathcal{A} \subseteq \text{cone}(\mathbf{F}_c^\star).$$

Despite the apparent simlicity of the theorem, the necessary and sufficient conditions are very difficult to check. Based on the theorem above, several sufficient conditions have been proposed. The most widely used condition is called the separability condition. Before introducing this condition (in its least restrictive version presented by Laurberg et al. [2008]), we need the following two definitions.

Definition 1 (Separability). *A nonnegative matrix $\mathbf{M} \in \mathbb{R}_+^{m\times n}$ is said to be* separable *if there is a m-by-m permutation matrix Π which verifies*

$$\mathbf{M} = \Pi \begin{pmatrix} \mathbf{D}_n \\ \mathbf{M}_0 \end{pmatrix},$$

where $\mathbf{D}_n$ is a n-by-n diagonal matrix with only strictly positive coefficients on the diagonal and zeros everywhere else, and $\mathbf{M}_0$ is a collection of the other $m-n$ rows of $\mathbf{M}$.

Definition 2 (Strongly Boundary Closeness). *A nonnegative matrix $\mathbf{M} \in \mathbb{R}_+^{m\times n}$ is said to be* strongly boundary close *if the following conditions are satisfied.*

1. $\mathbf{M}$ *is* boundary close*: for all $i, j \in \{1, ..., n\}, i \neq j$, there is a row $\mathbf{m}$ in $\mathbf{M}$ which satisfies $m_i = 0, m_j > 0$;*
2. *There is a permutation of $\{1, ..., n\}$ such that for all $i \in \{1, ..., n-1\}$, there are $n-i$ rows $\mathbf{m}^1, ..., \mathbf{m}^{n-i}$ in $\mathbf{M}$ which satisfy*
   (a) *$m_i^j = 0, \sum_{s=i+1}^n m_i^j > 0$ for all $j \in \{1, ..., n-i\}$;*
   (b) *the square matrix $(m_s^j)_{1\leq j\leq n-i, i+1\leq s\leq n}$ is of full rank $(n-i)$.*

Strongly boundary closeness demands, *modulo* a permutation in $\{1, ..., n\}$, that for each $1 \leq i \leq n-1$, there are $n-i$ rows $\mathbf{m}^1, ..., \mathbf{m}^{n-i}$ of $\mathbf{M}$ that have the following form,

$$\begin{pmatrix} \mathbf{m}^1 \\ \vdots \\ \mathbf{m}^{n-i} \end{pmatrix}^T = \left.\begin{pmatrix} \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ m_{i+1}^1 & \cdots & m_{i+1}^{n-i} \\ \vdots & \ddots & \vdots \\ m_n^1 & \cdots & m_n^{n-i} \end{pmatrix}\right\} \begin{array}{c} (i-1) \text{ first rows} \\ i\text{-th row is all zero} \\ \\ (n-i)\text{-by-}(n-i) \text{ full rank square matrix} \end{array} \tag{5}$$

These row vectors, $\mathbf{m}^1, ..., \mathbf{m}^{n-i}$, all have 0 on the i-th element, and its lower square matrix of is of full rank. There are therefore enough linearly independant points on each $n-1$-dimensional facet $\mathbb{R}^n_+$, which shows that cone$(\mathbf{M}^T)$ is somewhat maximal in $\mathbb{R}^n_+$.

Laurberg et al. [2008] proved the following:

Theorem 2 (Laurberg et al. [2008]). *If $\mathbf{F}_r$ is strongly boundary close, then the only simplicial cone with k generators in $\mathbb{R}_+^k$ containing cone$(\mathbf{F}_r^T)$ is $\mathbb{R}_+^k$. Moreover, if $\mathbf{F}_c$ is separable, then $\mathbf{V}^* = \mathbf{F}_r\mathbf{F}_c^T$ is the unique NMF of $\mathbf{V}^*$ up to permutation and scaling.*

2.2 Identification with side information

The NMF with linear row features, $\mathbf{V}^* = \mathbf{X}_r\mathbf{B}_r\mathbf{F}_c^T$, is said to be *unique*, if for all matrix pairs $(\tilde{\mathbf{B}}_r, \tilde{\mathbf{F}}_c) \in \mathbb{R}^{d_1\times k} \times \mathbb{R}^{n_2\times k}$ that verifies

$$\mathbf{X}_r\tilde{\mathbf{B}}_r \geq 0, \quad \tilde{\mathbf{F}}_c \geq 0, \quad \mathbf{V}^* = \mathbf{X}_r\tilde{\mathbf{B}}_r\tilde{\mathbf{F}}_c,$$

we have $\tilde{\mathbf{B}} = \mathbf{B}_r$, $\tilde{\mathbf{F}}_c = \mathbf{F}_c$ up to permutation of columns and scaling.

For a given full-rank matrix $\mathbf{X} \in \mathbb{R}^{n_1\times d_1}$, consider the following two sets of matrices:

$$E = \{\mathbf{M} \in \mathbb{R}_+^{n_1\times k} | \text{The columns of } \mathbf{M} \text{ are strongly boundary close}\};$$
$$F(\mathbf{X}) = \{\mathbf{M} \in \mathbb{R}_+^{n_1\times k} | \text{rank}(\mathbf{M}) = k, \mathbf{M}_i \in \text{span}(\mathbf{X}), \forall 1 \leq i \leq k\}.$$

Theorem 3. *If $E \bigcap F(\mathbf{X}_r) \neq \varnothing$, and $\mathbf{B}_r \in (\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T(E \bigcap F(\mathbf{X}_r))$, and $\mathbf{F}_c$ is separable, then the factorization $\mathbf{V}^* = \mathbf{X}_r\mathbf{B}_r\mathbf{F}_c^T$ is a unique.*

This theorem is proved by noticing that for $\mathbf{B}_r \in (\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T(E \bigcap F(\mathbf{X}_r))$, the nonnegative matrix $\mathbf{F}_r = \mathbf{X}_r\mathbf{B}_r$ is strongly boundary close. The factorization $(\mathbf{F}_r, \mathbf{F}_c)$ is therefore unique. The model identification follows immediately, since $\mathbf{X}_r$ is of full rank.

Example of $\mathbf{X}_r$ that verifies $E \bigcap F(\mathbf{X}_r) \neq \varnothing$ For this theorem to have practical consequences, one needs to find appropriate row features so that $E \bigcap F(\mathbf{X}_r) \neq \varnothing$.

Here we provide a family of matrices $\mathbf{X}_r$ so that $E \bigcap F(\mathbf{X}_r) \neq \varnothing$.

With a fixed $k \geq 2$, suppose that $\mathbf{X}_r$ has $k(k-1)/2$ columns, and at least $k(k-1)/2$ rows, with the first $k(k-1)/2+1$ rows defined as the following:

- the first row and column have 0 on the first entry and positive entries elsewhere;
- for $2 \leq i \leq k$, $\mathbf{X}_r$ has strictly positive entries on the first $((i-1)(i-2)/2+1)$ columns, from Row $(i-1)(i-2)/2+3$ to Row $(i-1)(i-2)/2+1+i$, and zero entries everywhere else. These $(k-1)$ rows are linearly independant.

Then we have $E \bigcap F(\mathbf{X}_r) \neq \varnothing$, because the following $k(k-1)/2$-by-k matrix $\mathbf{B}_r$ is in this set:

- for $1 \leq i \leq k$, $\mathbf{B}_K^*$ has i consecutive strictly positive entries on the i-th column, between Row $i(i-1)/2+1$ and Row $i(i-1)/2+i$.

The following matrices instantiate the case of $k = 4$:

$$\mathbf{B}_r = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{X}_r = \begin{pmatrix} 0 & 5 & 14 & 7 & 9 & 15 & 13 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 5 & 0 & 0 & 0 & 0 & 0 \\ 12 & 4 & 0 & 0 & 0 & 0 & 0 \\ 10 & 7 & 10 & 7 & 0 & 0 & 0 \\ 13 & 10 & 12 & 9 & 0 & 0 & 0 \\ 12 & 10 & 16 & 8 & 0 & 0 & 0 \\ : & : & : & : & : & : & : \end{pmatrix}, \mathbf{F}_r = \begin{pmatrix} 0 & 5 & 21 & 37 \\ 10 & 0 & 0 & 0 \\ 4 & 5 & 0 & 0 \\ 12 & 4 & 0 & 0 \\ 10 & 7 & 17 & 0 \\ 13 & 10 & 21 & 0 \\ 12 & 10 & 24 & 0 \\ : & : & : & : \end{pmatrix}.$$

If $E \bigcap F(\mathbf{X}) \neq \varnothing$, for any invertible matrix $\mathbf{R} \in \mathbb{R}^{K \times K}$, $E \bigcap F(\mathbf{XR}) \neq \varnothing$.

3 HALSX algorithm

In this section, we propose an algorithm to estimate the general nonnegative matrix factorization problem with side information, by solving (3). This general algorithm, called Hierarchical Alternating Least Squares with eXogeneous variables (HALSX), will be an extension to a popular NMF algorithm: Hierarchical Alternating Least Squares (HALS) (see Cichocki et al. [2007], Kim et al. [2014]).

Before discussing HALSX, we will first present a result on the local convergence of Gauss-Seidel algorithms. This result guarantees that any legitimate limiting points generated by HALSX are critical points of (3).

While discussing specific methods to estimate link functions, we will suppose that only row features are available, as a generalization to include column features will be fairly straightforward.

3.1 Relaxation of convexity assumption for the convergence of Gauss-Seidel algorithm

To show the local convergence of HALSX algorithm, we first extend a result concerning block nonlinear Gauss-Seidel algorithm by Grippo and Sciandrone [2000] (Proposition 4).

Consider the minimization problem,

$$\begin{aligned} \min \quad & g(x) \\ \text{s.t.} \quad & x \in X = X_1 \times X_2 \times ... \times X_m \subseteq \mathbb{R}^n, \end{aligned} \tag{6}$$

where g is a continuously differentiable real-valued function, and the feasible set X is the Cartesian product of closed, nonempty and convex subsets $X_i \subset \mathbb{R}^{n_i}$, for $1 \leq i \leq m$, with $\sum_i n_i = n$. The m-block Gauss-Seidel algorithm is:

Define formally the notion of componentwise quasiconvexity.

Algorithm 1 Gauss-Seidel algorithm

Initialize $x^0 \in X, t = 0$
while Stopping criterion is not satisfied **do**
  for $i = 1, 2, ..., m$ **do**
    Calculate $x_i^{t+1} = \arg\min_{y_i \in X_i} g(x_1^{t+1}, ..., y_i, ..., x_m^t)$
  end for
  Set $x_{t+1} = (x_1^{t+1}, ..., x_m^{t+1})$
  $t = t + 1$
end while

Definition 3. *Let $i \in \{1, 2, ..., m\}$. The function g is quasiconvex with respect to the i-th component on X if for every $x \in X$ and $y_i \in X_i$, we have*

$$g(x_1, x_2, ...tx_i + (1-t)y_i, ..., x_m) \leq \max\{g(x), g(x_1, x_2, ..., y_i, ...x_m)\}$$

for all $t \in [0, 1]$. g is said to be strictly quasiconvex with respect to the i-th component, if with the additional assumption that $y_i \neq x_i$, we have

$$g(x_1, x_2, ...tx_i + (1-t)y_i, ..., x_m) < \max\{g(x), g(x_1, x_2, ..., y_i, ...x_m)\}$$

for all $t \in]0, 1[$.

Grippo and Sciandrone [2000] showed that if g is strictly quasiconvex with respect to the first $m-2$ blocks of components on X, then a limiting point produced by a Gauss-Seidel algorithm is a local minimum.

This result is not directly applicable for the HALS algorithm. Typically, if $\mathbf{f}_{c,i}$, the $i-th$ column of $\mathbf{F}_c$, is identically zero, the loss function is completely flat respect to $\mathbf{f}_{c,i}$, the i-th column of $\mathbf{F}_r$. Therefore the loss function is not strictly quasiconvex. In order to avoid this scenario, Kim et al. [2014] suggests thresholding at a small positive number ϵ instead of at 0, when updating each column of the factor matrices.

In fact the convexity assumption of Grippo and Sciandrone [2000] can be slightly relaxed to directly apply to HALS, as demonstrated by the following proposition.

Theorem 4. *Suppose that the function g is quasiconvex with respect to x_i on X, for $i = 1, ..., m-2$. Suppose that some limit points $\bar{x}$ of $\{x^t\}$ verify that g is strictly quasiconvex with respect to x_i on $\{\bar{x}_1\} \times \{\bar{x}_2\} \times ... \times X_i \times, ... \times \{\bar{x}_m\}$, for $i = 1, ..., m-2$. Then every such limiting point is a local minimum of Problem (1).*

Compared to the result of Grippo and Sciandrone [2000], this shows that the strictness of convexity with respect to one block does not have to hold universally for feasible regions of other blocks. It only needs to hold at the limiting point.

This theorem can be established following the proof of Proposition 5 of Grippo and Sciandrone [2000], using the following lemma.

Lemma 1. *Suppose that the function g is quasiconvex with respect to x_i on X, for some $i \in \{1, ..., m\}$. Suppose that some limit points $\bar{y}$ of $\{y^t\}$ verify that g is strictly quasiconvex with respect to x_i on $\{\bar{y}_1\} \times \{\bar{y}_2\} \times ... \times X_i \times, ... \times \{\bar{y}_m\}$. Let $\{v^t\}$ be a sequence of vectors defined as follows:*

$$v_j^t = \begin{cases} y_j^t & \text{if } j \neq i, \\ \arg\min_{z_i \in X_i} g(y_1^t, ..., z_i, ..., y_m^t) & \text{if } j = i. \end{cases}$$

Then, if $\lim_{t\to+\infty} g(y^t) - g(v^t) = 0$, *we have* $\lim_{t\to+\infty} ||v_i^t - y_i^t|| = 0$. *That is* $\lim_{t\to+\infty} ||v^t - y^t|| = 0$.

Proof. (The proof of the lemma is based on Bertsekas and Tsitsiklis [1989].)

Suppose on the contrary that $||v_i^t - y_i^t||$ does not converge to 0. Define $\tau_k = ||v_i^t - y_i^t||$. Restricting to a subsequence, we can obtain that $\tau_k \geq \tau_0 > 0$. Define $s^t = \frac{v^t - y^t}{\tau_k}$. Notice that $\{s^t\}$ is of unit norm, and $v^t = y^t + \tau_k s^t$. Since $\{s^t\}$ is on the unit sphere, it has a converging subsequence. By restricting to a subsequence again, we could suppose that $\{s^t\}$ converges to $\bar{s}$.

For all $\epsilon \in [0, 1]$, we have $0 \leq \epsilon\tau_0 \leq \tau_k$, which implies $y^t + \epsilon\tau_0 s^t \in X$ is on the segment $[y^t, v^t]$. This segment has strictly positive dimension in the subspace corresponding to X_i.

By the definition of $\{v^t\}$, $g(v^t) \leq g(y_1^t, ..., z_i, ..., y_m^t)$, for all t, and for all $z_i \in X_i$. In particular,

$$g(v^t) \leq g(y^t + \epsilon\tau_0 s^t).$$

By quasiconvexity of g on X,

$$g(y^t + \epsilon\tau_0 s^t) \leq \max\{g(y^t), g(v^t)\} = g(y^t).$$

Taking the limit when t converges to $+\infty$ on both equalities, we obtain

$$g(\bar{y}) = \lim_{t\to+\infty} g(v^t) \leq \lim_{t\to+\infty} g(y^t + \epsilon\tau_0 s^t) = g(\bar{y} + \epsilon\tau_0\bar{s}) \leq \lim_{t\to+\infty} g(y^t) = g(\bar{y}).$$

In other words, $g(\bar{y} + \epsilon\tau_0\bar{s}) = g(\bar{y})$, $\forall\epsilon \in [0, 1]$, which contradicts the strict quasiconvexity of g on $\{\bar{y}_1\} \times \{\bar{y}_2\} \times ... \times X_i \times, ... \times \{\bar{y}_m\}$. □

3.2 HALSX algorithm

The general algorithm to solve (3) is formalized by Algorithm 2. When the measurement operator is identity, the feature matrices are identity matrices, and when only linear functions are allowed as link functions, Algorithm 2 is identical to HALS.

From Theorem 4, one deduces that every full-rank limiting point produced by the popular HALS algorithm is a critical point.

Corollary 1. *Every full-rank factorization produced by HALS is a critical point of the NMF problem.*

Proof. The HALS algorithm is a Gauss-Seidel algorithm applied to a continuously differentiable loss function $g(\mathbf{V}, \mathbf{F}_r, \mathbf{F}_c) \equiv \frac{1}{2}\|\mathbf{V} - \mathbf{F}_r\mathbf{F}_c\|_F^2$. The loss function g is convex with respect to $\mathbf{V}, \mathbf{F}_r$ and $\mathbf{F}_c$ separately. To apply Theorem 4 on a limiting point $(\bar{\mathbf{V}}, \bar{\mathbf{F}}_r, \bar{\mathbf{F}}_c)$ generated by HALS, one only needs to check that g is strictly quasiconvex respect to $2k-1$ blocks of the $2k+1$ blocks involved (namely, $\bar{\mathbf{V}}, \bar{\mathbf{f}}_{r,1}, ..., \bar{\mathbf{f}}_{r,k}$, and $\bar{\mathbf{f}}_{c,1}, ..., \bar{\mathbf{f}}_{c,k}$). Take $i \in \{1, ..., k\}$. If both $\bar{\mathbf{F}}_r$ and $\bar{\mathbf{F}}_c$ are of full-rank, then $\bar{\mathbf{f}}_{c,i}$ is not identically zero. Therefore, the loss function on the subproblem for $\mathbf{f}_{r,i}$

$$\begin{aligned} l(\mathbf{f}_{r,i}) =& \frac{1}{2}\|\bar{\mathbf{V}} - \sum_{j\neq i} \bar{\mathbf{f}}_{r,j}\bar{\mathbf{f}}_{c,j}^T - \mathbf{f}_{r,i}\bar{\mathbf{f}}_{c,i}^T\|_F^2 \\ =& \frac{1}{2}\|\bar{\mathbf{f}}_{c,i}\|^2\|\mathbf{f}_{r,i}\|^2 - 2\langle\bar{\mathbf{V}} - \sum_{j\neq i} \bar{\mathbf{f}}_{r,j}\bar{\mathbf{f}}_{c,j}^T, \quad \mathbf{f}_{r,i}\bar{\mathbf{f}}_{c,i}^T\rangle + \frac{1}{2}\|\bar{\mathbf{V}} - \sum_{j\neq i} \bar{\mathbf{f}}_{r,j}\bar{\mathbf{f}}_{c,j}^T\|_F^2 \end{aligned}$$

Algorithm 2 Hierarchical Alternating Least Squares with eXogeneous variables for NMF (HALSX)

Require: Measurement operator $\mathcal{A}$, measurements $\mathbf{b}$, rank $1 \leq k \leq \min\{n_1, n_2\}$, features $\mathbf{X}_r$ and $\mathbf{X}_c$, functional spaces F_r and F_c in which to search the link functions.

Initialize $\mathbf{F}_r^0, \mathbf{F}_c^0 \geq 0, t = 0$
while Stopping criterion is not satisfied **do**
  $\mathbf{V}^t = \arg\min_{\mathbf{V}|\mathcal{A}(\mathbf{V})=\mathbf{b}, \mathbf{V}\geq 0} \|\mathbf{V} - \mathbf{F}_r^t(\mathbf{F}_c^t)^T\|_F^2$
  $\mathbf{R}^t = \mathbf{V}^t - \mathbf{F}_r^t(\mathbf{F}_c^t)^T$
  for $i = 1, 2, ..., k$ **do**
    $\mathbf{R}^t = \mathbf{R}^t + \mathbf{f}_{r,i}^t(\mathbf{f}_{c,i}^t)^T$
    Calculate $f_{r,i}^{t+1} = \arg\min_{f\in F_r} \|\mathbf{R}^t - f(\mathbf{X}_r)(\mathbf{f}_{c,i}^t)^T\|_F^2$
    $\mathbf{f}_{r,i}^{t+1} = \max(0, f_{r,i}^{t+1}(\mathbf{X}_r))$
    $\mathbf{R}^t = \mathbf{R}^t - \mathbf{f}_{r,i}^{t+1}(\mathbf{f}_{c,i}^t)^T$
  end for
  for $i = 1, 2, ..., k$ **do**
    $\mathbf{R}^t = \mathbf{R}^t + \mathbf{f}_{r,i}^{t+1}(\mathbf{f}_{c,i}^t)^T$
    Calculate $f_{c,i}^{t+1} = \arg\min_{f\in F_c} \|\mathbf{R}^t - \mathbf{f}_{r,i}^{t+1} f(\mathbf{X}_c)^T\|_F^2$
    $\mathbf{f}_{c,i}^{t+1} = \max(0, f_{c,i}^{t+1}(\mathbf{X}_c))$
    $\mathbf{R}^t = \mathbf{R}^t - \mathbf{f}_{r,i}^{t+1}(\mathbf{f}_{c,i}^{t+1})^T$
  end for
  $t = t + 1$
end while
return $\mathbf{V}^t = \arg\min_{\mathbf{V}|\mathcal{A}(\mathbf{V})=\mathbf{b}, \mathbf{V}\geq 0} \|\mathbf{V} - \mathbf{F}_r^t(\mathbf{F}_c^t)^T\|_F^2$,
$\mathbf{F}_r^t \in \mathbb{R}_+^{n_1\times k}, f_{r,1}^t, ..., f_{r,k}^t \in F_r$,
$\mathbf{F}_c^t \in \mathbb{R}_+^{n_2\times k}, f_{c,1}^t, ..., f_{c,k}^t \in F_c$.

is strictly convex function with respect to $\mathbf{f}_{r,i}$, since $\|\bar{\mathbf{f}}_{c,i}\|^2 > 0$. Apply the same argument to all columns of $\bar{\mathbf{F}}_r$ and all lines of $\bar{\mathbf{F}}_c$ provides the $2k$ necessary blocks. □

What about HALSX algorithm

Corollary 2. *Every full-rank factorization produced by HALSX (Algorithm 2) is a critical point of Problem (3).*

Proof. We will need to show that each column iteration produces a strict minimum if the opposed column is not identically zero. This is probably the case in the same way as HALS. □

3.3 Designs and HALSX

At each iteration of Algorithm 2, we need to projects the working matrix $\mathbf{F}_r^t(\mathbf{F}_c^t)^T$ into the convex polytope defined by the measurements and nonnegativity:

$$\mathbf{V}^t = \arg \min_{\mathbf{V}|\mathcal{A}(\mathbf{V})=\mathbf{b}, \mathbf{V}\geq 0} \|\mathbf{V} - \mathbf{F}_r^t(\mathbf{F}_c^t)^T\|_F^2. \tag{7}$$

In general, the polytope projection can be obtained by alternating projection. Namely, we can alternate between:

- $\mathbf{V} = \mathbf{V} + \mathcal{A}^\dagger(\mathbf{b} - \mathcal{A}(\mathbf{V}))$;
- $v_{i,j} = \max(0, v_{i,j})$,

where $\mathcal{A}^\dagger$ is the right pseudo-inverse of $\mathcal{A}$, viewed as an N-by-n_1n_2 matrix.

For some measurement operators, there are specific efficient ways to solve (7).

- Matrix completion mask: $v_{i,j} = \begin{cases} b_l, & \text{if } \exists 1 \leq l \leq N, \mathbf{A}_l = \mathbf{e}_i\mathbf{e}_j^T; \\ \max(0, v_{i,j}), & \text{if not.} \end{cases}$
- Temporal aggregate mask: simplex projection (see Mei et al. [2016] for details).

3.4 Linear HALSX

In this section, we consider HALSX with numeric row features with linear row link functions. That is, given $\mathbf{X}_r$ and $\mathbf{b} = \mathcal{A}(\mathbf{V}^*)$, we need to estimate $\mathbf{B}_r$ and $\mathbf{F}_c$ so that $\mathbf{V}^* = \mathbf{X}_r\mathbf{B}_r\mathbf{F}_c^T$.

Following Algorithm 2, we need to update the columns of $\mathbf{B}_r$ at each iteration. At the t-th step, for $1 \leq i \leq k$, we solve the subproblem

$$\arg\min_{\mathbf{b}_{r,i}} \|\mathbf{R}_i^t - \mathbf{X}_r\mathbf{b}_{r,i}(\mathbf{f}_{c,i}^t)^T\|_F^2,$$

where $\mathbf{R}_i^t = \mathbf{V}^t - \sum_{j=1,j\neq i}^k \mathbf{X}_r\mathbf{b}_{r,l}\mathbf{f}_{c,l}^T$. This minimization problem has a closed-form solution:

$$\mathbf{b}_{r,i}^{t+1} = \frac{1}{\|\mathbf{f}_{c,i}^t\|_2^2}(\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T\mathbf{R}_i^t\mathbf{f}_{c,i}^t.$$

In order to accelerate the numerical algorithm, a QR decomputation of $\mathbf{X}_r = \mathbf{QR}$ is done before the iterations, where $\mathbf{Q}$ is an orthogonal matrix, and $\mathbf{R}$ is a square upper triangular matrix. When $\mathbf{X}_r$ is of full rank, $\mathbf{X}_r^T\mathbf{X}_r$ is invertible. We compute one time $(\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r = \mathbf{R}^{-1}\mathbf{Q}^T$ before the iterations, and use the result at each iteration.

3.5 HALSX with smoothing splines

The computation considered above can estimate an NMF with linear features fairly efficiently. However, in real applications, linear link functions are too restrictive. In the following sections, we will consider estimate general link functions.

In this section, we estimate link functions that are Generalized Additive Models (GAM, Wood [2006]). A Generalized Additive Model is a generalization to Generalized Linear Model (GLM) which includes non-linear additive components of the form:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\theta} + h_1(x_{i,1}) + h_2(x_{i,2}) + h_3(x_{i,3}, x_{i,4}) + ...$$

where $\mu_i = \mathbf{E}(Y_i)$ is the expected value of the exponential family response variable Y_i, $\mathbf{x}_i$ is the vector of features ($\mathbf{X}$ will be the matrix grouping the features of all individuals), $\boldsymbol{\theta}$ is the vector of parametric model componnets, g is a known, monotonic, twice-differentiable function, h_1, h_2, h_3, ..., are the non-linear functions to be estimated.

We use penalized regression spline to fit the GAMs. For $j = 1, 2, 3, ...$, define a spline basis $\mathbf{a}^j = (a_1^j, a_2^j, ...)$ in which h_j, the j-th component of the GAM, is to be estimated. Pratically, we search for h_j is the L-dimensional vector space

$$H(\mathbf{a}^j, L_j) = \{\sum_{l=1}^{L_j} \beta_l^j a_l^j | \boldsymbol{\beta}^j = (\beta_1^j, ...\beta_{L_j}^j) \in \mathbb{R}^{L_j}\}.$$

Noting by $\mathbf{X}^j = \{a_l^j(\mathbf{x}_i)\}_{i,l}$ the design matrix, for $h_j = \sum_{l=1}^{L_j} \beta_l^j a_l^j \in H(\mathbf{a}^j, L_j)$, an element of the functional space, we have

$$h_j(\mathbf{X}) = \mathbf{X}^j \boldsymbol{\beta}^j.$$

The whole model of g, can then be represented linearly:

$$\begin{aligned} g(\boldsymbol{\mu}) &= \mathbf{X}\theta + (\mathbf{X}^1, \mathbf{X}^2, ...) \begin{pmatrix} \boldsymbol{\beta}^1 \\ \boldsymbol{\beta}^2 \\ : \end{pmatrix} \\ &= \mathbf{X}\boldsymbol{\theta} + \sum_{j=1} \mathbf{X}^j \boldsymbol{\beta}^j. \end{aligned}$$

The dimension of $H(\mathbf{a}^j, L_j)$, L_j, controls the the smoothness of the functions to be estiamted. As little information is available on the degree of smoothness of the functions, we use a rather large L_j, and add a penalty on the wiggliness, $\int (h_j'')^2 dx$, as in Wood [2006]. The least squares estimator of this model is therefore

$$\arg \min_{\boldsymbol{\theta}, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, ...} \|g(\boldsymbol{\mu}) - \mathbf{X}\boldsymbol{\theta} - \sum_{j=1} \mathbf{X}^j \boldsymbol{\beta}^j\|^2 + \sum_{j=1} \lambda^j (\boldsymbol{\beta}^j)^T \mathbf{S}^j \boldsymbol{\beta}^j,$$

where λ_j is the penalization parameter of the j-th non-linear component, and $\mathbf{S}^j$ is a positive definite matrix depending on $\mathbf{X}$ and $\mathbf{a}^j$. The penalization parameter, λ^j, is chosen by a generalized cross validation criterion.

HALSX-GAM At each iteration of the algorithm, for $i = 1, ..., k$, we re-estimate the link function $f_{r,i}$ of the i-th column of $\mathbf{F}_r$ as a GAM.

The subproblem for i is the following

$$\arg \min_{\boldsymbol{\theta}, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, ...} \|\mathbf{R}_i^t - (\mathbf{X}_r \boldsymbol{\theta} + \sum_{j=1} \mathbf{X}^j \boldsymbol{\beta}^j)(\mathbf{f}_{c,i}^t)^T\|_F^2 + \sum_{j=1} \lambda^j (\boldsymbol{\beta}^j)^T \mathbf{S}^j \boldsymbol{\beta}^j. \tag{8}$$

With fixed penalization parameters λ_1, λ_2, ..., the optimization above can be solved by

$$\begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\beta^1} \\ \boldsymbol{\beta^2} \\ : \end{pmatrix}^{t+1} = \frac{1}{\|f_{c,i}^t\|^2} \begin{pmatrix} \mathbf{X}_r^T \mathbf{X}_r & \mathbf{X}_r^T \mathbf{X}^1 & \mathbf{X}_r^T \mathbf{X}^2 & \cdots \\ (\mathbf{X}^1)^T \mathbf{X}_r & (\mathbf{X}^1)^T \mathbf{X}^1 + \frac{\lambda^1}{\|f_{c,i}^t\|^2} \mathbf{S}^j & (\mathbf{X}^1)^T \mathbf{X}^2 & \cdots \\ : & : & \ddots & \cdots \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_r^T \\ (\mathbf{X}^1)^T \\ (\mathbf{X}^2)^T \\ : \end{pmatrix} \mathbf{R}_i^t \mathbf{f}_{c,i}^t.$$

In practice, we will use the GAM estimation routines implemented in the *R* package *mgcv* to choose the penalization parameter and estimate the model at the same time.

3.6 HALSX with other regression models

We can replicate the strategy above to work with other regression models. Instead of HALSX-GAM, we can have HALSX-Regression-Tree, HALSX-Random-Forest, or HALSX-Kernel-Regression.

3.7 Stopping criterion

The Karush–Kuhn–Tucker conditions (KKT) of (3) are,

$$\begin{aligned} &\mathbf{XB} \geq \mathbf{0}, \mathbf{F}_c \geq \mathbf{0}, \\ &\nabla_{\mathbf{F}_c} f(\mathbf{B}, \mathbf{F}_c) \geq \mathbf{0}, \mathbf{F}_c \circ \nabla_{\mathbf{F}_c} f(\mathbf{B}, \mathbf{F}_c) = \mathbf{0}, \\ &\exists \mathbf{U} \in \mathbb{R}_+^{M \times K}, \nabla_{\mathbf{B}} f(\mathbf{B}, \mathbf{F}_c) - \mathbf{X}'\mathbf{U} = \mathbf{0}, \mathbf{U} \circ \mathbf{XB} = \mathbf{0}, \end{aligned}$$

where $\mathbf{A} \circ \mathbf{B}$ is the entry-wise product (Hadamard product) for $\mathbf{A}, \mathbf{B}$ of the same dimension.

The objective of the following sections is to find a way to show that matrices $\mathbf{B}$, $\mathbf{F}_c$ verifying the local optimality conditions similar to KKT conditions of a modified problem (necessary for the limit of a gradient descent algorithm) is necessaril close to the true model.

3.8 What about Estimator (4)?

Before detailing Algorithm 3 which aims to solve (4), we will first develop the elemental HALS iteration in the context of (4) where no supplemental information is supplied for the factorization model, namely $\mathbf{X}_r = \mathbf{I}_{n_1}, \mathbf{X}_c = \mathbf{I}_{n_2}$. Indeed, when updating one column of $\mathbf{F}_r$, the sub-problem becomes: how to solve $\arg\min_{\mathbf{f}} \|\mathbf{b} - \mathcal{A}(\mathbf{f}(\mathbf{f}_c)^T)\|_2^2$?

We will use the fact that for all $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$,

$$\mathcal{A}(\mathbf{M}) = (\langle \mathbf{A}_i, \mathbf{M} \rangle)_{1 \leq i \leq N},$$

and for all $\mathbf{b} \in \mathbb{R}^N$, $\mathcal{A}^*$, the transpose of $\mathcal{A}$ is defined by

$$\mathcal{A}^*(\mathbf{b}) = \sum_{i=1}^{N} b_i \mathbf{A}_i.$$

Since

$$\frac{\partial}{\partial \mathbf{f}} \|\mathbf{b} - \mathcal{A}(\mathbf{f}(\mathbf{f}_c)^T)\|_2^2 = \mathcal{A}^*[\mathcal{A}(\mathbf{f}(\mathbf{f}_c)^T) - \mathbf{b}]\mathbf{f}_c,$$

the first order optimality condition $\frac{\partial}{\partial \mathbf{f}} \|\mathbf{b} - \mathcal{A}(\mathbf{f}(\mathbf{f}_c)^T)\|_2^2 0$ is therefore equivalent to

$$\mathcal{A}^*[\mathcal{A}(\mathbf{f}(\mathbf{f}_c)^T)]\mathbf{f}_c = \mathcal{A}^*[\mathbf{b}]\mathbf{f}_c.$$

The left-hand side of the equation can be written as

$$\begin{aligned}
\mathcal{A}^*[\mathcal{A}(\mathbf{f}(\mathbf{f}_c)^T)]\mathbf{f}_c &= (\sum_{i=1}^{N} \langle \mathbf{A}_i, \mathbf{f}(\mathbf{f}_c)^T) \rangle \mathbf{A}_i)\mathbf{f}_c \\
&= \sum_{i=1}^{N} \mathrm{Tr}(\mathbf{f}(\mathbf{A}_i\mathbf{f}_c)^T)(\mathbf{A}_i\mathbf{f}_c) \\
&= \sum_{i=1}^{N} (\mathbf{A}_i\mathbf{f}_c)\mathrm{Tr}((\mathbf{A}_i\mathbf{f}_c)^T\mathbf{f}) \\
&= \sum_{i=1}^{N} (\mathbf{A}_i\mathbf{f}_c)(\mathbf{A}_i\mathbf{f}_c)^T\mathbf{f},
\end{aligned}$$

which leads to the following symmetric n_1-by-n_1 system on $\mathbf{f}$:

$$(\sum_{i=1}^{N} (\mathbf{A}_i\mathbf{f}_c)(\mathbf{A}_i\mathbf{f}_c)^T)\mathbf{f} = \sum_{i=1}^{N} b_i \mathbf{A}_i\mathbf{f}_c,$$

or

$$\mathbf{f} = (\sum_{i=1}^{N} (\mathbf{A}_i\mathbf{f}_c)(\mathbf{A}_i\mathbf{f}_c)^T)^{-1} \sum_{i=1}^{N} b_i \mathbf{A}_i\mathbf{f}_c$$

This computation generalizes to linear exogeneous variables, with the optimality condition:

$$\boldsymbol{\beta} = (\sum_{i=1}^{N} (\mathbf{X}\mathbf{A}_i\mathbf{f}_c)(\mathbf{X}\mathbf{A}_i\mathbf{f}_c)^T)^{-1} \sum_{i=1}^{N} b_i \mathbf{A}_i\mathbf{f}_c.$$

When the matrices to be inversed in the these equations are not invertible, we will use the generalized inverse instead.

Using these elementary steps, we propose Algorithm 3 to solve Problem (4). Compared to Algorithm 2, Algorithm 3

- has one less block (the slack variable $\mathbf{V}$ is not present);
- checks the deviation with data more frequently;
- each subproblem is more costly because of the presence of $\mathcal{A}$ in the subproblem. As we will see in the detailed development of the computation, when N, the sample size (the dimension of image of $\mathcal{A}$) is large, each update involves rather costly computations.

Algorithm 3 Hierarchical Alternating Least Squares with eXogeneous variables for NMF (HALSX2)

Require: Measurement operator $\mathcal{A}$, measurements $\mathbf{b}$, rank $1 \leq k \leq \min\{n_1, n_2\}$, features $\mathbf{X}_r$ and $\mathbf{X}_c$, functional spaces F_r and F_c in which to search the link functions.

Initialize $\mathbf{F}_r^0, \mathbf{F}_c^0 \geq 0, t = 0$
while Stopping criterion is not satisfied **do**
  $\mathbf{R}^t = \mathbf{b} - \mathcal{A}(\mathbf{F}_r^t(\mathbf{F}_c^t)^T)$
  for $i = 1, 2, ..., k$ **do**
    $\mathbf{R}^t = \mathbf{R}^t + \mathcal{A}(\mathbf{f}_{r,i}^t(\mathbf{f}_{c,i}^t)^T)$
    Calculate $f_{r,i}^{t+1} = \arg\min_{f \in F_r} \|\mathbf{R}^t - \mathcal{A}(f(\mathbf{X}_r)(\mathbf{f}_{c,i}^t)^T)\|_2^2$
    $\mathbf{f}_{r,i}^{t+1} = \max(0, f_{r,i}^{t+1}(\mathbf{X}_r))$
    $\mathbf{R}^t = \mathbf{R}^t - \mathcal{A}(\mathbf{f}_{r,i}^{t+1}(\mathbf{f}_{c,i}^t)^T)$
  end for
  for $i = 1, 2, ..., k$ **do**
    $\mathbf{R}^t = \mathbf{R}^t + \mathcal{A}(\mathbf{f}_{r,i}^{t+1}(\mathbf{f}_{c,i}^t)^T)$
    Calculate $f_{c,i}^{t+1} = \arg\min_{f \in F_c} \|\mathbf{R}^t - \mathcal{A}(\mathbf{f}_{r,i}^{t+1} f(\mathbf{X}_c)^T)\|_2^2$
    $\mathbf{f}_{c,i}^{t+1} = \max(0, f_{c,i}^{t+1}(\mathbf{X}_c))$
    $\mathbf{R}^t = \mathbf{R}^t - \mathcal{A}(\mathbf{f}_{r,i}^{t+1}(\mathbf{f}_{c,i}^{t+1})^T)$
  end for
  $t = t + 1$
end while
return $\mathbf{F}_r^t \in \mathbb{R}_+^{n_1 \times k}, f_{r,1}^t, ..., f_{r,k}^t \in F_r$,
$\mathbf{F}_c^t \in \mathbb{R}_+^{n_2 \times k}, f_{c,1}^t, ..., f_{c,k}^t \in F_c$.

Complexity In order to solve Problem (4) by Algorithm (3), at each sub-iteration we need to calculate N n_1-by-n_1 matrices $((\mathbf{A}_i\mathbf{f}_c)(\mathbf{A}_i\mathbf{f}_c)^T)$, then inverse the sum of the these N matrices. While the computation of the sum is map-reducible, on a single-threaded machine, this can be vey computationally expensive when N is large. Effectively, this second algorithm has a multiplicative complexity of $O(kN\max(n_1,n_2)^2)$, while the first estimator has a complexity of $O(N\max(n_1,n_2)^2)+O(kn_1n_2)$ with general linear measurement operator. With linear measurement operators with efficient projection methods, the complexity could become $O(N)+O(kn_1n_2)$. This computation cost will be discussed more in detail in the experiements described in the next section.

4 Experiments

5 Conclusion

References

Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *The Journal of Machine Learning Research*, 10:803–826, 2009.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global Optimality of Local Search for Low Rank Matrix Recovery. *arXiv:1605.07221 [cs, math, stat]*, May 2016.

Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5):2359–2388, October 2012. doi: 10.1214/12-AOS1039.

Emmanuel J. Candès and Benjamin Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. doi: 10.1007/s10208-009-9045-5.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, December 2013. doi: 10.1093/biomet/ast036.

Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix Completion with Noisy Side Information. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3447–3455. Curran Associates, Inc., 2015.

Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In *Independent Component Analysis and Signal Separation*, pages 169–176. Springer, 2007.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, page None, 2003.

Rina Foygel, Michael Horrell, Mathias Drton, and John D. Lafferty. Nonparametric reduced rank regression. In *Advances in Neural Information Processing Systems*, pages 1628–1636, 2012.

Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations Research Letters*, 26 (3):127–136, 2000.

Prateek Jain and Inderjit S. Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

V. Kekatos, Y. Zhang, and G. B. Giannakis. Electricity Market Forecasting via Low-Rank Multi-Kernel Learning. *IEEE Journal of Selected Topics in Signal Processing*, 8(6):1182–1193, December 2014. doi: 10.1109/JSTSP.2014.2336611.

Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

Hans Laurberg, Mads Græsbøll Christensen, Mark D. Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on Positive Data: On the Uniqueness of NMF. *Computational Intelligence and Neuroscience*, 2008:1–9, 2008. doi: 10.1155/2008/764206.

Jiali Mei, Yohann De Castro, Yannig Goude, and Georges Hébrail. Nonnegative Matrix Factorization for Time Series Recovery From a Few Temporal Aggregates, 2016.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011. doi: 10.1214/10-AOS860.

Si Si, Kai-Yang Chiang, Cho-Jui Hsieh, Nikhil Rao, and Inderjit S. Dhillon. Goal-Directed Inductive Matrix Completion. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2016.

Raja Velu and Gregory C. Reinsel. *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer Science & Business Media, April 2013. ISBN 978-1-4757-2853-8. Google-Books-ID: dsfSBwAAQBAJ.

Simon Wood. *Generalized Additive Models: An Introduction with R*. CRC press, 2006.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup Matrix Completion with Side Information: Application to Multi-Label Learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2301–2309. Curran Associates, Inc., 2013.

Or Zuk and Avishai Wagner. Low-Rank Matrix Recovery from Row-and-Column Affine Measurements. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2012–2020, 2015.