Nonnegative Matrix Factorization for Time Series Recovery From a Few Temporal Aggregates

Jiali Mei [1 2] **Yohann De Castro** [1] **Yannig Goude** [1 2] **Georges Hébrail** [2]

Abstract

Motivated by electricity consumption reconstitution, we propose a new matrix recovery method using nonnegative matrix factorization (NMF). The task tackled here is to reconstitute electricity consumption time series at a fine temporal scale from measures that are temporal aggregates of individual consumption. Contrary to existing NMF algorithms, the proposed method uses temporal aggregates as input data, instead of matrix entries. Furthermore, the proposed method is extended to take into account individual autocorrelation to provide better estimation, using a recent convex relaxation of quadratically constrained quadratic programs. Extensive experiments on synthetic and real-world electricity consumption datasets illustrate the effectiveness of the proposed method.

1. Introduction

In this paper, we propose a new matrix recovery method using nonnegative matrix factorization (NMF, Lee & Seung (1999)) where matrix columns represent time series at a fine temporal scale. Moreover, only temporal aggregates of these time series are observed.

The method has its motivation in the context of electricity load balancing, where time series represent electric power consumption. To avoid failure in the electricity network, suppliers are typically required by transmission system operators (TSO) to supply as much electricity as their consumers consume at every moment. This mechanism is called balancing. In the context of an open electricity market, all market participants, such as suppliers, utility traders, and large consumers, have a balance responsibility: any imbalance caused within the perimeter of a participant is billed by the TSO. To calculate the imbalance caused by a market participant, one needs an estimation of the consumption and production within its perimeter at a small temporal scale, for example, half-hourly (RTE (2014), SVK (2016), REE (2016)).

However, for many customers (for instance residential) within a perimeter, electricity consumption is not recorded at that scale. Although smart meter readings may be recorded locally up to every minute, utility companies often have very limited access to such data, due to data transmission and processing costs and/or privacy issues. Following a fixed schedule, cumulative consumption of each meter is recorded by the utility company, for instance every day or every month. By differentiating consecutive readings, the utility obtains the consumption of a customer between two reading dates. Currently, TSOs use proportional rules to reconstitute consumption from these measurements, based on national consumption profiles adjusted by temperature. In this article, we develop an NMF-based matrix recovery method providing a solution to consumption reconstitution from such temporal aggregates.

Recent advances in matrix completion have made it clear that when a large number of individuals and features are involved, even partial data could be enough to recover much of lost information, thanks to the low-rank property (Candès & Recht, 2009): although the whole data matrix $\mathbf{V}^* \in \mathbb{R}^{T\times N}$ is only partially known, if $\mathbf{V}^* = \mathbf{WH}$, where $\mathbf{W} \in \mathbb{R}^{T\times K}, \mathbf{H} \in \mathbb{R}^{K\times N}$, with K much smaller than both T and N, one could recover $\mathbf{V}^*$ entirely under some conditions over the sampling process.

In this article, we adress electricity consumption reconstitution as a matrix recovery problem. Consider the electricity consumption of N consumers during T periods. Since consumption is always positive, the N time series are organized into a nonnegative matrix $\mathbf{V}^* \in \mathbb{R}_+^{T\times N}$. An entry of this matrix, $v^*_{t,n}$ represents, for example, the electricity consumption of Consumer n for Period t.

Information about consumption is revealed as meter readings which do not correspond to matrix entries but to cumulative sums of each column of $\mathbf{V}^*$: at a meter-reading date t, we observe that Consumer n has consumed $\sum_{i=1}^{t} v^*_{i,n}$

[1]EDF Lab Paris-Saclay, 91120 Palaiseau, France [2]LMO, Univ. Paris-Sud, CNRS, Universite Paris-Saclay, 91405 Orsay, France. Correspondence to: Jiali Mei <jiali.mei@u-psud.fr>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

since the first period. Several readings are available for each consumer.

An alternative matrix representation could be to define entries directly as the cumulative consumption since the first period. Again, this matrix has missing values and a matrix completion algorithm can be applied. However, this cumulative matrix has increasing columns, which is quite different from matrices considered in the standard matrix completion literature, where matrix completion error is typically bounded by the upper bound on matrix.

We represent meter readings as linear measures on the consumption matrix $\mathbf{V}^*$. Temporal aggregates are derived from meter readings by differentiating consecutive readings: we will consider these temporal aggregates as "observations" in the rest of the paper. We consider D scalar observations, represented by a data vector $\mathbf{a} \equiv \mathcal{A}(\mathbf{V}^*) \in \mathbb{R}_+^D$, where $\mathcal{A}$ is a D-dimensional linear operator. To recover $\mathbf{V}^*$ from $\mathbf{a}$, we look for a low-rank NMF of $\mathbf{V}^*$: $\mathbf{WH} \simeq \mathbf{V}^*$, where $\mathbf{W} \in \mathbb{R}_+^{T\times K}, \mathbf{H} \in \mathbb{R}_+^{K\times N}$. The columns of $\mathbf{W}$ are K nonnegative factors, which can be interpreted as typical profiles of the N time series, and the columns of $\mathbf{H}$ as the weights of each individual. The problem is formalized as the minimization of a quadratic loss function under nonnegativity and data constraints:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{W},\mathbf{H}} \quad & \ell(\mathbf{V},\mathbf{W},\mathbf{H}) = \|\mathbf{V}-\mathbf{WH}\|_F^2 \\ \text{s.t.} \quad & \mathbf{V}\geq\mathbf{0}, \quad \mathbf{W}\geq\mathbf{0}, \quad \mathbf{H}\geq\mathbf{0}, \quad \mathcal{A}(\mathbf{V})=\mathbf{a}, \end{aligned} \tag{1}$$

where $\mathbf{X} \geq \mathbf{0}$ (or $\mathbf{x} \geq \mathbf{0}$) means that the matrix $\mathbf{X}$ (or the vector $\mathbf{x}$) is element-wise nonnegative.

Note the difference between (1) and another potential estimator,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \quad & \hat{\ell}(\mathbf{W},\mathbf{H}) = \|\mathcal{A}(\mathbf{WH})-\mathbf{a}\|_2^2 \\ \text{s.t.} \quad & \mathbf{W}\geq\mathbf{0}, \quad \mathbf{H}\geq\mathbf{0}, \end{aligned} \tag{2}$$

studied in (Roughan et al., 2012). If $\mathbf{V}$ is a solution to (1), it satisfies exactly the measurement constraint, but is approximately low-rank, while $\mathbf{WH}$, a solution to (2) is exactly of low rank, but only matches the measurements approximately. Since in our application, the estimated time series matrix is to be used for billing, the match to metering data is essential. Therefore, we use (1) in this work.

1.1. Prior works

The measurement operator $\mathcal{A}$ is a special instance of the trace regression model (Rohde & Tsybakov, 2011) which generalizes the matrix completion setting. In matrix completion, each measurement is exactly one entry. Various forms of linear measurements other than matrix completion have been considered for matrix recovery without nonnegativity (Recht et al., 2010; Candès & Plan, 2011; Zuk & Wagner, 2015).

The NMF literature is generally focused on full observation (Gillis, 2014; Alquier & Guedj, 2016), or on matrix completion (Gillis & Glineur, 2011; Xu & Yin, 2013) Random projection measurements are used in an NMF context in (Pnevmatikakis & Paninski, 2013), where a maximum likelihood estimator is developed based on a specific generative model in neural imaging. The particular form of measurement operator considered here arises from meter reading, and can be used in other fields, such as Internet traffic matrix estimation (Roughan et al., 2012). Because of our choice of estimator (1) over estimator (2), we derive a novel algorithm for this measurement operator, which has a smaller time complexity than previously studied ones (more details in Section 2.1).

In real-world applications, global information such as temporal autocorrelation could be available in addition to measurements. Previous approaches combining matrix factorization and autoregressive structure are often focused on obtaining factors that are more smooth and/or sparse, both in NMF (Chen & Cichocki, 2005; Févotte & Idier, 2011; Smaragdis et al., 2014) and without nonnegativity (Udell et al., 2016; Yu et al., 2015). Our objective is different from these studies: we try to further improve the matrix recovery by constraining temporal correlation on individual time series (not factors). We use a recent convex relaxation of quadratically constrained quadratic programs (Ben-Tal & den Hertog, 2013) to deduce a closed-form projection step in this case.

We propose an algorithm to solve (1) in Section 2.1. To take into account individual autocorrelation, a second algorithm is proposed in Section 2.2. In Section 3, both algorithms are validated on synthetic and real electricity consumption datasets, compared to a linear benchmark and a state-of-art matrix completion method.

2. Reconstitution of time series with NMF

2.1. Iterative algorithm with simplex projection

We represent temporal aggregation by a linear operator $\mathcal{A}$. For each $1 \leq d \leq D$, the d-th measurement on $\mathbf{X}$, $\mathcal{A}(\mathbf{X})_d$, is the sum of several consecutive rows on one column of $\mathbf{X}$, that is,

$$\mathcal{A}(\mathbf{X})_d = \sum_{(t,n)\in I_d} x_{t,n},$$

where $I_d = \{(t,n)|t_0(d)+1 \leq t \leq t_0(d)+h(d), n = n_d\}$, is the index set over $h(d)$ consecutive periods of Consumer n_d, starting from Period $t_0(d)+1$. Each measurement covers a disjoint index set. All entries of $\mathbf{X}$ are not necessarily involved in the measurements.

A block Gauss-Seidel algorithm (Algorithm 1) is used to solve (1). We alternate by minimizing $\ell(\mathbf{V}, \mathbf{W}, \mathbf{H})$ over $\mathbf{W}, \mathbf{H}$ or $\mathbf{V}$, keeping the other two matrices fixed. Methods from classical NMF problems are used to update $\mathbf{W}$ and $\mathbf{H}$ (Kim et al., 2014). In the implementation, we use two variants that seem similarly efficient (more details in Section 3): Hierarchical Alternating Least Squares (HALS, Cichocki et al. (2007)), and a matrix-base NMF solver with Nesterov-type acceleration (NeNMF, Guan et al. (2012)).

When $\mathbf{W}$ and $\mathbf{H}$ are fixed, the optimization problem on $\mathbf{V}$ is equivalent to D simplex projection problems, one for each scalar measurement. For $1 \leq d \leq D$, we have to solve

$$\begin{aligned} \min_{\mathbf{v}_{I_d}} \quad & \|\mathbf{v}_{I_d} - \sum_{t=t_0(d)+1}^{t_0(d)+h(d)} w_t \mathbf{h}_{n_d}\|^2 \\ \text{s.t.} \quad & \mathbf{v}_{I_d} \geq 0, \quad \mathbf{v}'_{I_d}\mathbf{1} = b_d. \end{aligned} \tag{3}$$

The simplex projection algorithm introduced by Chen & Ye (2011) solves this subproblem efficiently. Define the operator, $\mathcal{P}_A$, as the orthogonal projection into the simplex $A \equiv \{\mathbf{X} \in \mathbb{R}_+^{T\times N} | \mathcal{A}(\mathbf{X}) = \mathbf{a}\}$. A is the intersection of the affine subspace $\{\mathbf{X} \in \mathbb{R}^{T\times N} | \mathcal{A}(\mathbf{X}) = \mathbf{a}\}$ and the first orthant. Projector $\mathcal{P}_A$ encodes the measurement data $\mathbf{a} = \mathcal{A}(\mathbf{V}^*)$. In Algorithm 1, we apply $\mathcal{P}_A$ to a working value of $\mathbf{V}$ in order to obtain its projection in A.

Contrary to previously studied algorithms (Roughan et al., 2012), by choosing estimator (1) over (2), the simplex projection step is separated from the classical NMF update steps in our algorithm. Instead of multiplying the rank and the complexity introduced by the number of measurements, we have an algorithm whose complexity is the sum of the two. In cases where the number of measurements is large, this difference can be crucial. [1]

Algorithm 1 Block coordinate descent for NMF from temporal aggregates

input $\mathcal{P}_A, 1 \leq K \leq \min\{T, N\}$
  Initialize $\mathbf{W}^0, \mathbf{H}^0 \geq 0, \mathbf{V}^0 = \mathcal{P}_A(\mathbf{W}^0\mathbf{H}^0), i = 0$
  while Stopping criterion is not satisfied **do**
    $\mathbf{W}^{i+1} =$ Update$(\mathbf{W}^i, \mathbf{H}^i, \mathbf{V}^i)$
    $\mathbf{H}^{i+1} =$ Update$(\mathbf{W}^{i+1}, \mathbf{H}^i, \mathbf{V}^i)$
    $\mathbf{V}^{i+1} = \mathcal{P}_A(\mathbf{W}^{i+1}\mathbf{H}^{i+1})$
    $i = i + 1$
  end while
output $\mathbf{V}^i \in A, \mathbf{W}^i \in \mathbb{R}_+^{T\times K}, \mathbf{H}^i \in \mathbb{R}_+^{K\times N}$

A classical stopping criterion in the NMF literature is based on Karush-Kuhn-Tucker (KKT) conditions on (1) (Gillis, 2014, Section 3.1.7). We calculate

$$\begin{aligned} \mathcal{R}(\mathbf{W})_{i,j} &= |(\mathbf{W}\mathbf{H} - \mathbf{V})\mathbf{H}')_{i,j}| \mathbb{1}_{\mathbf{W}_{i,j}\neq 0}, \\ \text{and } \mathcal{R}(\mathbf{H})_{i,j} &= |(\mathbf{W}'(\mathbf{W}\mathbf{H} - \mathbf{V}))_{i,j}| \mathbb{1}_{\mathbf{H}_{i,j}\neq 0}. \end{aligned}$$

The algorithm is stopped if $\|\mathcal{R}(\mathbf{W})\|_F^2 + \|\mathcal{R}(\mathbf{H})\|_F^2 \leq \epsilon$, for a small threshold $\epsilon > 0$.

Convergence to a stationary point has been proved for past NMF solvers with the full observation or the matrix completion setting (Guan et al. (2012); Kim et al. (2014)). Our algorithms have similar convergence property. Although the subproblems on $\mathbf{W}$ and $\mathbf{H}$ do not necessarily have unique optimum, the projection of $\mathbf{V}$ attains a unique minimizer. By Grippo & Sciandrone (2000, Proposition 5), the convergence to a stationary point is guaranteed.

Algorithm 1 can be generalized to other types of measurement operators $\mathcal{A}$, as long as a projection into the simplex defined by the data constraint $\mathcal{A}(\mathbf{X}) = \mathbf{a}$ and the positivity constraint can be efficiently computed.

2.2. From autocorrelation constraint to penalization

In addition to the measurements in $\mathbf{a}$, we have some prior knowledge on the temporal autocorrelation of the individuals. To take into account information about autocorrelation, we add a penalization term to the original matrix recovery problem, replacing (1) by:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{W},\mathbf{H}} \quad & \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2 - \lambda \sum_{n=1}^{N} \mathbf{v}'_n \mathbf{\Delta}_{\rho_n} \mathbf{v}_n \\ \text{s.t.} \quad & \mathbf{V} \geq 0, \quad \mathbf{W} \geq 0, \quad \mathbf{H} \geq 0, \quad \mathcal{A}(\mathbf{V}) = \mathbf{a}, \end{aligned} \tag{4}$$

where $\lambda \geq 0$ is a single fixed penalization parameter, and $\mathbf{\Delta}_{\rho_n}$ is a symmetric matrix precised shortly after. In the rest of this section, we first show by Theorem 1, that with an appropriately chosen value of λ, adding the penalization term $\mathbf{v}'_n \mathbf{\Delta}_{\rho_n} \mathbf{v}_n$ is equivalent to impose that the temporal autocorrelation of $\mathbf{v}_n$ to be at least equal to ρ_n, a prior threshold. Then we modify the Algorithm 1 to solve this penalized problem.

For $1 \leq n \leq N$, suppose that the lag-1 autocorrelation of Individual n's time series is at least equal to a threshold ρ_n (e.g. from historical data, excluded from observed temporal aggregates), that is,

$$\sum_{t=1}^{T-1} v_{t+1,n} v_{t,n} \geq \rho_n \sum_{t=1}^{T} v_{t,n}^2. \tag{5}$$

[1] This intuition is confirmed by the comparison of Algorithm 1 and our implementation of the algorithm proposed in (Roughan et al., 2012).

Notice that with the lag matrix,

$$\mathbf{\Delta} = \begin{pmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ 0 & 0 & 0 & \ddots & : \\ : & : & \ddots & \ddots & 1 \\ 0 & 0 & ... & 0 & 0 \end{pmatrix},$$

we have $\sum_{t=1}^{T-1} v_{t+1,n} v_{t,n} = \mathbf{v}_n' \mathbf{\Delta} \mathbf{v}_n$. Define $\mathbf{\Delta}_\rho \equiv \mathbf{\Delta} + \mathbf{\Delta}' - 2\rho\mathbf{I}$, for a threshold $-1 \leq \rho \leq 1$. Inequality (5) is then equivalent to

$$\mathbf{v}_n' \mathbf{\Delta}_{\rho_n} \mathbf{v}_n \geq 0. \tag{6}$$

Imposing (6) would require one to solve, at each iteration, N quadratically constrained quadratic programs (QCQP) of the form:

$$\begin{aligned} \min_{\mathbf{x}} \quad & ||\mathbf{x} - \mathbf{x}_0||^2 \\ \text{s.t.} \quad & \mathbf{x}'\mathbf{S}\mathbf{x} \geq 0, \end{aligned} \tag{7}$$

where $\mathbf{S}$ is a general symmetric matrix, not necessarily semi-definite positive. This means that the QCQP is in general a non-convex problem. Let $\boldsymbol{\delta}$ be the vector of eigenvalues of $\mathbf{\Delta}$. By eigendecomposition, $\mathbf{S} = \mathbf{U}'\mathbf{D}\mathbf{U}$. where the matrix $\mathbf{U}$ is orthogonal. The entries of $\boldsymbol{\delta}$ are the diagonal entries of $\mathbf{D}$. The following theorem justifies the choice of penalization term in (4), by showing with an appropriate λ, adding this penalization term is equivalent to imposing the autocorrelation constraint (5).

Theorem 1. *Suppose that δ_1, the largest eigenvalue of $\mathbf{S}$, is strictly positive. Suppose that $\mathbf{z}_0 \equiv \mathbf{U}\mathbf{x}_0$ has no zero component. Then there exists $0 \leq \lambda < \frac{1}{\delta_1}$, that verifies $\sum_{t=1}^{T} \delta_t \frac{z_{0,t}^2}{2(1-\lambda\delta_t)^2} = 0$, so that $\mathbf{x}^* \equiv (\mathbf{I} - \lambda\mathbf{S})^{-1}\mathbf{x}_0$ is an optimal solution of (7).*

Proof. We follow Ben-Tal & den Hertog (2013) to obtain a convex relaxation of (7).

Define $\mathbf{z} \equiv \mathbf{U}\mathbf{x}$, $\mathbf{z}_0 \equiv \mathbf{U}\mathbf{x}_0$, $y_t \equiv \frac{1}{2}z_t^2$, $\forall 1 \leq t \leq T$. Recall that $\delta_1 > 0$, and that $\forall t, 1 \leq t \leq T, z_{0,t} \neq 0$.

Problem (7) is equivalent to the non-convex problem

$$\begin{aligned} \min_{\mathbf{y},\mathbf{z}} \quad & \mathbf{1}'\mathbf{y} - \mathbf{z}_0'\mathbf{z} \\ \text{s.t.} \quad & -\boldsymbol{\delta}'\mathbf{y} \leq 0, \quad \frac{1}{2}z_t^2 = y_t, \quad \forall 1 \leq t \leq T. \end{aligned} \tag{8}$$

Now consider its convex relaxation

$$\begin{aligned} \min_{\mathbf{y},\mathbf{z}} \quad & \mathbf{1}'\mathbf{y} - \mathbf{z}_0'\mathbf{z} \\ \text{s.t.} \quad & -\boldsymbol{\delta}'\mathbf{y} \leq 0, \quad \frac{1}{2}z_t^2 - y_t \leq 0, \quad \forall 1 \leq t \leq T. \end{aligned} \tag{9}$$

By Ben-Tal & den Hertog (2013, Theorem 3), if $(\mathbf{z}^*, \mathbf{y}^*)$ is an optimal solution of (9), and if $\frac{1}{2}(z_t^*)^2 = y_t^*, \forall 1 \leq t \leq T$, then $(\mathbf{z}^*, \mathbf{y}^*)$ is also an optimal solution of (8), which makes $\mathbf{x}^* = \mathbf{U}'\mathbf{z}^*$ an optimal solution of (7).

We will look for such a solution to (9) by examining its first-order conditions of optimality. Problem (9) is convex, and it verifies the Slater condition: $\exists(\hat{\mathbf{y}}, \hat{\mathbf{z}}), -\boldsymbol{\delta}'\hat{\mathbf{y}} < 0, \frac{1}{2}\hat{z}_t^2 < \hat{y}_t, \forall 1 \leq t \leq T$. This is true, because $\delta_1 > 0$. We could choose an arbitrary value of $\hat{y}_1 > 0$ and strictly positive but small values for other components of $\hat{\mathbf{y}}$ so as to have $-\boldsymbol{\delta}'\hat{\mathbf{y}} < 0$, and $\hat{\mathbf{z}} = \mathbf{0}$. Thus, Problem (9) always has an optimal solution, because the objective function is coercive over the constraint. This shows the existence of $(\mathbf{z}^*, \mathbf{y}^*)$.

Now we show that $\frac{1}{2}(z_t^*)^2 = y_t^*, \forall 1 \leq t \leq T$. The KKT conditions of (9) are verified by $(\mathbf{z}^*, \mathbf{y}^*)$. In particular, there is some dual variable $\lambda \geq 0, \boldsymbol{\mu} \in \mathbb{R}_+^T$ that verifies,

$$\mathbf{1} - \lambda\boldsymbol{\delta} - \boldsymbol{\mu} = \mathbf{0}, \tag{10}$$

$$-\boldsymbol{\delta}'\mathbf{y}^* \leq 0, \tag{11}$$

$$\lambda\boldsymbol{\delta}'\mathbf{y}^* = 0, \tag{12}$$

$$-z_{0,t} + \mu_t z_t^* = 0, \quad \forall 1 \leq t \leq T, \tag{13}$$

$$\frac{1}{2}(z_t^*)^2 - y_t^* \leq 0, \quad \forall 1 \leq t \leq T, \tag{14}$$

$$\mu_t(\frac{1}{2}(z_t^*)^2 - y_t^*) = 0, \quad \forall 1 \leq t \leq T. \tag{15}$$

Since $z_{0,t} \neq 0$, we have $\mu_t \neq 0, z_t^* = \frac{1}{\mu_t} z_{0,t}, \forall 1 \leq t \leq T$ by (13). Therefore, by (15), $y_t^* = \frac{1}{2}(z_t^*)^2, \forall 1 \leq t \leq T$.

By (10), the values of $\mu_t = 1 - \lambda\delta_t$ can be deduced from that of λ. Since $\boldsymbol{\mu} > \mathbf{0}$, we obtain that $\lambda < \frac{1}{\delta_1}$.

By (13), $z_t^* = \frac{z_{0,t}}{1-\lambda\delta_t}, \forall 1 \leq t \leq T$. This shows that $\mathbf{x}^* = \mathbf{U}'\mathbf{z}^* = \mathbf{U}'(\mathbf{I} - \lambda\mathbf{D})^{-1}\mathbf{z}_0^* = (\mathbf{I} - \lambda\mathbf{S})^{-1}\mathbf{x}_0$ is an optimal solution for (7).

Theorem 1 shows that with a well chosen λ, the constraint in (7) can be replaced by a penalization.

There are in fact two cases. Either $\mathbf{x}_0$ verifies the constraint, in which case $\lambda = 0$, and $\mathbf{x}^* = \mathbf{x}_0$ is the solution. Otherwise, $\lambda > 0$. We replug the values of

$$y_t^* = \frac{1}{2}(z_t^*)^2 = \frac{z_{0,t}^2}{2(1-\lambda\delta_t)^2}$$

back into (12), and obtain that λ verifies

$$\sum_{t=1}^{T} \frac{\delta_t z_{0,t}^2}{2(1-\lambda\delta_t)^2} = 0.$$

□

When $\mathbf{W}$ and $\mathbf{H}$ are fixed, the subproblem of (4) on $\mathbf{V}$ can

be separated into N constrained problems of the form,

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}-\mathbf{x}_0\|^2 - \lambda \mathbf{x}' \mathbf{\Delta}_{\rho_n} \mathbf{x}, \\ \text{s.t.} \quad & \mathbf{A}_n \mathbf{x} = \mathbf{c}_n, \\ & \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{16}$$

where $\mathbf{x}_0$ is the n-th column of $\mathbf{WH}$, $\mathbf{c}_n$ is the observations on the n-th column, and $\mathbf{A}_n$ is a matrix which encodes the measurement operator over that column. The following theorem shows how to solve this problem.

Theorem 2. *Suppose that* $\mathbf{S}$ *is a symmetric matrix with eigenvalues* $\boldsymbol{\delta}$*, and* $\lambda_1 > 0$*. Suppose* $\mathbf{A} \in \mathbb{R}^{m,l}$ *a full-rank matrix with* $m \leq l$*,* $\mathbf{x}_0 \in \mathbb{R}^l$*,* $\mathbf{c} \in \mathbb{R}^m$*,* $\lambda \geq 0$*. Define* $\mathbf{Q} \equiv (\mathbf{I}-\lambda\mathbf{S})^{-1}\mathbf{A}'(\mathbf{A}(\mathbf{I}-\lambda\mathbf{S})^{-1}\mathbf{A}')^{-1}$*. If* $\lambda < \frac{1}{\delta_{\rho,1}}$*, then* $\mathbf{Qc} + (\mathbf{I}-\mathbf{QA})(\mathbf{I}-\lambda\mathbf{S})^{-1}\mathbf{x}_0$ *is a minimizer of*

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}-\mathbf{x}_0\|^2 - \lambda \mathbf{x}' \mathbf{S} \mathbf{x}, \\ \textit{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{c}, \end{aligned} \tag{17}$$

Proof. Let l be the dimension of $\mathbf{c}$. Define I_C as the indicator function for the constraint of (17), that is

$$\begin{aligned} I_C(\mathbf{x}) = 0, \text{ if } \mathbf{Ax} = \mathbf{c}, \\ \text{and } I_C(\mathbf{x}) = +\infty, \text{ if } \mathbf{Ax} \neq \mathbf{c}. \end{aligned}$$

Problem (17) is then equivalent to

$$\min_{\mathbf{x}} F(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{x}_0\|^2 - \frac{1}{2}\lambda \mathbf{x}'\mathbf{S}\mathbf{x} + I_C(\mathbf{x}). \tag{18}$$

The subgradient of (18) is $\partial F(\mathbf{x}) = \{\mathbf{x} - \mathbf{x}_0 - \lambda\mathbf{S}\mathbf{x} - \mathbf{A}'\boldsymbol{\epsilon} | \boldsymbol{\epsilon} \in \mathbb{R}^l\}$. When $\lambda < \frac{1}{\delta_1}$, (18) is convex. Therefore, $\mathbf{x}^*$ is a minimizer if and only if $0 \in \partial F(\mathbf{x})$, and $\mathbf{Ax}^* = \mathbf{c}$. That is, $\exists \boldsymbol{\epsilon} \in \mathbb{R}^l$,

$$\begin{aligned} (\mathbf{I}-\lambda\mathbf{S})\mathbf{x}^* - \mathbf{x}_0 - \mathbf{A}'\boldsymbol{\epsilon} = \mathbf{0}, \\ \mathbf{A}\mathbf{x}^* = \mathbf{c}. \end{aligned}$$

The vector $\boldsymbol{\epsilon}$ thereby verifies $\mathbf{A}(\mathbf{I}-\lambda\mathbf{S})^{-1}(\mathbf{x}_0 + \mathbf{A}'\boldsymbol{\epsilon}) = \mathbf{c}$.

The l-by-l matrix $\mathbf{A}(\mathbf{I}-\lambda\mathbf{S})^{-1}\mathbf{A}'$ is invertible, because l is smaller than m, and $\mathbf{A}$ is of full rank (because each measurement covers disjoint periods). Therefore,

$$\begin{aligned} \boldsymbol{\epsilon} &= (\mathbf{A}(\mathbf{I}-\lambda\mathbf{S})^{-1}\mathbf{A}')^{-1}(\mathbf{c} - \mathbf{A}(\mathbf{I}-\lambda\mathbf{S})^{-1}\mathbf{x}_0), \\ \mathbf{x}^* &= (\mathbf{I}-\lambda\mathbf{S})^{-1}(\mathbf{x}_0 + \mathbf{A}'\boldsymbol{\epsilon}) \\ &= \mathbf{Qc} + (\mathbf{I}-\mathbf{QA})(\mathbf{I}-\lambda\mathbf{S})^{-1}\mathbf{x}_0. \square \end{aligned}$$

In our particular problem, the eigenvalues of $\mathbf{\Delta}_{\rho_n}$ are

$$\delta_{\rho_n,t} = 2\cos(\frac{t}{T+1}\pi) - 2\rho_n,$$

with t taking every value from 1 to T. This means that for most of the autocorrelation threshold that we could need to impose ($-1 \leq \rho_n \leq 1$), $\mathbf{\Delta}_{\rho_n}$ has both strictly positive and strictly negative eigenvalues, allowing the above theorems to apply.

Both $\mathbf{I} - \lambda\mathbf{\Delta}_{\rho_n}$ and $\mathbf{A}_n(\mathbf{I}-\lambda\mathbf{\Delta}_{\rho_n})^{-1}\mathbf{A}_n'$ are invertible with $\lambda < \delta_{\rho_n,1}$. The matrix inversion only needs to be done once for each individual. After computing $\mathbf{Q}_n \equiv (\mathbf{I} - \lambda\mathbf{\Delta}_{\rho_n})^{-1}\mathbf{A}_n'(\mathbf{A}_n(\mathbf{I}-\lambda\mathbf{\Delta}_{\rho_n})^{-1}\mathbf{A}_n')^{-1}$, $\mathbf{Q}_n\mathbf{c}_n$ and $(\mathbf{I} - \mathbf{Q}_n\mathbf{A}_n)(\mathbf{I}-\lambda\mathbf{\Delta}_{\rho_n})^{-1}$ for each n, we use Algorithm 2 to solve (4).

Algorithm 2 Block coordinate descent for NMF from temporal aggregates and autocorrelation penalty

input $\rho_n, \mathbf{A}_n, \mathbf{Q}_n, \mathbf{Q}_n\mathbf{c}_n, \forall 1 \leq n \leq N$, and $1 \leq K \leq \min\{T, N\}$
  Initialize $\mathbf{W}^0, \mathbf{H}^0 \geq 0, \mathbf{V}^0 = \mathcal{P}_A(\mathbf{W}^0\mathbf{H}^0), i = 0$
  while Stopping criterion is not satisfied **do**
    $\mathbf{W}^{i+1} = \text{Update}(\mathbf{W}^i, \mathbf{H}^i, \mathbf{V}^i)$
    $\mathbf{H}^{i+1} = \text{Update}(\mathbf{W}^{i+1}, \mathbf{H}^i, \mathbf{V}^i)$
    for all $1 \leq n \leq N$ **do**
      $\mathbf{v}_n^{i+1} = (\mathbf{Q}_n\mathbf{c}_n + (\mathbf{I} - \mathbf{Q}_n\mathbf{A}_n)(\mathbf{I} - \lambda\mathbf{\Delta}_{\rho_n})^{-1}\mathbf{W}^{i+1}\mathbf{h}_n^{i+1})_+$
    end for
    $i = i + 1$
  end while
output $\mathbf{V}^i \in A, \mathbf{W}^i \in \mathbb{R}_+^{T\times K}, \mathbf{H}^i \in \mathbb{R}_+^{K\times N}$

Choosing λ An optimal value of λ could be calculated. Substituting the values of $\mathbf{y}^*$ in (12), shows that the optimal λ is a root of the polynomial $\sum_{t=1}^{T} \delta_{\rho,t}\frac{z_{0,t}^2}{2(1-\lambda\delta_{\rho,t})^2}$. The root-finding is too expensive to perform at every iteration. However, the optimal λ verifies

$$0 < \lambda < \frac{1}{\delta_{\rho,1}},$$

where $\delta_{\rho,1} = 2\cos(\frac{1}{T+1}\pi) - 2\rho$ is the biggest eigenvalue of $\mathbf{\Delta}_\rho$. This gives us a good enough idea about how large a λ to use. In the numerical experiments, we chose $\lambda = \min(1, \frac{1}{2\max_n \delta_{\rho_n,1}})$ in the penalization when the constraint in (7) is active, and $\lambda = 0$ (no penalization) when the constraint is verified by $\mathbf{x}_0$.

3. Experimental results

We use one synthetic dataset and three real-world electricity consumption datasets to evaluate the proposed algorithms. In each dataset, the individual autocorrelation is calculated on historical data from the corresponding datasets, not used for evaluation.

- **Synthetic data**: 20 independent Gaussian processes with Matern covariance function (shifted to be nonnegative) are sampled over 150 periods to form the

factor matrix $\mathbf{W}$. A 20-by-120 weight matrix $\mathbf{H}$ is generated by sampling from a standard normal distribution truncated at 0, independantly for each entry. The data matrix is obtained as $\mathbf{W} \times \mathbf{H}$ ($T = 150, N = 120$). This matrix is exactly of rank 20.

- **French electricity consumption** (proprietary dataset): daily consumption of 636 medium-voltage feeders gathering each around 1,500 consumers based near Lyon in France during 2012 ($T = 365, N = 636$).
- **Portuguese electricity consumption** (Trindade, 2016) daily consumption of 370 Portuguese clients during 2014 ($T = 365, N = 370$).
- **Electricity consumption of small Irish companies** (Commission for Energy Regulation, Dublin, 2011a;b) daily consumption of 426 small Irish companies during 200 days in 2010 ($T = 200, N = 426$).

For each individual in a dataset, we generate observations by selecting a number of observation periods. The temporal aggregates are the sum of the time series between two consecutive observation periods. The observation periods are chosen in two possible ways: periodically (at regular intervals with the first observation period sampled at random), or uniformly at random. The regular intervals for periodic observations are $p \in \{2, 3, 5, 7, 10, 15, 30\}$. This is motivated by the real application where meter readings are recorded regularly. With random observations, we use sampling rates that are equivalent to the regular intervals. That is, the number of observations D verifies $\frac{D}{TN} = \frac{1}{p} \in \{0.5, 0.33, 0.2, 0.14, 0.1, 0.07, 0.03\}$.

We apply the following methods to recover the data matrix from each set of sampled observations:

- **interpolation** Temporal aggregates are distributed equally over the covered periods.
- **softImpute** As an alternative method, we apply a state-of-art matrix completion algorithm to complete the cumulative matrix. The observed entries are the cumulative values of the column from the first period to the observation dates. We use a nuclear-norm minimization algorithm, implemented in the **R** package, **softImpute** (Mazumder et al. (2010)), to complete the cumulative consumption matrix, before differentiating each column to obtain recovered matrix. To choose the thresholding parameter, we use the warm start procedure documented in softImpute.
- **HALS**, and **NeNMF** These are the proposed matrix recovery algorithms using two classical $\mathbf{W}$ and $\mathbf{H}$ update implementations: HALS, and NeNMF. When autocorrelation penalization is used, we choose $\lambda = \min(1, \frac{1}{2 \max_n \delta_{\rho_n,1}})$, as explained in the previous section. The rank used in proposed algorithms is chosen by a 5-fold cross validation procedure: we split the observations randomly into 5 folds, and apply the algorithm to 4 of the 5 folds with ranks $2 \leq K \leq 30$. We then calculate the ℓ_2-distance between the temporal aggregates on the recovered matrix with the 1-fold holdout. Repeating this procedure onto the 5 folds separately, we choose the rank which minimizes the average ℓ_2-distance, to perform the algorithm on all observations.

With a recovered matrix $\mathbf{V}$ obtained in an algorithm run, we compute the relative root-mean-squared error (RRMSE):

$$\text{RRMSE}(\mathbf{V}, \mathbf{V}^*) = \frac{\|\mathbf{V} - \mathbf{V}^*\|_F}{\|\mathbf{V}^*\|_F}.$$

Each experiment (dataset, sampling scheme, sampling rate, recovery method, unpenalized or penalized) is run three times, and the average RRMSE is reported in Figure 1. The figure is zoomed to show the RRMSE of the proposed algorithms. Much higher error rates for reference methods are sometimes not shown.

On sample sets with random observation periods (lower panel), proposed methods (HALS and NeNMF, blue and purple lines), whether unpenalized (solid lines) or penalized (dashed lines), out-performs the interpolation benchmark (red solid lines) by large in all datasets. This is especially the case when the sampling rate is small, *i.e.* when the task is more difficult. On the Irish dataset (lower panel, furthest to the right), penalized HALS and NeNMF (dashed blue and purple lines) are an improvement to unpenalized HALS and NeNMF when the sampling rate is low.

With periodic observations (upper panel), the RRMSE is higher for every method. Proposed unpenalized methods, HALS and NeNMF (blue and purple solid lines) are equivalent to interpolation benchmark (red solid lines) for synthetic data, but sometimes worse for real datasets. Real electricity consumption has significant weekly periodicity, which is poorly captured by observations at similar periods. However, this shortcoming of the unpenalized method is more than compensated for by the penalization (dashed blue and purple lines).

We notice that penalized HALS and NeNMF consistently outperform interpolation with both observation schemes. This makes penalized methods particularly useful for the application of electricity consumption reconstituion, where it may be costly to install a random observation scheme, or to change the current periodic observation scheme to a random one.

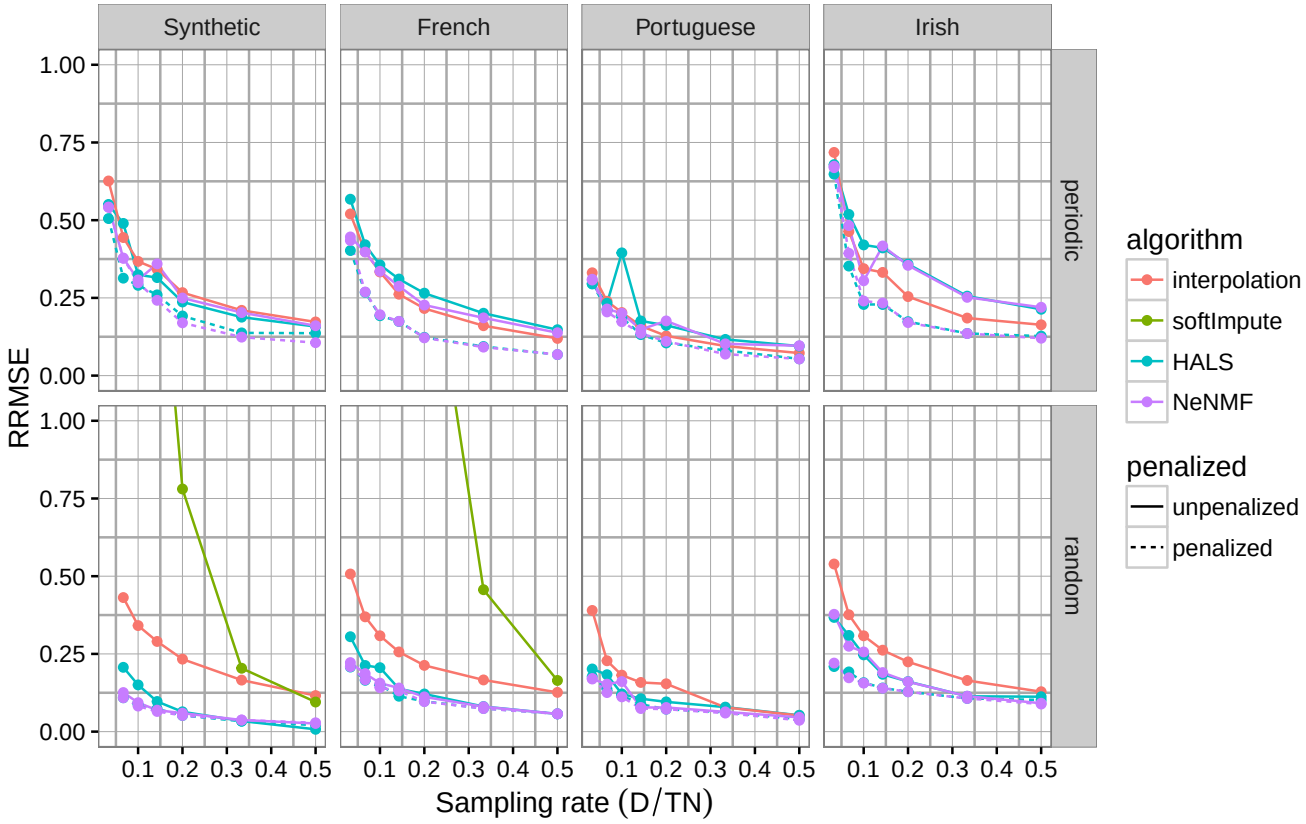

Figure 1. Mean RRMSE of the recovered matrices over three separate runs over the four datasets. On the samples with random observation periods, proposed methods (HALS and NeNMF, blue and purple lines, both penalized and unpenalized) out-performs the interpolation benchmark (solid red line). On the samples with regular observation periods, unpenalized HALS and NeNMF (solid blue and purple lines) are similar to the interpolation benchmark, whiled penalized HALS and NeNMF (dashed blue and purple lines) are an important improvement. The *softImpute* method (solid green line) only has comparable performance in two of the datasets, in the easiest task (50% sampling rate at random periods). In most cases, RRMSE of *softImpute* is larger than 100%.

It is also interesting to note that the rank chosen by the cross validation procedure is higher in higher sampling rate scenarios (Figure 2). This shows that the cross validation procedure is able to relax the rank constraint when more information is available in the data.

The traditional matrix completion method seems to fail in this application: softImpute (green solid lines) only has comparable results to interpolation or proposed methods in two of the four datasets, with 50% sampling rate in the random sampling scheme, which is the easiest case. In most cases, *softImpute* has an RRMSE much larger than 100%, and thus is not shown in the graphic. This indicates that the cumulative matrix considered in this application does not verify assumptions which guarantee matrix completion success.

4. Perspectives

Motivated by a new industrial application, we extended NMF to use temporal aggregates as input data, by adding a projection step into NMF algorithms. With appropriate projection algorithms, this approach could be further generalized to other types of data, such as disaggregating spatially aggregated data, or general linear measures. When such information is available, we introduce a penalization on individual autocorrelation, which improves the recovery performance of the base algorithm. This component can be generalized to larger lags (with a matrix $\mathbf{\Delta}$ with 1's further off the diagonal), or multiple lags (by adding several lag matrices together). It is also possible to generalize this approach to other types of expert knowledge through additional constraints on $\mathbf{V}$.

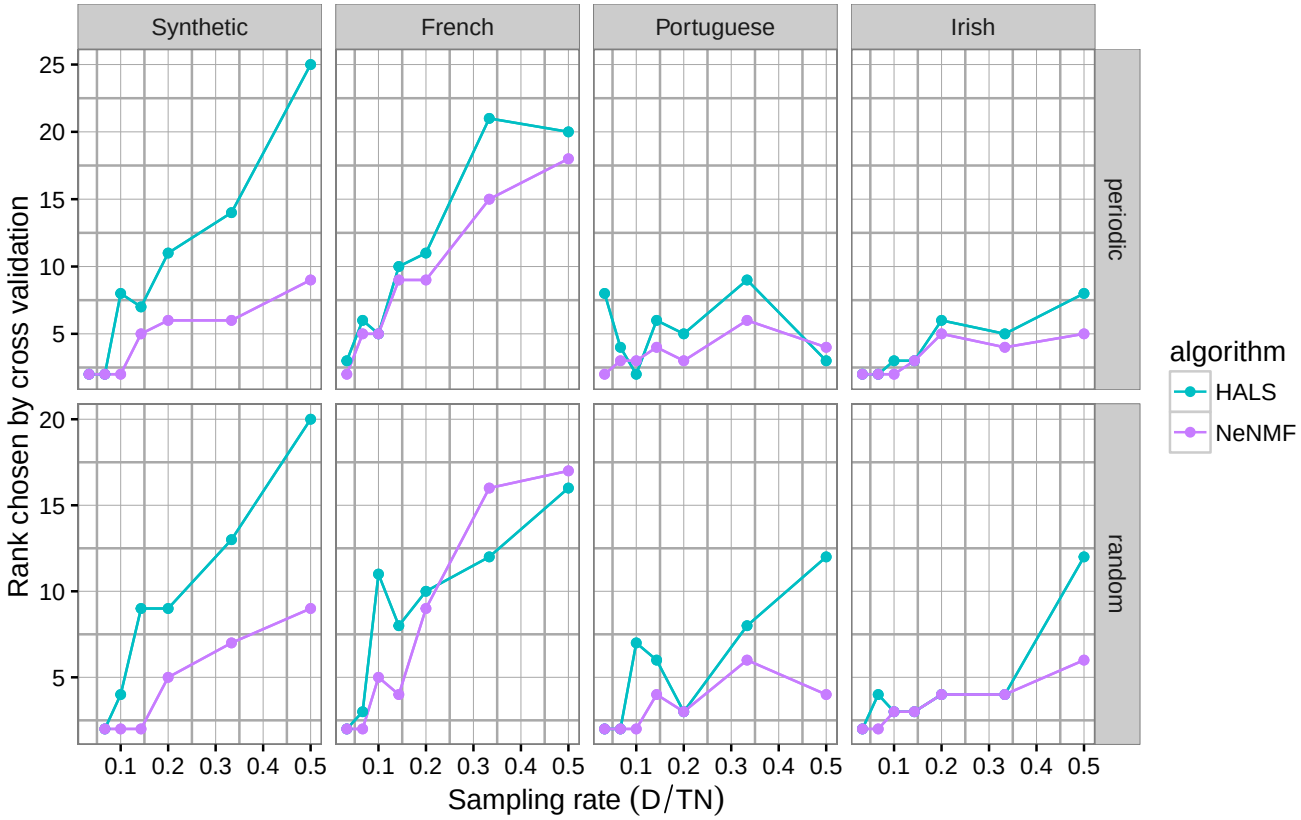

Figure 2. The rank chosen by the cross validation procedure generally increases with the sampling rate, for the four datasets. This shows that procedure is able to relax the rank constraint when more information is available in the data.

Acknowledgements

We thank Jean-Marc Azaïs for his input and the anonymous reviewers for pointing out applications with similar problem settings.

References

Alquier, Pierre and Guedj, Benjamin. An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization. *axXiv preprint arXiv:1601.01345*, 2016.

Ben-Tal, Aharon and den Hertog, Dick. Hidden conic quadratic representation of some nonconvex quadratic optimization problems. *Mathematical Programming*, 143(1-2):1–29, 2013. doi: 10.1007/s10107-013-0710-8.

Candès, Emmanuel J. and Plan, Yaniv. Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011. doi: 10.1109/TIT.2011.2111771.

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

Chen, Yunmei and Ye, Xiaojing. Projection Onto A Simplex. *arXiv preprint arXiv:1101.6081*, 2011.

Chen, Zhe and Cichocki, Andrzej. Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints. *Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep*, 68, 2005.

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

Commission for Energy Regulation, Dublin. Electricity smart metering customer behaviour trials findings report. 2011a.

Commission for Energy Regulation, Dublin. Results of electricity coast-benefit analysis, customer behaviour trials and technology trials commission for energy regulation. 2011b.

Févotte, Cédric and Idier, Jérôme. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

Gillis, Nicolas. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12:257, 2014.

Gillis, Nicolas and Glineur, François. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM Journal on Matrix Analysis and Applications*, 32 (4):1149–1165, 2011.

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

Guan, Naiyang, Tao, Dacheng, Luo, Zhigang, and Yuan, Bo. NeNMF: An Optimal Gradient Method for Nonnegative Matrix Factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012. doi: 10.1109/TSP.2012.2190406.

Kim, Jingu, He, Yunlong, and Park, Haesun. 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.

Lee, Daniel D. and Seung, H. Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Mazumder, Rahul, Hastie, Trevor, and Tibshirani, Robert. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

Pnevmatikakis, Eftychios A and Paninski, Liam. Sparse nonnegative deconvolution for compressive calcium imaging: Algorithms and phase transitions. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 1250–1258. Curran Associates, Inc., 2013.

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

REE. Balance responsible party. `https://www.esios.ree.es/en/glossary\#letterB`, 2016.

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

Roughan, M., Zhang, Y., Willinger, W., and Qiu, L. Spatio-Temporal Compressive Sensing and Internet Traffic Matrices (Extended Version). *IEEE/ACM Transactions on Networking*, 20(3):662–676, June 2012. doi: 10.1109/TNET.2011.2169424.

RTE. Balance Responsible Entity System. `http://clients.rte-france.com/lang/an/clients_producteurs/services_clients/dispositif_re.jsp`, May 2014.

Smaragdis, Paris, Févotte, Cédric, Mysore, Gautham J., Mohammadiha, Nasser, and Hoffman, Matthew. Static and Dynamic Source Separation Using Nonnegative Factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, 2014. doi: 10.1109/MSP.2013.2297715.

SVK. Balance responsibility. `http://www.svk.se/en/stakeholder-portal/Electricity-market/Balance-responsibility/`, October 2016.

Trindade, Artur. UCI Maching Learning Repository - ElectricityLoadDiagrams20112014 Data Set. `http://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014`, 2016.

Udell, Madeleine, Horn, Corinne, Zadeh, Reza, and Boyd, Stephen. Generalized Low Rank Models. *Foundations and Trends in Machine Learning*, 9(1), 2016. doi: 10.1561/2200000055.

Xu, Yangyang and Yin, Wotao. A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion. *SIAM Journal on Imaging Sciences*, 6 (3):1758–1789, 2013. doi: 10.1137/120887795.

Yu, Hsiang-Fu, Rao, Nikhil, and Dhillon, Inderjit S. High-dimensional Time Series Prediction with Missing Values. *arXiv preprint arXiv:1509.08333*, 2015.

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