Time series forecasting from partial observations via Non-negative Matrix Factorization

Anonymous authors
Paper under double-blind review

Abstract

In modern time series problems, one aims at forecasting multiple times series with possible missing and noisy values. In this paper, we introduce the Sliding Mask Method (SMM) for forecasting multiple nonnegative time series by means of nonnegative matrix completion: observed noisy values and forecast/missing values are collected into matrix form, and learning is achieved by representing its rows as a convex combination of a small number of nonnegative vectors, referred to as the archetypes. We introduce two estimates, the mask Archetypal Matrix factorization (mAMF) and the mask normalized Nonnegative Matrix Factorization (mNMF) which can be combined with the SMM method. We prove that these estimates recover the true archetypes with an error proportional to the noise. We use a proximal alternating linearized method (PALM) to compute the archetypes and the convex combination weights. We compared our estimators with state-of-the-art methods (Transformers, LSTM, SARIMAX...) in multiple time series forecasting on real data and obtain that our method outperforms them in most of the experiments.

1 Introduction

This article investigates forecasting multiple nonnegative times series with missing or noisy entries. We observe $N \geq 1$ time series $\mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(N)} \in \mathbb{R}^T$ over a period of time of length $T \geq 1$ and we would like to forecast the next $F \geq 1$ future values by means of matrix completion, see Figure 1.

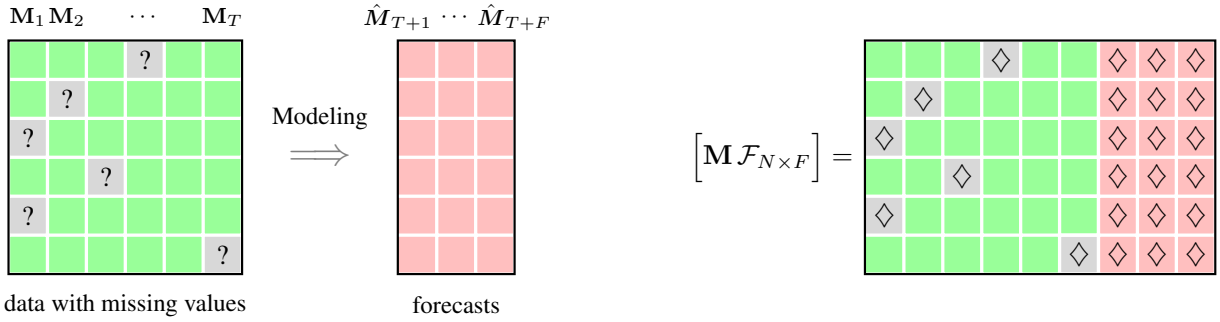

Figure 1: `[Left]` Consider noisy multiple time series observations (green) with possible missing entries (question mark) from $N \geq 1$ time series and their F forecast values in red. `[Right]` Matrix completion problem under consideration: missing and forecast values $\mathcal{F}_{N\times F}$ (gray and red diamonds) are not observed.

The matrix completion problem depicted on the right hand side of Figure 1 is ill-posed, it is not possible to complete the red values by any low-rank technique. To address this issue, we consider a linear transform $\mathbf{\Phi}$ defined as follows.

- We denote by $P \geq 1$ some time series length parameter referred to as the *periodicity*. However, our analysis does not assume any periodicity in the data and the practitioner is free to chose any value for P. If some periodicity appears in the data, she can choose P as a multiple of this periodicity. Note that our method works even if P is not chosen in this way.

- We partition the $T+F$ columns (right side of Figure 1) into B sub-blocks of length P so that $B = (T+F)/P$. Padding at most $P-1$ forecasts columns to the right of the matrix, one can assume, without loss of generality, that B is an integer.
- Now, the rows of the output matrix have length WP, gathering W consecutive row sub-blocks together. It amounts in sliding a window of length WP by jumps of length P on successive rows of the input matrix.
- We assume that W and P are such that $WP > F$. Up to row permutation, one obtain an output four blocks matrix whose bottom right block is the forecast values, see Figure 2.

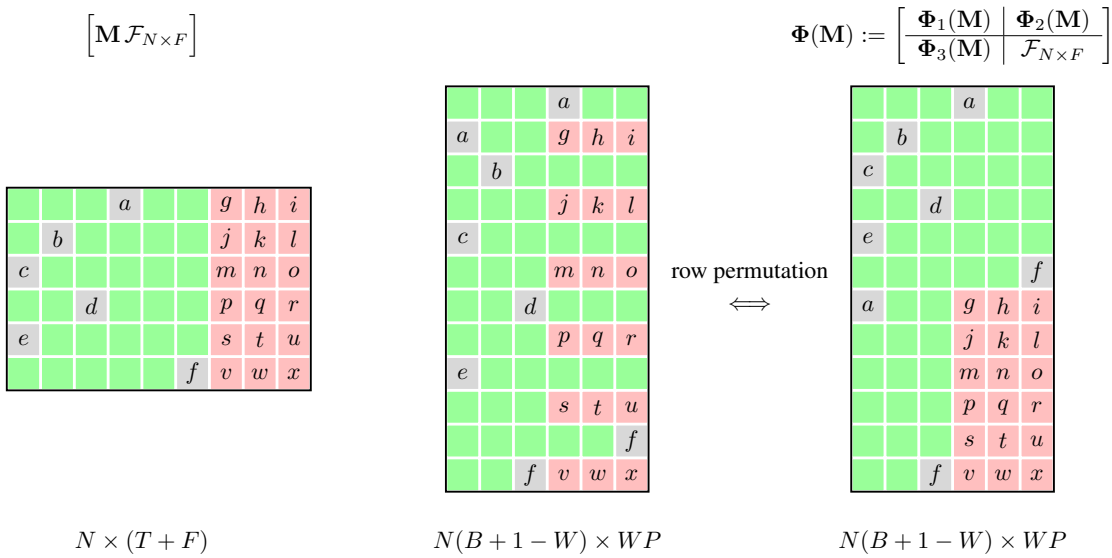

Figure 2: Given an input matrix of size $N \times (T+F)$, the output sliding mask matrix is composed of 4 blocks denoted by $\mathbf{\Phi}_1(\mathbf{M})$, $\mathbf{\Phi}_2(\mathbf{M})$, $\mathbf{\Phi}_3(\mathbf{M})$, and $\mathcal{F}_{N\times F}$, this latter being of size $N \times F$ with the F last columns of the input matrix. In this example, we consider a periodicity of $P = 3$, giving $B = 3$ sub-blocks per row of the input matrix and we gather $W = 2$ consecutive sub-blocks in an output row. To ease readability, we denote by $a, \ldots, f$ the missing values and by $g, \ldots, x$ the values to forecast. After row permutation, we obtain the output SMM. The matrix $\mathbf{\Phi}_1(\mathbf{M})$ (resp. $\mathbf{\Phi}_2(\mathbf{M})$, $\mathbf{\Phi}_3(\mathbf{M})$) is a sub-matrix of size $N(B-W) \times (WP-F)$ (resp. $N(B-W) \times F$, $N \times (WP-F)$) of the input matrix $\mathbf{M}$.

We define our matrix completion problems on the observation matrix $\mathbf{X}$ using the mask operator $\mathbf{T}$ defined below.

Definition 1 (Observation matrix $\mathbf{X}$) *Let $n := N(B+1-W)$ and $p := WP$ such that $p > F$. Given the* input *matrix $\mathbf{M}$ of size $N \times T$, define the* observation *matrix $\mathbf{X}$ as the matrix of size $n \times p$ obtained by dropping out the forecast and missing entries (gray and red entries in Figure 2) and keeping the observed values of $\mathbf{M}$ (green entries).*

Definition 2 (Mask operator $\mathbf{T}$) *Given any matrix $\mathbf{N}$ of size $n \times p$, the mask operator $\mathbf{T}(\mathbf{N})$ is defined by dropping out the forecast and missing entries (red and gray entries in Figure 2). Namely, we define*

$$\mathbf{T}(\mathbf{N}) = \left[\begin{array}{c|c}\mathbf{N}_1 & \mathbf{N}_2\\ \hline \mathbf{N}_3 & \mathbf{0}_{N\times F}\end{array}\right] \quad \textit{and} \quad \mathbf{T}^{\perp}(\mathbf{N}) = \left[\begin{array}{c|c}\mathbf{0} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{N}_4\end{array}\right] \quad \textit{where } \mathbf{N} = \left[\begin{array}{c|c}\mathbf{N}_1 & \mathbf{N}_2\\ \hline \mathbf{N}_3 & \mathbf{N}_4\end{array}\right]. \tag{1}$$

and $\mathbf{N}_1$ (resp. $\mathbf{N}_2$, $\mathbf{N}_3$ and $\mathbf{N}_4$) is a sub-matrix of size $N(B-W)\times(WP-F)$ (resp. $N(B-W)\times F$, $N\times(WP-F)$ and $N \times F$) of the input matrix $\mathbf{N}$.

While the algorithms presented in this paper can handle arbitrary missing entries by modifying the mask operator $\mathbf{T}$, our theoretical analysis focuses on the block-structured missing data pattern shown in (1). A rigorous analysis for more general patterns of missing values is an important direction for future work but is beyond the scope of this paper. With the above definitions, one has $\mathbf{X} = \mathbf{T}(\mathbf{\Phi}(\mathbf{M}))$, where $\mathbf{\Phi}(\mathbf{M})$ is given in Figure 2.

Definition 3 (mask normalized Nonnegative Matrix Factorization (mNMF)) *Given $K \geq 1$ referred to as the* nonnegative rank*, we introduce the mask normalized Nonnegative Matrix Factorization (mNMF):*

$$\min \left\{ \|\mathbf{N} - \mathbf{W}\mathbf{H}\|_F^2 \; : \; \mathbf{W}\mathbf{1} = \mathbf{1}, \mathbf{W} \geq \mathbf{0} \, , \; \mathbf{H} \geq \mathbf{0} \, , \; \mathbf{T}(\mathbf{N}) = \mathbf{X} \right\} , \tag{mNMF}$$

where $\mathbf{1}$ is the vector of ones, the weight matrix $\mathbf{W}$ *has size $(n \times K)$, $\mathbf{H}$ has size $(K \times p)$ and $\| \cdot \|_F$ is the Frobenius norm.*

Solutions $\mathbf{N}$ to (mNMF) are such that $\mathbf{T}(\mathbf{N}) = \mathbf{X}$, the values of $\mathbf{N}$ corresponding to the green entries in Figure 2 are equal to observed values.

Definition 4 (mask Archetypal Matrix Factorization (mAMF)) *We introduce the mask Archetypal Matrix Factorization (mAMF):*

$$\min \left\{ \|\mathbf{N} - \mathbf{W}\mathbf{H}\|_F^2 + \lambda \|\mathbf{H} - \mathbf{V}\mathbf{N}\|_F^2 \; : \; \mathbf{W} \geq \mathbf{0} \, , \; \mathbf{W}\mathbf{1} = \mathbf{1} \, , \; \mathbf{V} \geq \mathbf{0} \, , \; \mathbf{V}\mathbf{1} = \mathbf{1} \, , \; \mathbf{T}(\mathbf{N}) = \mathbf{X} \right\} , \tag{mAMF}$$

where $\lambda \geq 0$ is a tuning parameter and $\mathbf{V}$ has size $(K \times n)$.

Definition 5 (Normalization, nonnegative and archetype basis) *The matrix $\mathbf{W}$ satisfies the constraint $\mathbf{W} \geq \mathbf{0}$ and $\mathbf{W}\mathbf{1} = \mathbf{1}$, this being later referred to as normalization. Its rows $(w_{i,1}, \ldots, w_{i,K})$ are convex combination weights and each row of $\mathbf{W}\mathbf{H}$ is a convex combination of the K rows of $\mathbf{H}$. The matrix $\mathbf{H}$, referred to as the* nonnegative basis (resp. archetype basis) *in* (mNMF) *(resp.* (mAMF)*).*

In (mAMF) *as $\lambda \to \infty$, the* archetypes *(defined as the K rows of $\mathbf{H}$) are forced to lie in the convex combination of $\mathbf{N}$ by means of the matrix $\mathbf{V}$. Since $\mathbf{N}$ is a completion of the observation matrix $\mathbf{X}$ by the mask operator $\mathbf{N} = \mathbf{T}(\mathbf{X})$, the matrix $\mathbf{V}$ can be interpreted as the convex combination weights of the decomposition of the archetypes onto the rows of the observation matrix $\mathbf{X}$ and hence we recover the method of Cutler & Breiman (1994).*

Hence we get the following decomposition of the ith row of $\mathbf{W}\mathbf{H}$,

$$(\mathbf{W}\mathbf{H})^{(i)} = \sum_{k=1}^{K} w_{i,k} \mathbf{H}_k \, . \tag{2}$$

Once solved, the above matrix problems give forecast values to the original forecasting problem of time series by means of matrix $\hat{\mathbf{M}}$ defined below.

Definition 6 (Forecasts of the orignial problem) *Forecasts $\hat{\mathbf{M}} \in \mathbb{R}^{N \times F}$ are given by the bottom right $N \times F$ submatrix of $\mathbf{W}\mathbf{H}$, namely $\hat{\mathbf{M}}$ is the block 4 in the decomposition (1). The fourth block of $\mathbf{\Phi}(\mathbf{M})$ is depicted in red in Figure 2 which is the same block as $\mathcal{F}_{N \times F}$ in the original forecasting problem (letters g to x in Figure 2), hence $\hat{\mathbf{M}}$ can be interpreted as forecast values.*

1.1 Mask nonnegative matrix completion statistical guarantees

Our goal is to solve the following nonnegative matrix completion problem: We observe a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ containing the multiple time values, given by the transformation presented in Figure 2. The missing values and the forecast values are arbitrarily set to zero, *i.e.*, $\mathbf{T}^{\perp}(\mathbf{X}) = \mathbf{0}$. This choice is *not restrictive* since the values of $\mathbf{X}$ corresponding to the missing and forecast entries are *not observed* and our study is *insensitive* to the values of these entries. Our target is defined by the following best approximation of $\mathbf{X}$ through the mask operator $\mathbf{T}$.

Definition 7 (Best normalized nonnegative rank K approximation of $\mathbf{X}$) *Given a nonnegative rank K, we define the following best approximation of $\mathbf{X}$:*

$$\mathbf{X}_0 := \mathbf{W}_0 \mathbf{H}_0 \in \arg \min_{\substack{\mathbf{W}_0 \mathbf{1} = \mathbf{1} \\ \mathbf{W}_0 \geq \mathbf{0} \, , \, \mathbf{H}_0 \geq \mathbf{0}}} \left\{ \left\| \mathbf{X} - \mathbf{T}(\mathbf{W}_0 \mathbf{H}_0) \right\|_F^2 \right\} , \tag{3a}$$

where $\mathbf{W}_0 \in \mathbb{R}^{n \times K}$ and $\mathbf{H}_0 \in \mathbb{R}^{K \times p}$. As K grows, the approximation error $\|\mathbf{X} - \mathbf{T}(\mathbf{W}_0 \mathbf{H}_0)\|_F$ decreases. The matrix $\mathbf{X}_0$ is referred to as the best normalized nonnegative rank K approximation of $\mathbf{X}$.

The goal is to recover the matrices $\mathbf{W}_0$ (weights) and $\mathbf{H}_0$ (archetypes) from the observation matrix $\mathbf{X}$. The observation can be written as

$$\mathbf{X} = \mathbf{T}(\mathbf{X}_0) + \mathbf{F}\,, \tag{4}$$

where $\mathbf{F}$ is some additive error term, referred to as the noise.

Contributions Sliding Mask Method (SMM) inputs the forecast values and it can be viewed as a nonnegative matrix completion algorithm under low nonnegative rank assumption. This framework raises two issues. A first question is the uniqueness of the decomposition, also referred to as *identifiability* of the model. In Theorem 13, we introduce a new condition that ensures uniqueness from partial observation of the target matrix. An other challenge, as pointed out by Vavasis (2009) for instance, is that solving *exactly* NMF decomposition problem is NP-hard. Nevertheless NMF-type problems can be solved efficiently using (accelerated) proximal gradient descent method Parikh & Boyd (2013) for block-matrix coordinate descent in an *alternating projection scheme*, *e.g.*, Javadi & Montanari (2020a) and references therein. We rely on these techniques to introduce algorithms inputting the forecast values based on NMF decomposition, see Section 3. Theorem 16 complements the theoretical analysis by proving the robustness of NMF-type algorithms when entries are missing or corrupted by noise. Our main theoretical contributions are as follows:

- A uniqueness decomposition result (Theorem 13) showing that the decomposition $\mathbf{W}_0\mathbf{H}_0$ is unique *given partial observations*, namely

$$\text{If} \quad \mathbf{T}(\mathbf{W}\mathbf{H}) = \mathbf{T}(\mathbf{W}_0\mathbf{H}_0) \quad \text{then} \quad (\mathbf{W}, \mathbf{H}) \equiv (\mathbf{W}_0, \mathbf{H}_0)\,, \tag{$\mathbb{P}_{\mathrm{u}}$}$$

  where $\equiv$ means up to positive scaling and permutation[1].

- A robustness result (Theorem 16 and Corollary 17) showing that (mNMF) and (mAMF) recover $\mathbf{H}_0$ and $\mathbf{W}_0$ with an error proportional to $\|\mathbf{F}\|_F$ (hence we recover $\mathbf{X}_0$ with the same precision).

Our analysis is completed by an algorithmic and numerical study that

- introduces a Proximal Alternating Linearized Minimization (PALM) method to solve (mAMF) and shows that PALM reaches a stationary point (Theorem 19).

- reports a performance improvement of (mNMF) and (mAMF) against state-of-the-art algorithms on real datasets for RRMSE and RMPE (Table 3). The relative root-mean-squared error (RRMSE) and the relative mean-percentage error (RMPE) are defined by

$$\text{RRMSE} = \frac{\|\mathbf{M}_F - \mathbf{M}_F^\star\|_F}{\|\mathbf{M}_F^\star\|_F}\,,\ \text{RMPE} = \frac{\|\mathbf{M}_F - \mathbf{M}_F^\star\|_1}{\|\mathbf{M}_F^\star\|_1}\,.$$

  where $\mathbf{M}_F^\star$ are the true values and $\mathbf{M}_F$ the forecasts (see Section 4).

Comments on low rank modeling and periodicity in time series Sparse or Low-Rank representations are ubiquitous in applications and well studied in the literature. In our analysis a time series is cut into several smaller W sub-blocks time series with the same length $p = WP$. For instance, observing sales over a period of one year, one can consider 52 weekly time series (one per week). These observations are the rows of our observed matrix $\mathbf{X}$. The normalized nonnegative low rank hypothesis assumes that the p-length multiple time series of the dataset can be decomposed as a sum of K basis time series $\mathbf{H}$ plus an error term. Of course, this error term can incorporate the model approximation error as depicted in (3a). The K basis time series $\mathbf{H}$ are learned on the entire dataset $\mathbf{X}$. This technique can be seen as dimension reduction, each observation can be summarized as K weights $\mathbf{W}$ such that the resulting convex combination of basis time series (2) is a good approximation of the observation $\mathbf{X}$.

The low rank hypothesis can be interpreted as a periodicity assumption. Indeed, if the time series are exactly periodic with period p, then the rank of the data matrix $\mathbf{X}$ is at most p. While P is a free parameter, the model's performance

[1] if an entry-wise nonnegative pair $(\mathbf{W}, \mathbf{H})$ is given then $(\mathbf{W}\mathbf{P}\mathbf{D}, \mathbf{D}^{-1}\mathbf{P}^\top\mathbf{H})$ is also a nonnegative decomposition $\mathbf{W}\mathbf{H} = \mathbf{W}\mathbf{P}\mathbf{D} \times \mathbf{D}^{-1}\mathbf{P}^\top\mathbf{H}$, where $\mathbf{D}$ scales and $\mathbf{P}$ permutes the columns (resp. rows) of $\mathbf{W}$ (resp. $\mathbf{H}$)

Algorithms	mAMF		mNMF		RFR		EXP		SARIMAX		LSTM		GRU	
Metrics	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE
daily electricity	14.42%	36.85%	15.86%	46.66%	12.16%	47.78%	11.25%	43.83%	9.85%	43.16%	12.42%	46.49%	12.03%	45.90%
weekly electricity	14.80%	17.50%	11.09%	13.79%	7.25%	8.61%	10.07%	7.98%	9.05%	**7.42%**	27.85%	15.64%	26.04%	15.92%
gas	**21.71%**	**18.55%**	37.46%	42.79%	66.80%	71.61%	63.35%	68.16%	45.58%	52.83%	62.97%	68.38%	62.87%	67.90%
Istanbul	15.67%	17.80%	**14.18%**	16.77%	15.37%	18.32%	15.46%	18.64%	14.75%	17.01%	16.22%	20.96%	20.01%	26.87%
ETTh1	**10.24%**	15.23%	12.30%	14.16%	12.96%	17.98%	12.37%	13.65%	13.36%	15.94%	14.86%	18.78%	14.71%	18.85%
ETTh2	9.42%	13.07%	**4.87%**	**6.66%**	6.47%	7.60%	14.06%	13.67%	12.76%	13.03%	14.17%	13.75%	14.44%	14.36%
ETTm1	10.12%	15.22%	**9.94%**	**12.25%**	12.81%	17.42%	11.45%	14.20%	12.29%	16.45%	13.39%	17.96%	14.13%	18.63%
ETTm2	8.19%	11.65%	**5.08%**	7.41%	5.81%	**7.16%**	13.18%	12.88%	13.16%	12.95%	14.29%	13.89%	14.46%	14.03%
electricity1	**6.59%**	13.17%	11.11%	15.28%	12.75%	16.09%	38.27%	34.44%	>100.00%	>100.00%	8.51%	10.13%	7.19%	**9.61%**
electricity2	**8.09%**	16.82%	8.82%	**12.17%**	12.05%	15.67%	47.40%	40.36%	43.38%	38.98%	9.05%	12.83%	10.30%	13.28%
electricity3	10.57%	13.95%	12.43%	14.04%	12.45%	14.14%	40.37%	33.48%	37.05%	33.01%	10.70%	11.62%	10.70%	11.02%
electricity4	11.02%	24.30%	25.07%	29.71%	23.16%	19.50%	54.42%	43.63%	63.08%	46.59%	12.18%	13.88%	**9.53%**	**11.05%**
electricity5	9.52%	19.05%	**7.72%**	**15.48%**	25.96%	26.84%	56.76%	49.31%	*	*	21.92%	28.79%	20.73%	27.02%
electricity6	10.11%	17.04%	14.30%	18.62%	13.81%	16.26%	51.87%	37.35%	52.10%	40.56%	7.58%	11.74%	**7.13%**	**10.32%**
electricity7	**8.34%**	16.75%	37.49%	30.03%	29.51%	22.96%	53.00%	45.95%	*	*	17.74%	14.72%	16.55%	**14.19%**
electricity8	**10.03%**	17.49%	23.81%	20.59%	19.33%	17.98%	36.83%	40.46%	38.54%	41.23%	12.16%	**15.73%**	13.89%	17.08%
electricity9	19.45%	38.90%	21.15%	41.72%	18.18%	37.53%	35.90%	38.65%	>100.00%	>100.00%	18.00%	37.77%	18.80%	38.21%
electricity10	**5.13%**	12.53%	5.40%	11.29%	12.11%	13.42%	33.88%	34.89%	36.55%	38.92%	7.66%	10.25%	7.77%	9.94%
synthetic1	6.40%	9.30%	5.81%	**8.01%**	5.79%	9.44%	**5.73%**	9.32%	5.76%	8.81%	6.67%	11.87%	6.77%	11.89%
synthetic2	19.04%	20.28%	20.35%	25.30%	17.12%	20.24%	21.09%	25.87%	28.41%	35.53%	21.55%	29.04%	21.48%	28.94%
low-noise	**0.10%**	**0.26%**	0.10%	0.26%	8.65%	22.76%	16.97%	48.02%	0.19%	0.33%	16.52%	46.03%	16.76%	46.46%
medium-noise	2.41%	5.23%	**1.92%**	**4.81%**	8.42%	21.95%	15.94%	44.25%	1.97%	4.94%	15.66%	42.68%	15.67%	42.98%
high-noise	12.69%	28.04%	**10.39%**	**26.43%**	11.73%	30.26%	13.02%	33.27%	15.24%	27.37%	13.03%	33.67%	12.97%	33.47%

Algorithms	BasisFormer		Autoformer		iTransformer		PatchMLP		TimeMixer	
Metrics	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE	RRMSE	RMPE
daily electricity	**7.56%**	6.64%	39.65%	80.25%	28.85%	61.63%	8.30%	**6.39%**	32.15%	68.83%
weekly electricity	8.76%	9.07%	41.34%	80.93%	37.27%	75.07%	**6.99%**	7.57%	44.07%	82.74%
gas	57.45%	52.10%	99.37%	>100.00%	>100.00%	>100.00%	–	–	–	–
Istanbul	14.83%	**12.54%**	80.00%	89.26%	>100.00%	>100.00%	18.77%	18.93%	>100.00%	>100.00%
ETTh1	14.57%	**13.61%**	52.37%	34.35%	42.74%	30.95%	19.94%	25.41%	36.92%	24.39%
ETTh2	54.66%	53.66%	66.85%	73.31%	65.10%	69.41%	53.59%	53.68%	56.72%	60.34%
ETTm1	13.58%	12.31%	49.72%	34.94%	36.77%	28.11%	21.28%	26.48%	36.94%	23.70
ETTm2	55.52%	54.95%	59.79%	66.20%	64.78%	68.69%	53.89%	53.38%	57.21%	61.53%
electricity1	26.93%	28.19%	49.69%	83.08%	36.84%	80.29%	11.97%	12.29%	58.63%	91.57%
electricity2	35.38%	39.73%	49.00%	52.77%	38.62%	47.73	10.42%	11.48%	56.70%	55.77%
electricity3	34.30%	37.25%	52.13%	90.29%	38.04%	68.80%	**9.13%**	**9.12%**	62.73%	79.92%
electricity4	39.42%	40.66%	46.88%	81.34%	44.36%	85.28%	15.52%	15.80%	54.53%	90.54%
electricity5	46.22%	49.60%	49.16%	43.74%	55.43%	44.32%	13.94%	13.36%	53.67%	42.78%
electricity6	45.50%	46.86%	50.51%	73.42%	46.08%	79.51%	18.58%	17.00%	50.77%	75.83%
electricity7	40.17%	43.20%	86.87%	>100.00%	84.60%	>100.00%	16.72%	17.27%	85.16%	>100.00%
electricity8	30.64%	30.99%	62.89%	98.00%	47.05%	81.41%	13.62%	14.21%	62.71%	95.12%
electricity9	34.88%	35.85%	40.79%	27.79%	25.32%	26.44%	**11.32%**	**11.08%**	54.67%	29.59%
electricity10	29.78%	31.79%	54.90%	>100.00%	35.75%	>100.00%	8.15%	**8.51%**	66.22%	>100.00%
synthetic1	8.06%	12.32%	> 100.00%	>100.00%	>100.00%	>100.00%	6.52%	11.74%	96.27%	94.14%
synthetic2	31.32%	47.88%	59.75%	83.66%	60.08%	>100.00%	**1.33%**	**7.53%**	55.61%	69.21%
low-noise	23.71%	51.94%	95.23%	80.40%	7.18%	13.73%	10.28%	27.49%	>100.00%	>100.00%
medium-noise	21.68%	48.48%	83.34%	73.79%	16.49%	30.96%	9.66%	27.36%	>100.00%	>100.00%
high-noise	18.44%	45.77%	>100.00%	>100.00%	94.25%	>100.00%	13.49%	39.92%	>100.00%	>100.00%

Figure 3: Best results in **bold**, and second best results underlined. Some values have not been evaluated due to numerical errors on matrix inversion, reported as a star symbol $*$ or because the time-series length is too short for long-term forecasting with default parameters, reported as a bar symbol –. Datasets include: electricity consumption (daily, weekly, electricity1-10), gas sensor measurements (gas), stock exchange returns (Istanbul), electricity transformer temperatures (ETTh1, ETTh2, ETTm1, ETTm2), and various synthetic datasets, see Section 4 for details.

depends on P aligning with a quasi-periodic, low-rank structure in the data. Our experiments in Section 4 show that this approach is effective on real-world datasets. In practice, time series are not exactly periodic, but they can be approximated as a sum of few periodic components plus some noise. This is the rationale behind Fourier analysis and wavelet analysis for time series. The low rank hypothesis can be seen as a nonnegative and adaptive generalization of Fourier analysis where the basis time series $\mathbf{H}$ are learned from data.

The relevance of such a hypothesis on real data cannot be proven beforehand. Our numerical study on real data shows that we improve results in prediction, better than standard methods in time series analysis: Seasonal AutoRegressive Integrated Moving Average with eXogenous variables model (SARIMAX), EXPonential moving average (EXP), Random Forest Regressor (RFR), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), BasisFormer (Attention-based Time Series Forecasting with Learnable and Interpretable Basis). It suggests that the low rank assumption is reasonable for the datasets studied in the paper.

1.2 Related Works

Our work intersects with several research areas, including the theory of Nonnegative Matrix Factorization (NMF), its application to time-series analysis, and methods for handling missing data.

NMF Uniqueness and Our Contribution The uniqueness of NMF decompositions is a cornerstone of its theoretical understanding. Foundational work by Thomas (1974) and subsequent analyses by Donoho & Stodden (2004); Laurberg et al. (2008) and Recht et al. (2012) have established conditions under which NMF yields a unique solution, often relying on geometric properties of the data matrix. Our work builds upon this literature but extends it to the *challenging setting of partially-observed data* with a specific block-wise missing structure. While previous results require the full data matrix, our theoretical guarantees (Section 2.2) are tailored to the matrix completion scenario inherent to our SMM framework, representing a key theoretical novelty.

Time-Series Forecasting Models The field of time-series forecasting is dominated by statistical and deep learning models. Classical methods like SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous variables) assume linear dependencies and specific seasonal patterns. In contrast, deep learning models such as LSTMs (Long Short-Term Memory networks), BasisFormer (Ni et al., 2023), Autoformer (Wu et al., 2021), iTransformer (Liu et al., 2024), PatchMLP (Tang & Zhang, 2025) and TimeMixer (Wang et al., 2024) learn complex, non-linear temporal dependencies from large amounts of data. These models often operate as "black boxes". Our SMM framework offers a different paradigm: it assumes that time-series segments can be represented as a convex combination of a few learned, interpretable basis vectors (archetypes). This low-rank hypothesis is fundamentally different from the auto-regressive or attention-based mechanisms of other models and provides inherent interpretability, as demonstrated in our experiments (Section 4).

NMF for Missing Data The problem of applying NMF to data with missing values is not new and many of these approaches are purely algorithmic. Our primary contribution is that the SMM framework itself—a structured method for converting a time-series forecasting problem into a matrix completion problem. Our theoretical analysis provides guarantees for this specific structure, which general-purpose NMF-for-missing-data algorithms do not offer. The present work assumes block-wise missing structures and provides uniqueness and robustness guarantees in this context, which is novel compared to prior works that often assume random missingness without specific structural patterns. General patterns of missing data are not covered by our analysis and remain an open research question. However, our algorithmic framework can be adapted to other missing data patterns, although without the same theoretical guarantees.

NMF-based time-series analysis Our work is distinct from previous NMF-based time-series analysis by Mei et al. (2017); Mei et al. (2018). While Mei et al. also use NMF, their work focuses on recovering high-resolution time series from temporal aggregates (disaggregation) and leveraging side information. For example, they might recover individual household consumption from a neighborhood's total consumption. Our SMM framework is fundamentally different. It operates by creating a matrix of sliding windows from the time series, thereby transforming the forecasting problem into one of finding a low-rank representation of these segments. The goal is to learn archetypal segment patterns for forecasting, not to disaggregate a signal.

Robustness of archetypal analysis have been studied in Javadi & Montanari (2020a) for simplicial polyhedral cone approximation of a dataset, denoted in data matrix form by $\mathbf{X} \in \mathbb{R}^{n \times p}$ in this paper. This paper extends this latter analysis to the case where some data entries might be missing and some data blocks are not observed (forecast, red values in Fig. 2).

Notation Denote by $\mathbf{A}^\top$ the transpose of matrix $\mathbf{A}$. Denote by $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^\top \mathbf{A})$ the Frobenius norm of $\mathbf{A}$. We use $\mathbb{R}_+^{n \times p}$ to denote $n \times p$ nonnegative matrices. It would be useful to consider the columns description $A_k \in \mathbb{R}^{n_1}$ of matrix $\mathbf{A} = [A_1 \cdots A_{n_2}]$ and the row decomposition $A^{(k)} \in \mathbb{R}^{n_2}$ of a matrix $\mathbf{A}$ using $\mathbf{A}^\top = [(A^{(1)})^\top \cdots (A^{(n_1)})^\top]$ for $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$. Notation $A_{i,j}$ indicates the elements of matrix $\mathbf{A}$; $[n]$ represents the set $\{1, 2, \ldots, n\}$; $\mathbf{1}_d$ is the all-ones vector of size d; and $\mathbb{1}_{\mathcal{A}}$ is the indicator function of $\mathcal{A}$, such that $\mathbb{1}_{\mathcal{A}} = 0$ if condition $\mathcal{A}$ is verified, ∞ otherwise. *Unless otherwise stated, all the proofs are given in Supplement Material.*

2 Uniqueness and estimation guarantees

2.1 The train and test paradigm, link with forecasting multiple nonnegative time series

The model under consideration is presented in Equations (3). Our goal is to estimate the K-best normalized non-negative approximation $\mathbf{X}_0$, defined in Equation (3a), from the partial and noisy observation $\mathbf{X}$. We denote by $\mathbf{X}^\star$ is the *mask* of $\mathbf{X}_0$, namely

$$\mathbf{X}^\star := \mathbf{T}(\mathbf{X}_0) = \left[\begin{array}{c|c} \mathbf{X}_1^\star & \mathbf{X}_2^\star \\ \hline \mathbf{X}_3^\star & \mathbf{0}_{N\times F} \end{array}\right] , \tag{5a}$$

where $\mathbf{X}_1^\star \in \mathbb{R}^{(n-N)\times(p-F)}$, $\mathbf{X}_2^\star \in \mathbb{R}^{(n-N)\times F}$, and $\mathbf{X}_3^\star \in \mathbb{R}^{N\times(p-F)}$ are blocks of $\mathbf{X}_0$. Note that $\mathbf{X} = \mathbf{X}^\star + \mathbf{F}$, where $\mathbf{F}$ is the *noise* term, see Equation (4).

These blocks can be be gathered into a *train and test paradigm*. Note that we observe the full sub-matrix $\mathbf{T}_{\text{train}}(\mathbf{X}_0) := [\mathbf{X}_1^\star\ \mathbf{X}_2^\star]$ which we refer to as the training part. We would like to predict (forecast) the $\mathbf{0}_{N\times F}$ block of the sub-matrix $\mathbf{T}_{\text{test}}(\mathbf{X}_0) := [\mathbf{X}_3^\star\ \mathbf{0}_{N\times F}]$ which we refer to as the test part of our observation $\mathbf{X}$. Looking at Figure 2, we define

$$\mathbf{T}_T(\mathbf{X}_0) := \left[\begin{array}{c} \mathbf{X}_1^\star \\ \mathbf{X}_3^\star \end{array}\right] \quad \text{and} \quad \mathbf{T}_F(\mathbf{X}_0) := \left[\begin{array}{c} \mathbf{X}_2^\star \\ \mathbf{0}_{N\times F} \end{array}\right] . \tag{5b}$$

Our notation (subscripts T and F) stems from the sliding mask method for multiple time series forecast. Note that $\mathbf{T}_T(\mathbf{X})$ gathers all the information observed up to time T, and we would like to forecast the $\mathbf{0}_{N\times F}$ block of $\mathbf{T}_F(\mathbf{X})$. Now, we know by design that $\mathbf{X}_0 := \mathbf{W}_0\mathbf{H}_0$. Hence, denoting $\mathbf{H}_0 =: [\mathbf{H}_{0T}\ \mathbf{H}_{0F}]$, and $\mathbf{W}_0^\top =: [\mathbf{W}_{0_{\text{train}}}^\top\ \mathbf{W}_{0_{\text{test}}}^\top]$, we get that

$$\begin{aligned} \mathbf{T}_{\text{train}}(\mathbf{X}_0) &= \mathbf{W}_{0_{\text{train}}}\mathbf{H}_0 , & \mathbf{X}_3^\star &= \mathbf{W}_{0_{\text{test}}}\mathbf{H}_{0T} , \\ \mathbf{T}_T(\mathbf{X}_0) &= \mathbf{W}_0\mathbf{H}_{0T} , & \mathbf{X}_2^\star &= \mathbf{W}_{0_{\text{train}}}\mathbf{H}_{0F} . \end{aligned} \tag{5c}$$

In light of Figures 1 and 2, the multiple forecasts $\hat{\mathbf{M}}_{T+1}, \ldots, \hat{\mathbf{M}}_{T+F}$ can be given a best normalized nonnegative rank K approximation by $\mathbf{W}_{0_{\text{test}}}\mathbf{H}_{0F}$. Observe that an estimation of $\mathbf{W}_{0_{\text{test}}}$ gives the weights learnt on the test sub-matrix while an estimation of $\mathbf{H}_{0F}$ is the forecast of the archetypes, see the decomposition (2).

2.2 Uniqueness from partial observations

When we observe the full matrix $\mathbf{X}_0 = \mathbf{W}_0\mathbf{H}_0$, the issue on uniqueness has been addressed under some sufficient conditions on $\mathbf{W}, \mathbf{H}$, *e.g.*, *Strongly boundary closeness* of Laurberg et al. (2008), *Complete factorial sampling* of Donoho & Stodden (2004), and *Separability* of Recht et al. (2012). A necessary and sufficient condition exists as given by the following theorem. We recall that that the K-dimensional positive orthant is the set $\{x \in \mathbb{R}^K : x_i \geq 0\,,\ \forall i \in [K]\}$ and a K-simplicial cone is the set which the conic hull of K linearly independent vectors of $\mathbb{R}^K$.

Theorem 8 (Thomas (1974)) *The decomposition $\mathbf{X}_0 := \mathbf{W}_0\mathbf{H}_0$ is unique up to permutation and positive scaling of columns (resp. rows) of $\mathbf{W}_0$ (resp. $\mathbf{H}_0$)* **if and only if** *the K-dimensional positive orthant is the only K-simplicial cone verifying $\text{Cone}(\mathbf{W}_0^\top) \subseteq \mathcal{C} \subseteq \text{Cone}(\mathbf{H}_0)$ where $\text{Cone}(\mathbf{A})$ is the cone generated by the rows of $\mathbf{A}$.*

Our first assumption is following.

Assumption 1 *In the set given by the union of sets:*

$$\{\mathcal{C} \,:\, \text{Cone}({\mathbf{W}_{0_{\text{train}}}}^\top) \subseteq \mathcal{C} \subseteq \text{Cone}(\mathbf{H}_0)\} \bigcup \{\mathcal{C} \,:\, \text{Cone}(\mathbf{W}_0^\top) \subseteq \mathcal{C} \subseteq \text{Cone}(\mathbf{H}_{0T})\} , \tag{$\mathbb{A}_1$}$$

the nonnegative orthant is the only K-simplicial cone. Note that this assumption is implied by the following one: In the set

$$\{\mathcal{C} \,:\, \text{Cone}({\mathbf{W}_{0_{\text{train}}}}^\top) \subseteq \mathcal{C} \subseteq \text{Cone}(\mathbf{H}_{0T})\} \tag{$\mathbb{A}_1'$}$$

the nonnegative orthant is the only K-simplicial cone.

Remark 9 *This assumption adapts the necessary and sufficient condition for NMF uniqueness from Thomas (1974) to our partial observation setting. The standard condition requires the positive orthant to be the only simplicial cone $\mathcal{C}$ such that* $\mathrm{Cone}(\mathbf{W}_0^\top) \subseteq \mathcal{C} \subseteq \mathrm{Cone}(\mathbf{H}_0)$. *In our case, since we only observe parts of the data matrix, we need to ensure uniqueness based on partial information about the factors* $\mathbf{W}_0$ *and* $\mathbf{H}_0$. *The union of sets in* ($\mathbb{A}_1$) *ensures that we can uniquely identify the factors from the observed training data (*$\mathbf{T}_{\mathrm{train}}(\mathbf{X}_0)$*) and the observed past data (*$\mathbf{T}_T(\mathbf{X}_0)$*), which is a stronger requirement to guarantee identifiability in the matrix completion framework.*

Remark 10 *Assumption 1 imposes implicit constraints on the dimensions of the problem and the nonnegative rank K. For the condition to be non-trivial, the number of rows in the matrices generating the cones must be sufficient. Specifically, for* $\mathrm{Cone}({\mathbf{W}_{0\mathrm{train}}}^\top)$, *the number of training samples,* $n-N$ must be at least K. *Similarly, for* $\mathrm{Cone}(\mathbf{H}_{0T})$, *the number of observed time steps,* $p-F$ must be at least K. *These conditions ensure that the cones are full-dimensional in* $\mathbb{R}^K$, *which is a prerequisite for the uniqueness argument to hold.*

We consider the following standard definition.

Definition 11 (Javadi & Montanari (2020a)) *For a matrix* $\mathbf{A} \in \mathbb{R}^{n' \times p'}$, *let* $\mathrm{conv}(\mathbf{A})$ *denote the convex hull of its rows. The internal radius of* $\mathrm{conv}(\mathbf{A})$, *denoted* $\mu(\mathbf{A})$, *is the radius of the largest* $(K-1)$*-dimensional ball contained within the affine hull of* $\mathrm{conv}(\mathbf{A})$. *We say that* $\mathrm{conv}(\mathbf{A})$ *has an internal radius* μ *if* $\mu(\mathbf{A}) = \mu$.

Our second main assumption is the following.

Assumption 2 *Assume that*

$$\mathrm{conv}(\underbrace{\mathbf{T}_T(\mathbf{X}_0)}_{=\mathbf{W}_0\mathbf{H}_{0T}}) \text{ and } \mathrm{conv}(\underbrace{\mathbf{T}_{\mathrm{train}}(\mathbf{X}_0)}_{=\mathbf{W}_{0\mathrm{train}}\mathbf{H}_0}) \text{ have internal radius at least } \mu > 0\,. \tag{$\mathbb{A}_2$}$$

Remark 12 *Assumption* ($\mathbb{A}_2$) *implies that the convex hulls of the data points,* $\mathrm{conv}(\mathbf{W}_0\mathbf{H}_{0T})$ *and* $\mathrm{conv}(\mathbf{W}_{0\mathrm{train}}\mathbf{H}_0)$, *are not flat, meaning they are full-dimensional within the affine subspace they span. We uncover the same constraints as in the previous remark:* $n-N$ must be at least K *(imposed by the dimension of the rows of* $\mathbf{T}_T(\mathbf{X}_0)$*) and* $p-F$ must be at least K *(imposed by the number of the rows of* $\mathbf{T}_{\mathrm{train}}(\mathbf{X}_0)$*).*

Theorem 13 *Assumption 1 implies Property* ($\mathbb{P}_{\mathrm{u}}$). *Moreover, if* ($\mathbb{A}_1$) *and* ($\mathbb{A}_2$) *holds,* $\mathbf{T}(\mathbf{W}\mathbf{H}) = \mathbf{T}(\mathbf{W}_0\mathbf{H}_0)$ *and* $\mathbf{W}_0\mathbf{1} = \mathbf{W}\mathbf{1} = \mathbf{1}$ *then* $(\mathbf{W}, \mathbf{H}) = (\mathbf{W}_0, \mathbf{H}_0)$ *up to permutation of columns (resp. rows) of* $\mathbf{W}$ *(resp.* $\mathbf{H}$*), and there is no scaling.*

Corollary 14 *If decomposition of* $\mathbf{X}_1 = \mathbf{W}_{0\mathrm{train}}\mathbf{H}_{0T}$ *is unique then* ($\mathbb{P}_{\mathrm{u}}$) *holds.*

Proof. By Theorem 8, observe that ($\mathbb{A}_1'$) is a necessary and sufficient condition for the uniqueness of the decomposition $\mathbf{X}_1 = \mathbf{W}_{0\mathrm{train}}\mathbf{H}_{0T}$. Observe that ($\mathbb{A}_1'$) implies ($\mathbb{A}_1$) and invoke Theorem 8 and Theorem 13. ■

This corollary shows that the uniqueness of the decomposition of the top left block $\mathbf{X}_1$ (which is fully observed) implies the uniqueness of normalized decomposition of $\mathbf{X}_0$ given partial observations (the bottom right block is not observed).

Limitations and Discussion

It is important to note that the uniqueness theory presented here relies on a specific structure of missing data, namely the block-wise missing pattern corresponding to the matrix completion problem for recommender systems. Our analysis leverages the fact that certain submatrices are fully observed.

The extension of these uniqueness guarantees to scenarios with arbitrary or unstructured missing data patterns is a non-trivial challenge. Such cases would require different theoretical tools, as the problem can no longer be reduced to the uniqueness of fully-observed sub-decompositions. This constitutes an important direction for future research.

2.3 Robustness under partial observations

The second issue is *robustness to noise*. To the best of our knowledge, all the results addressing this issue assume that the noise error term is small enough, *e.g.*, Laurberg et al. (2008), Recht et al. (2012), or Javadi & Montanari (2020a).

In this paper, we extend these stability result to the nonnegative matrix completion framework (partial observations) and we also assume that noise term $\|\mathbf{F}\|_F$ is small enough.

In the normalized case (*i.e.*, $\mathbf{W}\mathbf{1} = \mathbf{1}$), both issues (uniqueness and robustness) can be handled with the notion of α-uniqueness, introduced by Javadi & Montanari (2020a). This notion does not handle the matrix completion problem we are addressing. To this end, let us introduce the following notation. Given two matrices $\mathbf{A} \in \mathbb{R}^{n_a \times p}$ and $\mathbf{B} \in \mathbb{R}^{n_b \times p}$ with same row dimension, and $\mathbf{C} \in \mathbb{R}^{n_a \times n_b}$, define the divergence $\mathcal{D}(\mathbf{A}, \mathbf{B})$ as

$$\begin{aligned}\mathcal{D}(\mathbf{A}, \mathbf{B}) &:= \min_{\mathcal{C} \geq \mathbf{0}\,,\, \mathcal{C}\mathbf{1}_{n_b} = \mathbf{1}_{n_a}} \sum_{a=1}^{n_a} \left\| A^{(a)} - \sum_{b=1}^{n_b} C_{ab} B^{(b)} \right\|_F^2 , \\ &= \min_{\mathcal{C} \geq \mathbf{0}\,,\, \mathcal{C}\mathbf{1}_{n_b} = \mathbf{1}_{n_a}} \|\mathbf{A} - \mathbf{C}\mathbf{B}\|_F^2 . \end{aligned} \tag{6a}$$

which is the squared distance between rows of $\mathbf{A}$ and $\mathrm{conv}(\mathbf{B})$, the convex hull of rows of $\mathbf{B}$. For $\mathbf{B} \in \mathbb{R}^{n \times p}$ define

$$\widetilde{\mathcal{D}}(\mathbf{A}, \mathbf{B}) := \min_{\substack{\mathbf{C} \geq \mathbf{0}\,,\, \mathbf{C}\mathbf{1}_n = \mathbf{1}_{n_a} \\ \mathbf{T}(\mathbf{N} - \mathbf{B}) = 0}} \|\mathbf{A} - \mathbf{C}\mathbf{N}\|_F^2 . \tag{6b}$$

Definition 15 ($\mathbf{T}_\alpha$-unique, Javadi & Montanari (2020a)) *Given $\mathbf{X}_0 \in \mathbb{R}^{n \times p}$, $\mathbf{W}_0 \in \mathbb{R}^{n \times K}$, and $\mathbf{H}_0 \in \mathbb{R}^{K \times p}$, the factorization $\mathbf{X}_0 = \mathbf{W}_0 \mathbf{H}_0$ is $\mathbf{T}_\alpha$-unique with parameter $\alpha > 0$ if for all $\mathbf{H} \in \mathbb{R}^{K \times p}$ with $\mathrm{conv}(\mathbf{X}_0) \subseteq \mathrm{conv}(\mathbf{H})$:*

$$\widetilde{\mathcal{D}}(\mathbf{H}, \mathbf{X}_0)^{1/2} \geq \widetilde{\mathcal{D}}(\mathbf{H}_0, \mathbf{X}_0)^{1/2} + \alpha \left\{ \mathcal{D}(\mathbf{H}, \mathbf{H}_0)^{1/2} + \mathcal{D}(\mathbf{H}_0, \mathbf{H})^{1/2} \right\} . \tag{6c}$$

Our third main assumption is given by:

Assumption 3 *Assume that*

$$\mathbf{X}_0 = \mathbf{W}_0 \mathbf{H}_0 \text{ is } \mathbf{T}_\alpha\text{-unique} \tag{$\mathbb{A}_3$}$$

Theorem 16 (Archetypes estimation) *If ($\mathbb{A}_2$) and ($\mathbb{A}_3$) hold then there exist positive reals Δ and Λ (depending on $\mathbf{X}_0$) such that, for all $\mathbf{F}$ such that $\|\mathbf{F}\|_F \leq \Delta$ and $0 \leq \lambda \leq \Lambda$, any solution $(\widehat{\mathbf{W}}, \widehat{\mathbf{H}})$ to (mAMF) (if $\lambda \neq 0$) or (mNMF) (if $\lambda = 0$) with observation (4) is such that:*

$$\sum_{\ell \leq K} \min_{\ell' \leq K} \|\mathbf{H}_{0\ell} - \widehat{\mathbf{H}}_{\ell'}\|_2^2 \leq c \, \|\mathbf{F}\|_F^2 ,$$

where $c > 0$ is a constant depending only on $\mathbf{W}_0$ and $\mathbf{H}_0$

By Theorem 16, when the noise is sufficiently small, there exists a permutation σ on $[K]$ such that

$$\|\mathbf{H}_0 - \hat{\mathbf{H}}_\sigma\|_F^2 := \sum_{\ell \leq K} \|\mathbf{H}_{0\ell} - \hat{\mathbf{H}}_{\sigma(\ell)}\|_2^2 \leq c \, \|\mathbf{F}\|_F^2 \tag{6d}$$

where $\hat{\mathbf{H}}_\sigma$ is a permutation of the row of $\hat{\mathbf{H}}$.

Corollary 17 (Estimation Error Bound for $\mathbf{W}$) *Under the assumptions of Theorem 16 with $\mu > 0$ the internal radius of Assumption 2. Let $\hat{\mathbf{H}}$ be the estimator satisfying $\|\hat{\mathbf{H}} - \mathbf{H}_0\|_F \leq c\|\mathbf{F}\|_F$ as in (6d), the error of the weight matrix $\hat{\mathbf{W}}$ is bounded by:*

$$\|\hat{\mathbf{W}} - \mathbf{W}_0\|_F \leq c' \|\mathbf{F}\|_F \tag{6e}$$

where $c' > 0$ is a constant depending only on $\mathbf{W}_0$ and $\mathbf{H}_0$.

The proof of this corollary can be found in Appendix B.3.

3 Solving masked nonnegative/archetypal matrix factorization

3.1 Alternating Least Squares for (mNMF)

The basic algorithmic framework for matrix factorization problems is *Block Coordinate Descent* (BCD) method, which can be straightforwardly adapted to (mNMF) (see Supplement Material). BCD for (mNMF) reduces to *Alternating Least Squares* (ALS) algorithm (see Algorithm 4 in Appendix), when an alternative minimization procedure is performed and matrix $\mathbf{WH}$ is projected onto the linear subspace $\mathbf{T}(\mathbf{N}) = \mathbf{X}$ by means of operator $\mathcal{P}_{\mathbf{X}}$, as follows:

$$\mathbf{N} := \mathcal{P}_{\mathbf{X}}(\mathbf{WH}) : \mathbf{T}(\mathbf{N}) = \mathbf{X} \text{ and } \mathbf{T}^{\perp}(\mathbf{N}) = \mathbf{WH}\,.$$

Hierarchical Alternating Least Squares (HALS) is an ALS-like algorithm obtained by applying an exact coordinate descent method Gillis (2014). Moreover, an accelerated version of HALS is proposed in Gillis & Glineur (2012) (see Supplement Material).

3.2 Projected Gradient for (mAMF)

Proximal Alternating Linearized Minimization (PALM) method, introduced in Bolte et al. (2014) and applied to AMF by Javadi & Montanari (2020a), can be also generalized to (mAMF) (see Algorithm 1). In the following, $\mathcal{P}_{\mathrm{conv}(\mathbf{A})}$ is the projection operator onto $\mathrm{conv}(\mathbf{A})$ and $\mathcal{P}_{\Delta}$ is the projection operator onto the $(N-1)$-dimensional standard simplex Δ^N. The two projections can be efficiently computed by means of, *e.g.*, Wolfe algorithm Wolfe (1976) and active set method Condat (2016) respectively.

Algorithm 1 PALM for mAMF

1: **Initialization**: chose $\mathbf{H}^0, \mathbf{W}^0 \geq \mathbf{0}$ such that $\mathbf{W}^0\mathbf{1} = \mathbf{1}$, set $\mathbf{N}^0 := \mathcal{P}_{\mathbf{X}}(\mathbf{W}^0\mathbf{H}^0)$ and $i := 0$.
2: **while** stopping criterion is not satisfied **do**
3: $\quad \widetilde{\mathbf{H}}^i := \mathbf{H}^i - \frac{1}{\gamma_1^i}\mathbf{W}^{i\top}\left(\mathbf{W}^i\mathbf{H}^i - \mathbf{N}^i\right)$ ▷ Gradient step on $\mathbf{H}$, objective first term
4: $\quad \mathbf{V}^{i+1}$ such that $\mathcal{P}_{\mathrm{conv}(\mathbf{N}^i)}(\widetilde{\mathbf{H}}^i) = \mathbf{V}^{i+1}\mathbf{N}^i$ ▷ Projection of $\widetilde{\mathbf{H}}^i$ onto $\mathrm{conv}(\mathbf{N}^i)$ by Wolfe algorithm
5: $\quad \mathbf{H}^{i+1} := \widetilde{\mathbf{H}}^i - \frac{\lambda}{\lambda+\gamma_1^i}\left(\widetilde{\mathbf{H}}^i - \mathcal{P}_{\mathrm{conv}(\mathbf{N}^i)}(\widetilde{\mathbf{H}}^i)\right)$ ▷ Gradient step on $\mathbf{H}$, objective second term
6: $\quad \mathbf{W}^{i+1} := \mathcal{P}_{\Delta}\left(\mathbf{W}^i - \frac{1}{\gamma_2^i}\left(\mathbf{W}^i\mathbf{H}^{i+1} - \mathbf{N}^i\right)\mathbf{H}^{i+1\top}\right)$ ▷ Projected gradient step on $\mathbf{W}$
7: $\quad \mathbf{N}^{i+1} := \mathcal{P}_{\mathbf{X}}\left(\mathbf{N}^i + \frac{1}{\gamma_3^i}\left(\mathbf{W}^{i+1}\mathbf{H}^{i+1} - \mathbf{N}^i\right)\right)$ ▷ Projected gradient step on $\mathbf{N}$
8: $\quad i := i + 1$
9: **end while**

Remark 18 *Strictly speaking, Algorithm 1 is an approximation of the exact PALM framework as described in Bolte et al. (2014) because the so-called regularization term contains a coupling term $\lambda\mathcal{D}(\mathbf{H},\mathbf{N})$ between blocks $\mathbf{H}$ and $\mathbf{N}$ that is not part of the so-called mooth function. Algorithm 1 addresses this by splitting the update of $\mathbf{H}$ into a gradient step on the smooth function (Step 3) followed by a correction step using the archetypal projection (Step 5), and updating $\mathbf{N}$ via a projected gradient step on the smooth function (Step 7). We analyze the convergence properties assuming this scheme approximates the minimization with respect $\mathbf{N}$ at each step. We could have made a loop over $\mathbf{N}$ to ensure convergence, at the price of computing time. Further details are given in Appendix B.4.*

Theorem 19 *Let $\varepsilon > 0$. Let $L_H(\mathbf{W}) = \|\mathbf{W}^\top\mathbf{W}\|_F$, $L_W(\mathbf{H}) = \|\mathbf{H}\mathbf{H}^\top\|_F$, and $L_N = 1$. If the step sizes satisfy $\gamma_1^i > L_H(\mathbf{W}^i)$, $\gamma_2^i > \max\{L_W(\mathbf{H}^{i+1}), \varepsilon\}$, and $\gamma_3^i > 1$, then Algorithm 1, viewed as an approximation of a Proximal Alternating Linearized Minimization (PALM) scheme, generates a sequence $(\mathbf{H}^i, \mathbf{W}^i, \mathbf{N}^i)$ that converges to a stationary point of the objective function of* (mAMF).

Proof. Proof is given in Supplement Material. ■

Remark 20 *Note that* (mAMF) *objective (with the $\lambda\|\mathbf{H} - \mathbf{VN}\|_F^2$ or $\lambda\mathcal{D}(\mathbf{H},\mathbf{N})$ term) is not a standard Nonnegative Least-Squares problem, making HALS inapplicable. In the following algorithms $\mathbf{H}, \mathbf{W}$ updates (Steps 5 & 6) are proximal-gradient steps. Further details are given in Appendix B.4.*

Finally, the inertial PALM (iPALM) method, introduced for NMF in Pock & Sabach (2016), is generalized to (mAMF) in Algorithm 2.

Algorithm 2 iPALM for mAMF

Initialization: $\mathbf{H}^0$, $\mathbf{W}^0 \geq 0$ such that $\mathbf{W}^0\mathbf{1} = \mathbf{1}$, set $\mathbf{N}^0 := \mathcal{P}_{\mathbf{X}}(\mathbf{W}^0\mathbf{H}^0)$, $\mathbf{H}^{-1} := \mathbf{H}^0$, $\mathbf{W}^{-1} := \mathbf{W}^0$, $\mathbf{N}^{-1} := \mathbf{N}^0$, and $i := 0$.
while stopping criterion is not satisfied **do**
  $\mathbf{H}_1^i := \mathbf{H}^i + \alpha_1^i\left(\mathbf{H}^i - \mathbf{H}^{i-1}\right)$, $\mathbf{H}_2^i := \mathbf{H}^i + \beta_1^i\left(\mathbf{H}^i - \mathbf{H}^{i-1}\right)$ ▷ Inertial $\mathbf{H}$
  $\widetilde{\mathbf{H}}^i := \mathbf{H}_1^i - \frac{1}{\gamma_1^i}\mathbf{W}^{i\top}\left(\mathbf{H}_2^i\mathbf{W}^i - \mathbf{N}^i\right)$ ▷ Gradient step on $\mathbf{H}$, objective first term
  $\mathbf{V}^{i+1}$ such that $\mathcal{P}_{\mathrm{conv}(\mathbf{N}^i)}(\tilde{\mathbf{H}}^i) = \mathbf{V}^{i+1}\mathbf{N}^i$ ▷ Projection of $\tilde{\mathbf{H}}^i$ onto $\mathrm{conv}(\mathbf{N}^i)$ by Wolfe algorithm
  $\mathbf{H}^{i+1} := \widetilde{\mathbf{H}}^i - \frac{\lambda}{\lambda+\gamma_1^i}\left(\widetilde{\mathbf{H}}^i - \mathcal{P}_{\mathrm{conv}(\mathbf{N}^i)}(\tilde{\mathbf{H}}^i)\right)$ ▷ Gradient step on $\mathbf{H}$, objective second term
  $\mathbf{W}_1^i := \mathbf{W}^i + \alpha_2^i\left(\mathbf{W}^i - \mathbf{W}^{i-1}\right)$, $\mathbf{W}_2^i := \mathbf{W}_1^i + \beta_2^i\left(\mathbf{W}^i - \mathbf{W}^{i-1}\right)$ ▷ Inertial $\mathbf{W}$
  $\mathbf{W}^{i+1} := \mathcal{P}_{\Delta}\left(\mathbf{W}_1^i - \frac{1}{\gamma_2^i}\left(\mathbf{W}_2^i\mathbf{H}^{i+1} - N^i\right)\mathbf{H}^{i+1\top}\right)$ ▷ Projected gradient step on $\mathbf{W}$
  $\mathbf{N}_1^i := \mathbf{N}_1^i + \alpha_3^i\left(\mathbf{N}^i - \mathbf{N}^{i-1}\right)$, $\mathbf{N}_2^i := \mathbf{N}_1^i + \beta_3^i\left(\mathbf{N}^i - \mathbf{N}^{i-1}\right)$ ▷ Inertial $\mathbf{N}$
  $\mathbf{N}^{i+1} := \mathcal{P}_{\mathbf{X}}\left(\mathbf{N}_1^i + \frac{1}{\gamma_3^i}\left(\mathbf{W}^{i+1}\mathbf{H}^{i+1} - \mathbf{N}_2^i\right)\right)$ ▷ Projected gradient step on $\mathbf{N}$
  $i := i + 1$
end while

Remark 21 *If, for all iterations i, $\alpha_1^i = \alpha_2^i = 0$ and $\beta_1^i = \beta_2^i = 0$, iPALM reduces to PALM.*

Stopping criterion For (mNMF), KKT conditions regarding matrix $\mathbf{W}$ are the following (see Supplement Material):

$$\mathbf{W} \circ \left((\mathbf{W}\mathbf{H} - \mathbf{N})\mathbf{H}^\top + t\,\mathbf{1}_K^\top\right) = 0\,.$$

By complementary condition, it follows that, $\forall j$, $t_i = ((\mathbf{W}\mathbf{H} - \mathbf{N})\mathbf{H}^\top)_{i,j}$. Hence, we compute t_i by selecting, for each row $W^{(i)}$, any positive entry $W_{i,j} > 0$.

Remark 22 *Numerically to obtain a robust estimate of t_i, we can average the corresponding values calculated per entry $W_{i,j}$.*

Let $\varepsilon_{\mathbf{W}}$, $\varepsilon_{\mathbf{H}}$, and $\varepsilon_{\mathbf{R}}$ be three positive thresholds. The stopping criterion for the previous algorithms consists of a combination of:

1. the maximum number of iterations;
2. the Frobenius norm of the difference of $\mathbf{W}$ and $\mathbf{H}$ at two consecutive iterations, *i.e.*, the algorithm stops if $\|\mathbf{W}^{i+1} - \mathbf{W}^i\|_F \leq \varepsilon_{\mathbf{W}} \;\wedge\; \|\mathbf{H}^{i+1} - \mathbf{H}^i\|_F \leq \varepsilon_{\mathbf{H}}$;
3. a novel criterion based on KKT condition, *i.e.*, the algorithm stops if it holds that

$$\|\mathbf{R}(\mathbf{W}^{i+1})\|_F + \|\mathbf{R}(\mathbf{H}^{i+1})\|_F \leq \varepsilon_{\mathbf{R}}\,,$$

   where matrices $\mathbf{R}(\mathbf{W})$ and $\mathbf{R}(\mathbf{H})$ are defined as

$$\mathbf{R}(\mathbf{W})_{i,j} := |(\mathbf{W}\mathbf{H} - \mathbf{N})\mathbf{H}^\top)_{i,j} + t_i|\mathbb{1}_{\{W_{i,j}\neq 0\}}$$
$$\text{and } \mathbf{R}(\mathbf{H})_{i,j} := |\mathbf{W}^\top(\mathbf{W}\mathbf{H} - \mathbf{N}))_{i,j}|\mathbb{1}_{\{H_{i,j}\neq 0\}}$$

   respectively.

3.3 Large-scale dataset

Assume the observed matrix $\mathbf{X} = \mathbf{T}(\mathbf{\Phi}(\mathbf{M}))$ is large-scaled, namely one has to forecast a large number N of times series (*e.g.* more than $100,000$) and possibly a large number of time stamps T. The strategy, described in Section 1.3.1

in Cichocki et al. (2009) for NMF, is to learn the $\mathbf{H} \in \mathbb{R}^{K\times T}$ matrix from a submatrix $\mathbf{N}_r \in \mathbb{R}^{r\times T}$ of $K \leq r \ll N$ rows of $\mathbf{N} \in \mathbb{R}^{n\times T}$, and to learn the $\mathbf{W} \in \mathbb{R}^{N\times K}$ matrix from a sub-matrix $\mathbf{N}_c \in \mathbb{R}^{N\times c}$ of $K \leq c \ll T$ columns of $\mathbf{N} \in \mathbb{R}^{N\times T}$. We denote by $\mathbf{H}_c$ the submatrix of $\mathbf{H}$ given by the columns appearing in $\mathbf{N}_c$ and $\mathbf{W}_r$ the sub-matrix of $\mathbf{H}$ given by the columns appearing in $\mathbf{N}_c$.

This strategy can be generalized to (mNMF) and (mAMF). For (mNMF) this generalization is straightforward, and for (mAMF) one need to change Steps 3-5 in Algorithm 1 as follows:

$$\begin{aligned}
\widetilde{\mathbf{H}}^i &:= \mathbf{H}^i - \frac{1}{\gamma_1^i}(\mathbf{W}_r^i)^\top \left(\mathbf{W}_r^i\mathbf{H}^i - \mathbf{N}_r^i\right) \\
\mathbf{H}^{i+1} &:= \widetilde{\mathbf{H}}^i - \frac{\lambda}{\lambda+\gamma_1^i}\left(\widetilde{\mathbf{H}}^i - \mathcal{P}_{\mathrm{conv}(\mathbf{N}^i)}(\tilde{\mathbf{H}}^i)\right) \\
\mathbf{W}^{i+1} &:= \mathcal{P}_\Delta\left(\mathbf{W}^i - \frac{1}{\gamma_2^i}\left(\mathbf{W}^i\mathbf{H}_c^{i+1} - \mathbf{N}_c^i\right)(\mathbf{H}_c^{i+1})^\top\right).
\end{aligned}$$

The same approach is used for Algorithm 2.

4 Numerical Experiments

We tested SMM on real-world datasets. Matrix $\mathbf{H}^0$ is initially selected as in Javadi & Montanari (2020a). Each row of matrix $\mathbf{W}^0$ is generated randomly in the corresponding standard simplex. For SMM we implemented both HALS for (mNMF) and iPALM for (mAMF).

Moreover, we have compared our method with other classically-designed mainstream time series forecasting methods such as *Random Forest Regression* (RFR) and *EXPonential smoothing* (EXP), *Long Short-Term Memory* (LSTM) and *Gated Recurrent Units* (GRU) deep neural networks with preliminary data standardization Shewalkar et al. (2019), and *Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors* (SARIMAX) models Douc et al. (2014).

The interested reader may find a Github repository on numerical experiments at **[link redacted to comply with double blind reviewing]**

4.1 Real-world datasets

The numerical experiments refer to the following real-world datasets: weekly and daily electricity consumption datasets of 370 Portuguese customers during the period 2011-2014, Trindade (2015); twin gas measurement dataset of five replicates of an 8-MOX gas sensor, Fonollosa (2016); Istanbul Stock Exchange returns with seven other international indexes for the period 2009-2011, Akbilgic (2013); daily electricity transformer temperature (ETT) measurements, Zhou et al. (2020). Figure 3 reports the cross-validated RRMSE and RMPE on observed values obtained during computational tests for each method.

In the majority of the cases, our method is the best or second best among all the approaches for all the dataset we tested in terms of RRMSE and RMPE indices (except for the "weekly electricity" dataset), and there is no other method performing better.

(mAMF) seems to be the most promising algorithm in terms of performances for the first five datasets, while (mNMF) is the best method for the last four ETT datasets.

4.2 Comparison with SOTA method

We performed additional computational experiments to compare our NMF-based methodology with state-of-the-art time series transformer models, which are suitable for large-scale time series forecasting problems. In particular, we consider the BasisFormer model recently described in Ni et al. (2023). We consider the same electricity dataset as in Ni et al. (2023) and split the whole dataset into 10 small sets of 960 time steps each. We collect our performance statistics, namely RRMSE and RMPE, on the original unscaled datasets. Note that in Ni et al. (2023), the performance statistics reported are the absolute errors on the scaled dataset obtained

by applying `sklearn.preprocessing.StandardScaler` to the original data. For BasisFormer, we run the code in the repository https://github.com/nzl5116190/Basisformer. Moreover we compare also against previous state-of-the-art methods, such as Autoformer as in Wu et al. (2021), iTransformer as in Liu et al. (2024), PatchMLP as in Tang & Zhang (2025) and TimeMixer as in Wang et al. (2024), for which we run the code in repositories https://github.com/thuml/Autoformer, https://github.com/thuml/iTransformer, https://github.com/TangPeiwang/PatchMLP, and https://github.com/kwuking/TimeMixer, respectiverly.

As shown in Figure 3, our method outperforms the SOTA methodology and is competitive against the other methodologies (in particular, with respect to the deep learning approaches which seem the most promising methods for these datasets). We also perform additional computational experiments on scaled datasets, collecting our performance indices on relative errors and absolute errors as in Ni et al. (2023), and we obtain the same dominance results.

4.3 Synthetic datasets

Further computational experiments have been performed by considering additional synthetic datasets. In particular, we generated three datasets by replicating $1,000$ short time series (with 10 time periods) 10 times and adding white noise multiplied by a constant factor σ to each time series entry separately. We choose $\sigma \in \{0.005, 0.1, 1\}$. We refer to the these datasets as "low noise", "medium noise", and "high noise", respectively.

An additional synthetic dataset has been generated considering few probability vectors and computing the entire matrix $\mathbf{W}$ by randomly choosing a probability vector and adding white noise. A completely randomly generated matrix $\mathbf{H}$ is multiplied by $\mathbf{W}$ to obtain the whole matrix $\mathbf{M}^* := \mathbf{WH}$. We refer to this dataset as "synthetic1".

Finally, the last synthetic dataset is obtained by generating a matrix $\mathbf{H}$ by replicating a small time series (with 50 time periods) 100 times and adding white noise multiplied by a constant factor $\sigma = 1$ and matrix $\mathbf{W}$ of suitable dimensions, whose rows are uniformly distributed over the corresponding dimensional simplex. Then, we set the matrix $\mathbf{M}^* := \mathbf{WH}$. We refer to this last dataset as "synthetic2".

Figure 3 reports the cross-validated RRMSE and RMPE indices referring to synthetically generated datasets. The more pronounced the periodicity of the time series or of the archetypes, the better the performances of our proposed NMF-like methods: in this case, the more realistic the hypothesis that the whole dataset can be expressed as convex combinations of a few archetypes, having a low-rank representation.

Broader Impact Statement

The development of interpretable machine learning models for time-series forecasting has significant potential for positive societal impact. Our work, the Sliding Mask Method (SMM), contributes to this area by providing a framework that is not only effective for prediction but also offers insights into the underlying patterns driving the data.

One of the primary application areas for this research is in resource management and planning. For example, utility companies can use our method to forecast electricity demand. Accurate forecasts help in optimizing power generation, reducing waste, and preventing outages, which benefits both the economy and the environment. The interpretability of SMM is particularly valuable here, as it can help identify and understand different consumption patterns (e.g., residential vs. industrial archetypes), allowing for more targeted demand-response strategies. Similarly, in retail, forecasting sales data can lead to more efficient inventory management, reducing waste from overstocking and economic losses from understocking.

While the potential impacts are largely positive, it is important to consider potential negative consequences. Like any forecasting tool, models can be misused if their limitations are not understood. For instance, relying on a model that assumes quasi-periodicity for a dataset with a strong, unexpected trend could lead to poor decisions. We have attempted to be transparent about these limitations in our discussion. Furthermore, the application of forecasting in financial markets, while a potential use case, carries inherent risks. Our model is not designed for high-frequency trading and should not be used as such without a thorough understanding of its assumptions.

5 Discussion and Conclusion

In this paper, we introduced the Sliding Mask Method (SMM), a framework that leverages Nonnegative Matrix Factorization for time-series forecasting. Our theoretical analysis provides uniqueness guarantees for the underlying decomposition in a structured matrix completion setting, and our experiments demonstrate its practical effectiveness. This concluding section synthesizes our findings to answer a crucial question: *When should a practitioner choose SMM?*

Based on our analysis and experimental results, our method is particularly well-suited for datasets with the following characteristics:

- **Non-negativity:** The time-series values must be non-negative, as this is a fundamental constraint of the NMF model.
- **Quasi-periodicity and Low-Rank Structure:** The method performs best when the time series exhibits quasi-periodic patterns. The core assumption of SMM is that segments of the time series can be effectively approximated by a low-rank model, i.e., as combinations of a few archetypal patterns. Datasets like electricity consumption and sales data, which often have daily, weekly, or seasonal cycles, are prime candidates.
- **Interpretability is Valued:** A key advantage of SMM is the interpretability of its results. The learned basis vectors ($\mathbf{H}$) represent archetypal time-series segments, and the weights ($\mathbf{W}$) show how each individual segment is composed of these archetypes. This provides insights into the underlying data-generating process that "black-box" models like LSTMs or Transformers cannot offer.

Conversely, our method may not be the optimal choice in other scenarios. For instance, as suggested by our experiments on synthetic data, SMM is less effective for time series dominated by strong, non-periodic linear trends. In such cases, models explicitly designed to handle trends, such as SARIMAX or other regression-based techniques, may be more appropriate.

In summary, SMM provides a powerful and interpretable tool for a specific but important class of time-series forecasting problems. Future work could focus on extending the theoretical guarantees to more general missing data patterns and incorporating mechanisms to handle non-periodic components within the NMF framework.

References

O. Akbilgic. Istanbul Stock Exchange. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C54P4J.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014.

A. Cichocki and R. Zdunek. *NMFLAB for signal processing*, 2006.

A. Cichocki, R. Zdunek, A.H. Phan, and S.-I. Amari. *Nonnegative matrix and tensor factorization: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley and Sons, 2009.

L. Condat. Fast projection onto the simplex and the l_1 ball. *Mathematical Programming*, 158(1–2):575–585, 2016.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. K. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems 16*, pp. 1141–1148. MIT Press, 2004.

R. Douc, E. Moulines, and D. Stoffer. *Nonlinear time series: Theory, methods, and applications with R examples*. Chapman & Hall/CRC, 2014.

J. Fonollosa. Twin gas sensor arrays. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C5MW3K.

N. Gillis. The why and how of nonnegative matrix factorization. In J.A.K. Suykens, M. Signoretto, and A. Argyriou (eds.), *Regularization, optimization, kernels, and support vector machines*, Machine Learning and Pattern Recognition Series, pp. 257–291. Chapman & Hall/CRC, 2014.

N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4):1085–1105, 2012. doi: 10.1162/NECO_a_00256. URL https://doi.org/10.1162/NECO_a_00256.

N. Gillis and A. Kumarg. Exact and heuristic algorithms for semi-nonnegative matrix factorization. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1404–1424, 2015.

H. Javadi and A. Montanari. Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association*, 115(530):896–907, 2020a.

H. Javadi and A. Montanari. Supplement To "Non-negative Matrix Factorization via Archetypal Analysis". *Journal of the American Statistical Association*, 115(530):896–907, 2020b. URL https://doi.org/10.1080/01621459.2019.1594832.

H. Laurberg, M.G. Christensen, M.D. Plumbley, L.K. Hansen, and S.H. Jensen. Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, 2008:1–9, 2008.

Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long. iTransformer: inverted transformers are effective for time series forecasting. *International Conference on Representation Learning 2024 (ICLR 2024)*, 2024.

J. Mei, Y. De Castro, Y. Goude, and G. Hébrail. Nonnegative matrix factorization for time series recovery from a few temporal aggregates. In *Proceedings of the 34th International Conference on Machine Learning*. JMLR: W&CP, 2017.

J. Mei, Y. De Castro, Y. Goude, J.-M. Azaïs, and G. Hébrail. Nonnegative matrix factorization with side information for time series recovery and prediction. *IEEE Transactions on Knowledge and Data Engineering*, 31(3):493–506, 2018.

Z. Ni, H. Yu, S. Liu, J. Li, and W. Lin. BasisFormer: Attention-based time series forecasting with learnable and interpretable basis. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

T. Pock and S. Sabach. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787, 2016. doi: 10.1137/16M1064064. URL https://doi.org/10.1137/16M1064064.

B. Recht, C. Re,, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1214–1222. Curran Associates, Inc., 2012.

A. Shewalkar, D. Nyavanandi, and S.A. Ludwig. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, 9:235–245, 2019.

P. Tang and W. Zhang. Unlocking the power of patch: Patch-Based MLP for long-term time series forecasting. *The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*, 2025.

L. Thomas. Solution to problem 73–14, rank factorizations of nonnegative matrices. *SIAM Review*, 16(1):393–394, 1974.

A. Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86.

S.A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009. ISSN 1052–6234.

S. Wang, H. Wu, X. Shi, T. Hu, H. Luoand L. Ma, J.Y. Zhang, and J. Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. *International Conference on Representation Learning 2024 (ICLR 2024)*, 2024.

T. Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11:128–149, 1976.

H. Wu, J. Xu, J. Wang, and M. Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, 2020.