Journal of Scientific Computing (2024) 100:29
https://doi.org/10.1007/s10915-024-02565-7

Supervised Low-Rank Semi-nonnegative Matrix Factorization with Frequency Regularization for Forecasting Spatio-temporal Data

Keunsu Kim[1] · Hanbaek Lyu[2] · Jinsu Kim[1] · Jae-Hun Jung[1]

Received: 14 November 2023 / Revised: 22 April 2024 / Accepted: 4 May 2024 /
Published online: 14 June 2024

Abstract
We propose a novel methodology for forecasting spatio-temporal data using supervised semi-nonnegative matrix factorization (SSNMF) with frequency regularization. Matrix factorization is employed to decompose spatio-temporal data into spatial and temporal components. To improve clarity in the temporal patterns, we introduce a nonnegativity constraint on the time domain along with regularization in the frequency domain. Specifically, regularization in the frequency domain involves selecting features in the frequency space, making an interpretation in the frequency domain more convenient. We propose two methods in the frequency domain: soft and hard regularizations, and provide convergence guarantees to first-order stationary points of the corresponding constrained optimization problem. While our primary motivation stems from geophysical data analysis based on GRACE (Gravity Recovery and Climate Experiment) data, our methodology has the potential for wider application. Consequently, when applying our methodology to GRACE data, we find that the results with the proposed methodology are comparable to previous research in the field of geophysical sciences but offer clearer interpretability.

Keywords Time-series · Dictionary learning · Matrix factorization · Fourier transform · Regularization

Mathematics Subject Classification 65F22 · 65F55 · 86A04

✉ Jae-Hun Jung
jung153@postech.ac.kr

Keunsu Kim
keunsu@postech.ac.kr

Hanbaek Lyu
hlyu@math.wisc.edu

Jinsu Kim
jinsukim@postech.ac.kr

[1] The Department of Mathematics, POSTECH, Pohang 37673, Republic of Korea

[2] The Department of Mathematics, University of Wisconsin-Madison, Wisconsin 53706, USA

1 Introduction

Time-series data analysis is primarily divided into two domains: time domain analysis and frequency domain analysis, also referred to as spectral analysis. Spectral analysis specifically examines the periodic properties of time-series data and summarizes them through the spectral density function. The primary objective of spectral analysis is to estimate or infer the spectral density function. In this paper, we propose a method that combines the usual spectral analysis and nonnegative matrix factorization.

Matrix factorization is a technique used to decrease the rank of a matrix by factorizing it into two low-rank matrices and extracting the latent features of the data. A low-rank matrix means that each column (row) is linearly dependent, or in other words, there is a correspondence for each column (row). Among matrix factorizations, nonnegative matrix factorization involves factorizing a nonnegative matrix into two nonnegative matrices. In [13], the nonnegative constraint is used to achieve interpretability. However, in general, datasets may have negative values, so we need to relieve the nonnegative constraint on factorization. Semi-nonnegative matrix factorization is a relaxation technique motivated by K-means clustering, as described in [8].

Matrix factorization is an important technique used in this paper for separating the spatial and temporal components of spatio-temporal data. The concept of breaking down complex problems into multiple components has a long history in mathematical research. For instance, the following statement by René Descartes aligns with our statement—To divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution.[1]

By decomposing the data into lower-dimensional matrices, matrix factorization can reveal underlying patterns that are not immediately apparent from the raw data. This approach is similar to the separation of variables in partial differential equations, which is used to simplify the equations and make them easier to solve. In particular, the separation of variables allows us to break down a complex equation into multiple simpler equations, each of which can be solved independently. Matrix factorization also follows the similar decomposition method. Matrix factorization is a unsupervised learning process, but for the prediction problem, we need a supervised learning, particularly in the form of semi-nonnegative matrix factorization. Supervised Nonnegative Matrix Factorization is proposed in [1] to solve the classification problem. In [1] it is proposed to use regularization in the loss function of nonnegative matrix factorization, which is a cross-entropy term between the ground truth and the class label. In contrast, our proposed method is a supervised semi-nonnegative matrix factorization with regularization in the frequency domain for forecasting spatio-temporal data. The traditional approach involves employing regularization in the physical domain. However, numerous time-series data characterized by time periodicity, such as those derived from geophysical applications exhibit distinct characteristics in terms of frequency. In other words, the patterns observed in the physical domain display periodicity aligned with Earth's motion, often with a noticeable yearly frequency. Utilizing such frequency information for regularization yields convenience in terms of interpretation. Moreover, for forecasting problems, incorporating frequency helps in predicting the spatial patterns that undergo periodic changes. We introduce a novel approach using matrix factorization with regularization in the frequency domain for forecasting problems.

[1] Discourse on the method of rightly conducting the reason and seeking the truth in the sciences, René Descartes (1637) edited by Charles W. Eliot, P.F. Collier & Son, 1909, New York. Transcribed by Andy Blunden. https://www.marxists.org/reference/archive/descartes/1635/discourse-method.htm.

For the proposed method, we explore two approaches for the proposed regularization method in the frequency domain: soft regularization and hard regularization. The soft regularization method permits the presence of noisy or insignificant frequencies in the regularization process, whereas the hard regularization method enforces such frequencies to be eliminated. Thus, for soft regularization, it is not necessary to pre-specify the frequencies to be removed, but it cannot completely eliminate specific frequencies. We use a high level of block coordinate descent (BCD) and the projected subgradient descent method for the implementation. Further we prove the convergence properties of BCD in this paper. The difficulty arises from the fact that the Fourier transform operates in the complex domain, and the given penalty is not always differentiable. In hard frequency regularization, we apply the results from the three operator splitting method. This algorithm can entirely remove specific frequencies, but it requires prior knowledge to specify which frequencies to remove. To address these challenges with hard frequency regularization, we introduce a heuristic method for solving hard frequency regularization. The heuristic method allows for the removal of specific frequencies without prior knowledge, but there is no guarantee of convergence.

This paper is composed of the following sections. In Sect. 2 we propose supervised semi-nonnegative matrix factorization with a novel regularization method, that is, soft/hard regularization in the frequency domain. In Sect. 3 we present the algorithms to solve the problem proposed in Sect. 2. In Sect. 4 we provide theoretical convergence analysis of the algorithms introduced in Sect. 3. In Sect. 5 we apply our proposed method to synthetic data to clarify the differences between the proposed method and existing methods, specifically focusing on matrix factorization and regularization methods. Finally, in Sect. 6 we apply the proposed method to the GRACE data for forecasting and compare the results with those by the methods currently available in the geophysical community.

2 Methods

2.1 Problem Setup and Low-Rank Spatio-temporal Model

Consider two fixed integer time T and T_{tot}, $T_{tot} > T \geq 0$ and suppose that we have tensor data $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{A\times B\times T_{tot}}$. We view each of $\mathcal{X}$ and $\mathcal{Y}$ as a 'spatio-temporal' data, where $\mathcal{X}[a, b, t]$ encodes the observed value at 'spatial location' (a, b) at 'time' t. One can view $\mathcal{X}$ as the evolution of a real-valued observable (e.g., gravitational force) in a spatial region encoded as a $A \times B$ matrix over the period of discrete time $\{0, \ldots, T_{tot} - 1\}$, but with the 'temporal slices' $\mathcal{X}[:, :, t]$, for times between T and $T_{tot} - 1$, missing. Similarly, we view $\mathcal{Y}$ as the evolution of an auxiliary observable (e.g., precipitation) over the same spatial region over the longer period of time $\{0, \ldots, T_{tot} - 1\}$. The main question we address in this paper is the following: Can we find spatio-temporal patterns from the joint tensor data $(\mathcal{X}, \mathcal{Y})$ and use them to predict $\mathcal{X}$ over the missing period of $[T, T_{tot})$?

Our approach is based on the following assumption that there are r latent spatial patterns and each temporal slice of $\mathcal{X}$ and $\mathcal{Y}$ is their nonnegative linear combination. Such r spatial patterns can be encoded as two tensors $\mathcal{W}, \mathcal{W}' \in \mathbb{R}^{A\times B\times r}$. We hypothesize that each temporal slice $\mathcal{X}[:, :, t]$ and $\mathcal{Y}[:, :, t]$ can be approximated by some linear combination of the r spatial patterns $\mathcal{W}[:, :, s]$ and $\mathcal{W}'[:, :, s]$ for $s = 0, \ldots, r - 1$. That is, there exist *common* and *nonnegative* coefficients $h_{s,t} \geq 0$ for $s = 0, \ldots, r - 1$ such that

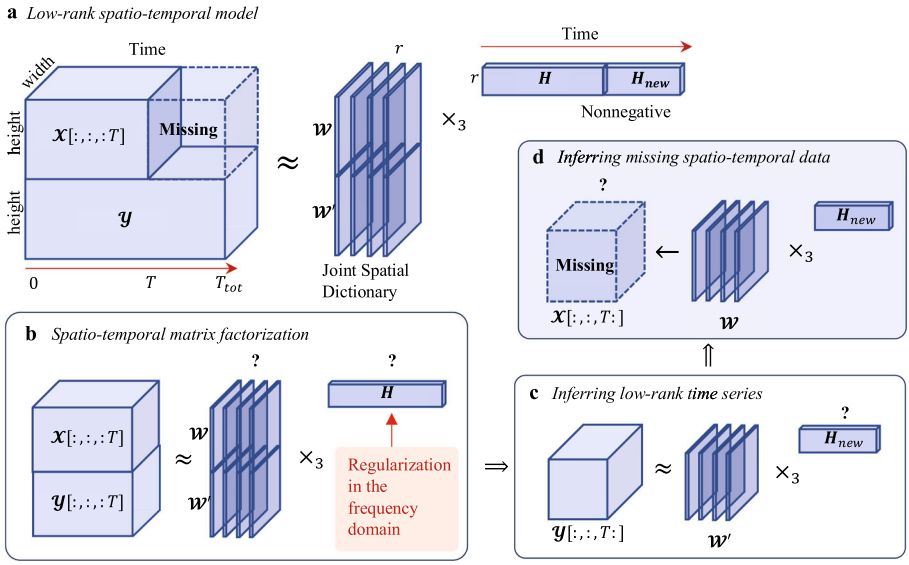

Fig. 1 (**a**) Schematic illustration of the low-rank spatio-temporal model. (**b**) From the joint spatio-temporal data during the period $[0, T)$, we learn the latent spatial patterns $(\mathcal{W}, \mathcal{W}')$ and low-rank time-series $\mathbf{H}$, where the temporal structure of $\mathbf{H}$ is regularized in the frequency domain. (**c**) Using the auxiliary data $\mathcal{Y}$ and the latent dictionary $\mathcal{W}'$, we infer the low-rank time-series $\mathbf{H}_{\text{new}}$ during the missing period $[T, T_{\text{tot}})$. (**d**) The missing data $\mathcal{X}[:, :, T:]$ is then inferred as $\mathcal{W} \times_3 \mathbf{H}_{\text{new}}$

$$\mathcal{X}[:, :, t] \approx \sum_{s=0}^{r-1} \mathcal{W}[:, :, s] h_{s,t}, \quad \mathcal{Y}[:, :, t] \approx \sum_{s=0}^{r-1} \mathcal{W}'[:, :, s] h_{s,t}. \tag{1}$$

Hence the pair of $(\mathcal{W}[:, :, s], \mathcal{W}'[:, :, s])$ encodes possible spatial patterns of the two observables recorded in $\mathcal{X}, \mathcal{Y}$ that can be *simultaneously observed* in the spatial region $A \times B$. The underlying latent statio-temporal model can be concisely re-stated as the following tensor factorization equations:

$$\mathcal{X} \approx \mathcal{W} \times_3 \widetilde{\mathbf{H}}, \quad \mathcal{Y} \approx \mathcal{W}' \times_3 \widetilde{\mathbf{H}}, \tag{2}$$

where $\times_3$ denotes the mode-3 tensor-matrix product and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times T}$ and $\widetilde{\mathbf{H}}[s, t] = h_{s,t}$ for $t \in [0, T_{tot})$ encode the rank-r representation of the joint time-series $(\mathcal{X}, \mathcal{Y})$ over the basis of $(\mathcal{W}, \mathcal{W}')$. We call (2) the low-rank spatio-temporal model (See Fig. 1a for an illustration.)

Remark 1 We remark that extending our model for multiple auxiliary spatio-temporal data $\mathcal{Y}_0, \ldots, \mathcal{Y}_{S-1} \in \mathbb{R}^{A \times B \times T_{tot}}$ is straightforward by simply concatenating them into a single $AS \times B \times T_{tot}$ tensor $\mathcal{Y}$ and use the same framework as in (2) except that the 'height' of $\mathcal{Y}$ and $\mathcal{W}'$ is now AS instead of A. For the simplicity of the discussion, we keep the setting $S = 1$. But later in our experiments, we use an instance of $S = 3$ where $\mathcal{Y}_0$ for normalized precipitation ranging between 0 and 1, $\mathcal{Y}_1$ for normalized temperature, and $\mathcal{Y}_2$ for normalized total water storage from the GLDAS Noah model.

2.2 Infering the Missing Spatio-temporal Data Given the Latent Spatial Patterns

The unknown parameters for the low-rank spatio-temporal model (2) are the latent spatial patterns $(\mathcal{W}, \mathcal{W}') \in \mathbb{R}^{2A\times B\times r}$ and the low-rank nonnegative time-series $\widetilde{\mathbf{H}} \in \mathbb{R}^{r\times T_{tot}}_{\geq 0}$. Learning these parameters simultaneously should be carefully formulated through constrained non-convex optimization problems, which we will discuss in the following section. Here we first discuss how we can infer the missing spatio-temporal data $\mathcal{X}[:, :, T:] \in \mathbb{R}^{A\times B\times(T_{\text{tot}}-T)}$ by assuming that we have the latent spatial patterns $(\mathcal{W}, \mathcal{W}')$. Here the notation T: means all the elements starting from index T to the end. In a similar way, $: T$ means all the elements until index $T-1$ excluding T. This can be done through the following three steps as below:

$$\text{(Encoding)} \quad \mathbf{H}'_{\text{new}} \leftarrow \underset{\mathbf{U}\in\mathbb{R}^{r\times T_{\text{tot}}}_{\geq 0}}{\arg\min} \|\mathcal{Y} - \mathcal{W}' \times_3 \mathbf{U}\|_F^2 \tag{3}$$

$$\text{(Slicing)} \quad \mathbf{H}_{\text{new}} \leftarrow \mathbf{H}'_{\text{new}}[:, T:] \tag{4}$$

$$\text{(Prediction)} \quad \mathcal{X}[:, :, T:] \leftarrow \mathcal{W} \times_3 \mathbf{H}_{\text{new}}, \tag{5}$$

where the subscript F denotes the Frobenius norm.

Namely, according to the low-rank spatio-temporal model (2), the auxiliary data $\mathcal{Y}$ during the missing period $[T, T_{\text{tot}})$ is represented by the linear model $\mathcal{Y} \approx \mathcal{W}' \times_3 \mathbf{H}_{\text{new}}$ for some nonnegative matrix $\mathbf{H}_{\text{new}}$. Hence, by solving the nonnegative least-squares problem in (3), we can infer the low-rank time-series $\mathbf{H}_{\text{new}}$. Now $\mathbf{H}_{\text{new}}$ is also used to represent the missing data $\mathcal{X}[:, :, T:]$ over the corresponding latent spatial patterns in $\mathcal{W}$. Thus, we can infer the missing spatio-temporal data as in (5). See Fig. 1c, d for an illustration.

Remark 2 It is crucial to construct $\mathbf{H}_{\text{new}}$ by considering the entire period of $\mathcal{Y}$ in (3) to accurately infer the missing data. While it may seem doable to use $\mathcal{Y}[:, :, T:]$ for this purpose, it is important to note that in the frequency domain, only $\lfloor \frac{T}{2} \rfloor + 1$ frequencies are present for the time-series of length T. Consequently, relying solely on the test period information $[T, T_{tot})$ of $\mathcal{Y}$ to construct $\mathbf{H}_{\text{new}}$ could result in a loss of frequency information. For instance, when the testing period spans 10 months, Encoding step of (3) solely with this period would fail to capture the annual patterns. However, by considering the entire duration, even within those 10 months, we can infer the annual patterns reasonably.

2.3 Spatio-temporal Matrix Factorization with Frequency Regularization

In this section, we introduce and propose the matrix factorization method with frequency regularization. In general, due to the nature of time-series data, specific frequencies often hold significant physical meanings. Thus, analysis with these frequencies can greatly facilitate the meaningful interpretation of the underlying physics in the given data. Consequently, regularization in the frequency domain may be more advantageous for matrix factorization than regularizing in the time domain. In this section, we aim to propose methods that incorporate this idea.

Decompose the low-rank time-series $\widetilde{\mathbf{H}}$ as $[\mathbf{H}, \mathbf{H}_{\text{new}}]$, where $\mathbf{H} \in \mathbb{R}^{r\times T}_{\geq 0}$ and $\mathbf{H}_{\text{new}} \in \mathbb{R}^{r\times T_{tot}-T}_{\geq 0}$. With a slight abuse of notation, we will denote $\mathcal{X}[:, :, : T]$ and $\mathcal{Y}[:, :, : T]$ as $\mathcal{X}$ and $\mathcal{Y}$, respectively, in this section. The main task of this method is to learn the joint latent spatial patterns $(\mathcal{W}, \mathcal{W}')$ and the common nonnegative low-rank time-series $\mathbf{H}$ from the joint observed time-series data $(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{2A\times B\times T}$. Here, we make the assumption that the observations were gathered over a sufficient period of time, and we consider the loss of frequency information noted in Remark 2 is insignificant. The necessary duration for

sufficient observation time can vary depending on the specific data and the temporal pattern under consideration. For instance, for the GRACE data, as discussed in Sect. 6, a minimum of 12 months of observations is required to discern the expected annual pattern. With this assumption we formulate the problem as the following spatio-temporal matrix factorization problem over $[0, T)$:

$$\begin{aligned} &\text{minimize} \quad \|\mathcal{X} - \mathcal{W} \times_3 \mathbf{H}\|_F^2 + \xi \|\mathcal{Y} - \mathcal{W}' \times_3 \mathbf{H}\|_F^2 + \psi(\mathbf{H}), \\ &\text{subject to} \quad \mathcal{W}, \mathcal{W}' \in \mathbb{R}^{A \times B \times r} \text{ and } \mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times T} \end{aligned} \tag{6}$$

where $\xi \geq 0$ is a *supervision parameter* and $\psi(\mathbf{H})$ is a penalty term for $\mathbf{H}$, which will take one of the following choices below:

(**a**) (Ridge penalty) $\psi(\mathbf{H}) = \lambda \|\mathbf{H}\|_F^2$
(**b**) (Lasso penalty) $\psi(\mathbf{H}) = \lambda \|\mathbf{H}\|_1$
(**c**) (Soft frequency regularization) $\psi(\mathbf{H}) = \lambda \|\widehat{\mathbf{H}}\|_{1,M}$ ($\triangleright$ $\|\cdot\|_{1,M}$ =Minkowski 1-norm)
(**d**) (Hard frequency regularization) $\psi(\mathbf{H}) = 0$ if $\widehat{\mathbf{H}}$ only uses a prespecified set of frequencies and ∞ otherwise.

Here $\lambda \geq 0$ is a regularization parameter and $\widehat{\mathbf{H}}$ denotes the Fourier transform of $\mathbf{H}$.

The rationale behind regularizing the temporal structure of $\mathbf{H}$ in (6) is as follows. The observed data ($\mathcal{X}$ as well as $\mathcal{Y}$) may contain noise so we should avoid overfitting the latent spatial patterns to accurately represent the observation. Instead of directly regularizing the latent spatial patterns, we find it more effective to regularize the associated low-rank time-series representation $\mathbf{H}$ so that it is sparse in the frequency domain. That is, our guiding principle is that the observed spatio-temporal data admits a low-rank time-series representation $\mathbf{H}$, which uses a small number of dominating frequencies. Our intuition is largely inspired by analyzing geospatial data (e.g., GRACE data, see Sec. 6.1), where typically the annual or semi-annual spatio-temporal patterns are dominating, and other frequencies are often under-represented. Options (**a**-**b**) above use standard Ridge and Lasso (least absolute shrinkage and selection operator) penalties commonly used in the machine learning and statistics literature [11]. (We opt for using the Lasso penalty being a convex relaxation of the computationally more challenging ℓ_0-penalty [17].) Note that both penalties regularize $\mathbf{H}$ in the time domain, whereas the options (**c**-**d**) regularize the Fourier transform $\widehat{\mathbf{H}}$ of $\mathbf{H}$ so the regularization is done directly in the frequency domain. We will discuss the latter two options in more detail in the next section.

Notice that we referred to (6) as a matrix factorization problem while it involves factorizing tensors $\mathcal{X}$ and $\mathcal{Y}$. This is because one can matricize these tensors as well as the latent spatial patterns $(\mathcal{W}, \mathcal{W}')$ to reformulate (6) as a matrix factorization problem. Namely, let $\mathbf{X} := (\mathcal{X}_{(3)})^T \in \mathbb{R}^{AB \times T}$ denote the matrix by matricizing the tensor $\mathcal{X}$ along the time mode, where $\mathcal{X}_{(3)} \in \mathbb{R}^{T \times AB}$ denotes the mode-3 unfolding of $\mathcal{X}$. Similarly, denote $\mathbf{Y} := (\mathcal{Y}_{(3)})^T$, $\mathbf{W} := (\mathcal{W}_{(3)})^T$, and $\mathbf{W}' := (\mathcal{W}'_{(3)})^T$. Then we have $\|\mathcal{X} - \mathcal{W} \times_3 \mathbf{H}\|_F = \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F$ and $\|\mathcal{Y} - \mathcal{W}' \times_3 \mathbf{H}\|_F = \|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F$. So henceforth, we will consider the following supervised semi-nonnegative matrix factorization (SSNMF) problem instead of (6): (denoting $d := AB$)

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi \|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2 + \psi(\mathbf{H}) \\ &\text{subject to} \quad \mathbf{W}, \mathbf{W}' \in \mathbb{R}^{d \times r} \text{ and } \mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times T}. \end{aligned} \tag{7}$$

3 Algorithms

3.1 Soft Frequency Regularization

Spectral analysis is popularly used in time-series analysis. It uncovers the underlying periodic patterns present in the given time-series data. This method revolves around the concept of the spectral density function, which is a probability density function describing the distribution of the modulus of frequencies across different frequencies in the data. The primary goal of spectral analysis is to accurately estimate the spectral density function using the sampled time-series data. The discrete Fourier transform is widely utilized for such estimation. However, the expectation value of the periodogram at a certain frequency does not converge to the spectral density function at that frequency even when dealing with a large number of samples. To mitigate this issue, smoothing in the frequency domain is commonly employed in practical time-series analysis [5]. We aim to integrate spectral analysis in classical time-series data analysis with SSNMF. That is, instead of analyzing the time-series data corresponding to each spatial data, we utilize regularization from the perspective of analyzing in the frequency domain of the time-series data obtained through the dimensional reduction of the given spatio-temporal data using SSNMF. In Proposition 5, a mismatch term between $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ can disrupt the inference of the periodic pattern of $\mathbf{X}$. Therefore, regularization in the frequency domain may help mitigate the impact of this disparity in forecasting.

Definition 1 (*Fourier transform*) For a matrix $\mathbf{A} \in \mathbb{R}^{r\times T}$, we regard its rows $\mathbf{A}[s]$ for $s = 0, \ldots, r-1$ as a univariate time-series where the column index corresponds to time. With this interpretation, we can define the *Fourier transform* of $\mathbf{A}$, denoted as $\widehat{\mathbf{A}}$, to be the $r \times T$ matrix of complex coefficients defined as

$$\widehat{\mathbf{A}}[s,k] := \frac{1}{T}\sum_{f=0}^{T-1} \mathbf{A}[s,\, f]e^{-2\pi i f k/T} \tag{8}$$

for $0 \le s \le r-1$, $0 \le k \le T-1$. Denote by $\mathcal{F}_T$ the $T \times T$ *Fourier matrix* whose (f,k) coordinate equals $e^{-2\pi i f k/T}/T$. Then we can write

$$\widehat{\mathbf{A}} = \mathbf{A}\mathcal{F}_T. \tag{9}$$

In the case of Lasso, when we reduce the number of features, it simplifies the interpretation of each parameter. Similarly, when we introduce sparsity in the frequency domain, it simplifies the interpretation of the patterns in the temporal data obtained through matrix factorization. Therefore, we propose the following statio-temporal decomposition problem with frequency regularization:

$$\min_{\substack{\mathbf{W},\mathbf{W}'\in\mathbb{R}^{d\times r}\\ \widehat{\mathbf{H}}\in\mathbb{C}^{r\times T},\\ \widehat{\mathbf{H}}_{sk}=\overline{\widehat{\mathbf{H}}}_{s,T-k},\widehat{\mathbf{H}}\mathcal{F}_T^{-1}\ge 0}} \|\widehat{\mathbf{X}}[:,:T] - \mathbf{W}\widehat{\mathbf{H}}\|_F^2 + \xi\|\widehat{\mathbf{Y}}[:,:T] - \mathbf{W}'\widehat{\mathbf{H}}\|_F^2 + \lambda\|\widehat{\mathbf{H}}\|_{1,M}. \tag{10}$$

Here, $\xi, \lambda \ge 0$ are tuning parameters and for a complex matrix $C \in \mathbb{C}^{r\times T}$, $\|C\|_{1,M}$ denotes its *Minkowski 1-norm* defined as

$$\|C\|_{1,M} := \sum_{a,b}\|C[a,b]\|_{1,M},\ \ \|x+iy\|_{1,M} := |x| + |y|. \tag{11}$$

The condition $\widehat{\mathbf{H}}_{sk} = \overline{\widehat{\mathbf{H}}}_{s,T-k}$ is needed because we want $\mathbf{H}$ to be a real matrix and the condition $\widehat{\mathbf{H}}\mathcal{F}_T^{-1} \ge 0$ is applied to the nonnegative constraint on $\mathbf{H}$.

Note that by the Parseval's identity, $\|\widehat{\mathbf{X}} - \mathbf{W}\widehat{\mathbf{H}}\|_F^2 = \|(\mathbf{X} - \mathbf{W}\mathbf{H})\mathcal{F}_T\|_F^2 = \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$. Therefore problem (10) is equivalent to

$$\min_{\substack{\mathbf{H}\in\mathbb{R}_{\geq 0}^{r\times T} \\ \mathbf{W},\mathbf{W}'\in\mathbb{R}^{d\times r}}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2 + \lambda\|\mathbf{H}\mathcal{F}_T\|_{1,M}. \tag{12}$$

Moreover, $\|\widehat{\mathbf{H}}\|_{1,M} = \frac{1}{T}\|\mathbf{H}\|_1 + \sum_{s,k=1} \|\widehat{\mathbf{H}}[s,k]\|_{1,M}$ holds. This relation tells us that the Minkowski 1-norm contains $L1$-regularization term.

Algorithm 1 below takes the form of block coordinate descent for solving SSNMF (7) for three instances of regularization: Ridge and Lasso regularization in the time domain; and soft regularization in the frequency domain (see (12)).

Algorithm 1 SSNMF with Ridge, Lasso and soft frequency regularization

1: **Input:** $\mathbf{X} \in \mathbb{R}^{d\times T}$ (Data) and $\mathbf{Y} \in \mathbb{R}^{d\times T_{tot}}$ (Auxiliary data);

2: **Variables:** $N \in \mathbb{N}$ (iterations), $L \in \mathbb{N}$ (sub-iterations); $\lambda \geq 0$ (regularization parameter);

3: **Regularizer:** $\psi(\mathbf{H}) = \|\mathbf{H}\|_F^2$ (Ridge) or $\|\mathbf{H}\|_1$ (Lasso) or $\|\mathbf{H}\mathcal{F}_T\|_{1,M}$ (soft frequency regularization)

4: *Initialize:* $\mathbf{W}_0, \mathbf{W}'_0 \in \mathbb{R}^{d\times r}$ and $\mathbf{H}_0 \in \mathbb{R}^{r\times T}$

5: **for** $j = 1, \cdots, N$ **do**

6:

$$\mathbf{H}_j \leftarrow \underset{\mathbf{H}\in\mathbb{R}_{\geq 0}^{r\times T}}{\arg\min} \left\| \begin{bmatrix} \mathbf{X} \\ \sqrt{\xi}\mathbf{Y}[:,:T] \end{bmatrix} - \begin{bmatrix} \mathbf{W}_{j-1} \\ \sqrt{\xi}\mathbf{W}'_{j-1} \end{bmatrix} \mathbf{H} \right\|_F^2 + \lambda\psi(\mathbf{H}) \tag{13}$$

(▷ Use projected subgradient descent. Sub-iteration L is required here.)

$$\mathbf{W}_j \leftarrow \underset{\mathbf{W}\in\mathbb{R}^{d\times r}}{\arg\min} \ ||\mathbf{X} - \mathbf{W}\mathbf{H}_j||_F^2 \quad \left(\triangleright\ \mathbf{W}_j = \mathbf{X}{\mathbf{H}_j}^T(\mathbf{H}_j{\mathbf{H}_j}^T)^{-1}\right) \tag{14}$$

$$\mathbf{W}'_j \leftarrow \underset{\mathbf{W}'\in\mathbb{R}^{d\times r}}{\arg\min} \ ||\mathbf{Y}[:,:T] - \mathbf{W}'\mathbf{H}_j||_F^2 \quad \left(\triangleright\ \mathbf{W}'_j = \mathbf{Y}[:,:T]{\mathbf{H}_j}^T(\mathbf{H}_j{\mathbf{H}_j}^T)^{-1}\right) \tag{15}$$

7: **end for**

8: (Encoding)

$$\mathbf{H}'_{\text{new}} \leftarrow \underset{\mathbf{H}\in\mathbb{R}_{\geq 0}^{r\times T_{tot}}}{\arg\min} \ ||\mathbf{Y} - \mathbf{W}'\mathbf{H}||_F^2 + \frac{\lambda}{\xi}\psi(\mathbf{H}) \tag{16}$$

(▷ Use projected subgradient descent. Sub-iteration L is required here.)

$$\mathbf{H}_{\text{new}} \leftarrow \mathbf{H}'_{\text{new}}[:, T:] \tag{17}$$

9: **return** $\mathbf{W}_N, \mathbf{W}'_N, \mathbf{H}_N, \mathbf{H}_{\text{new}}$.

3.2 Hard Constraints in the Frequency Domain

Lasso introduces sparsity in the time domain, allowing us to select proper features. Similarly, we employ soft frequency regularization to introduce sparsity in the frequency domain. However, in our experiments (Examples 3 and 4), we observed that the soft constraint reduces the utilization of less dominant frequencies but does not completely eliminate them. As a result, the soft constraint in the frequency domain partially fulfills our objective. This partial

achievement led us to introduce the hard constraint in the frequency domain, enabling us to selectively eliminate specific frequencies. Our algorithm is based on the three operator splitting problem, and we have proven its convergence. However, for this algorithm specific frequencies that need to be eliminated should be predetermined in advance. And determining these frequencies necessitates prior knowledge of temporal patterns. Therefore, we introduce a heuristic method to address the hard constraint problem, which does not rely on prior knowledge. By doing so, we can achieve a clear temporal pattern related to the corresponding spatial information. This additional step enhances our ability to identify and interpret the meaningful patterns in the data.

Consider the three operator splitting problem to find $x \in \mathbb{R}^n$ such that

$$\min_{x\in\mathbb{R}^n} f(x) + g(x) + h(x). \tag{18}$$

If g and h are indicator functions on the convex domain $\mathcal{G}$ and $\mathcal{H}$, respectively, then the three operator splitting problem becomes

$$\min_{x\in\mathbb{R}^n} f(x) \text{ subject to } x \in \mathcal{G} \cap \mathcal{H}. \tag{19}$$

The method for finding a value of x is referred to as the following proposition. Define the ergodic sequences $\bar{x}_t = \frac{1}{t+1}\sum_{\tau=0}^{t} x_\tau$. To prioritize satisfying the nonnegative condition over suppressing specific frequencies, we set

$$f(\mathbf{H}) = \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2, \tag{20}$$

$$g(\mathbf{H}) = \begin{cases} 0, & \text{if } \mathbf{H} \geq 0 \\ \infty, & \text{Otherwise} \end{cases}, \tag{21}$$

$$h_R(\mathbf{H}) = \begin{cases} 0, & \text{if } \mathbf{H} \in \mathcal{F}_T^{-1}(\mathbb{C}^{r\times R} \times \{0\}^{T-R}) \\ \infty, & \text{Otherwise} \end{cases}, \tag{22}$$

here, the subscript R represents the number of frequencies to remain. If the priority is to suppress specific frequencies rather than to satisfy the nonnegative condition, the functions g and h_R should be interchanged.

Instead of Algorithm 2, we found experimentally that the following algorithm also converges for updating $\mathbf{H}$, even though there is no convergence guarantee. Moreover, unlike the previous method, the advantage of this method is that it does not require specifying the frequencies to be removed in advance. Furthermore, Algorithm 2 should also be carefully applied during the encoding process. In Example 1, we illustrate an example of time-series data with the same period but different lengths. If we choose Algorithm 2 in Algorithm 3 (in the step of (27)), we should consider the difference of the time-series length when obtaining $\mathbf{H} \in \mathbb{R}^{r\times T}$ and $\mathbf{H}'_{\text{new}} \in \mathbb{R}^{r\times T_{tot}}$. If we remove specific frequencies during the coding process (27), how can we remove frequencies during the encoding process (30)? Assume that the training period is $T = 132$ months, and the total period (including both the training and test periods) is $T_{tot} = 163$ months. For example, in coding process the 11th Fourier coefficient corresponds to a 12 month cycle but, during the encoding process, the 11th Fourier coefficient corresponds to the 14.8 month cycle. This raises the question of how to handle frequency removal in the encoding process.

Algorithm 2 Three Operator Splitting [23]

1: **Input:** Initial point $y_0 \in \mathbb{R}^{r\times T}$; $R \in \mathbb{N}$ (the number of remaining frequencies)

2: **for** $j = 0, \cdots, N$ **do**

3:

$$\mathbf{H}_j = \underset{\mathbf{H}\in\mathbb{R}^{r\times T}}{\arg\min} \left\{ \gamma_j g(\mathbf{H}) + \frac{1}{2}\|y_j - \mathbf{H}\|^2 \right\} \tag{23}$$

$$\mathbf{G}_j = \underset{\mathbf{G}\in\mathbb{R}^{r\times T}}{\arg\min} \left\{ \gamma_j h_R(\mathbf{G}) + \frac{1}{2}\|2\mathbf{H}_j - y_j - \gamma_j \nabla f(\mathbf{H}_j) - \mathbf{G}\|^2 \right\} \tag{24}$$

$$y_{j+1} = y_j - \mathbf{H}_j + \mathbf{G}_j \tag{25}$$

$$\gamma_j = \frac{1}{\sqrt{\sum_{\tau=0}^{j-1} \|\nabla f(\mathbf{H}_\tau)\|^2}} \tag{26}$$

4: **end for**

5: **return** Ergodic sequence $\overline{\mathbf{H}}_N$.

Algorithm 3 SSNMF with hard constraint in the frequency domain

1: **Input:** $\mathbf{X} \in \mathbb{R}^{d\times T}$ (Data) and $\mathbf{Y} \in \mathbb{R}^{d\times T_{tot}}$ (Auxiliary data);

2: **Variables:** $N \in \mathbb{N}$ (iterations); $\lambda \geq 0$ (regularization parameter); $R \in \mathbb{N}$ (the number of remaining frequencies);

3: *Initialize:* $\mathbf{W}_0 \in \mathbb{R}^{d\times r}$ and $\mathbf{H}_0 \in \mathbb{R}^{d\times T}$

4: **for** $j = 1, \cdots, N$ **do**

5:

$$\mathbf{H}_j \leftarrow \text{Algorithm 2 or Algorithm 4} \tag{27}$$

$$\mathbf{W}_j \leftarrow \underset{\mathbf{W}\in\mathbb{R}^{d\times r}}{\arg\min} \ ||\mathbf{X} - \mathbf{W}\mathbf{H}_j||_F^2 \quad \left(\triangleright \text{ i.e. } \mathbf{W}_j = \mathbf{X}\mathbf{H}_j^T(\mathbf{H}_j\mathbf{H}_j^T)^{-1}\right) \tag{28}$$

$$\mathbf{W}'_j \leftarrow \underset{\mathbf{W}'\in\mathbb{R}^{d\times r}}{\arg\min} \ ||\mathbf{Y}[:,:T] - \mathbf{W}'\mathbf{H}_j||_F^2 \quad \left(\triangleright \text{ i.e. } \mathbf{W}'_j = \mathbf{Y}[:,:T]\mathbf{H}_j^T(\mathbf{H}_j\mathbf{H}_j^T)^{-1}\right) \tag{29}$$

6: **end for**

7:

$$\mathbf{H}'_{\text{new}} \leftarrow \text{Algorithm 2 or Algorithm 4} \tag{30}$$

$$\text{for } f(\mathbf{H}) = \|\mathbf{Y} - \mathbf{W}'_N\mathbf{H}\|_F^2 \tag{31}$$

$$\mathbf{H}_{\text{new}} \leftarrow \mathbf{H}'_{\text{new}}[:, T:] \tag{32}$$

8: **return** $\mathbf{W}_N, \mathbf{W}'_N, \mathbf{H}_N, \mathbf{H}_{\text{new}}$.

Example 1 For $M = 2T$, let $f(t) = \cos\frac{2\pi t}{T}$ for $t = 0, \ldots, T-1$ and $F(t) = \cos\frac{4\pi t}{M}$ for $t = 0, \ldots, M-1$, i.e. f is a partial information of F. Then $\hat{f}(k) = \frac{1}{T}\sum_{t=0}^{T-1} f(t)e^{-2\pi itk/T} = \frac{1}{2T}\sum_{t=0}^{T-1} e^{2\pi it(1-k)/T} + e^{-2\pi it(1+k)/T}$ implies that

$$\hat{f}(k) = \begin{cases} \frac{1}{2}, & \text{if } k = 1, T-1 \\ 0, & \text{otherwise} \end{cases},$$

and similarly

$$\hat{F}(k) = \begin{cases} \frac{1}{2}, & \text{if } k = 2, 2T-2 \\ 0, & \text{otherwise} \end{cases}.$$

In the frequency domain, the length of time-series distorts the frequency information. Therefore we should consider the difference in length between $\mathbf{H}$ (coding process) and $\mathbf{H}_{\text{new}}$ (encoding process).

The following Algorithm 4 is our heuristic approach to avoid the above problems.

Algorithm 4 Alternating projected gradient descent

1: **Input:** Initial point $\mathbf{H}_0 \in \mathbb{R}^{r\times T}$;

2: **Variable:** $R \in \mathbb{N}$ (the number of remaining frequencies);

3: **for** $j = 0, \cdots, N$ **do**

4: **for** $s = 0, \cdots, r-1$ **do**

5:

$$\mathcal{K}_j^s = \left\{k_1^s(j), \cdots, k_R^s(j), T-k_1^s(j), \cdots, T-k_R^s(j)\right\} \quad (33)$$

(set of preserving frequency indexes)

$$\mathbf{H}_j[s,t] \leftarrow \sum_{k\in\mathcal{K}_j^s} \widehat{\mathbf{H}}_j[s,k]e^{2\pi ikt/T}; \text{ delete frequencies (projection step)} \quad (34)$$

6: **end for**

7:

$$\mathbf{H}_j \leftarrow \mathbf{H}_j - \gamma_j \nabla f(\mathbf{H}_j), \text{ where } f(\mathbf{H}) = \left\| \begin{bmatrix} \mathbf{X} \\ \sqrt{\xi}\mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{W} \\ \sqrt{\xi}\mathbf{W}' \end{bmatrix} \mathbf{H} \right\|_F^2 \text{ (gradient step)} \quad (35)$$

$$\mathbf{H}_j \leftarrow \max\left\{0, \mathbf{H}_j\right\} \text{ (projection step)} \quad (36)$$

$$\gamma_j = \frac{1}{j+1} \cdot \frac{\nabla f(\mathbf{H}_j)}{(\mathbf{W}^T\mathbf{W}+1)} \quad (37)$$

8: **end for**

9: **return** $\mathbf{H}_N$.

In (34) of Algorithm 4, unlike Algorithm 2, it is not necessary to remove the same frequency at each iteration. In our experiments, for each $s \in \{0, \ldots, r-1\}$, we removed a few(hyperparameter) of the frequencies with low spectrum power from the temporal data $\mathbf{H}[s]$. In Sect. 6, we preserve the top R frequencies of the power spectrum density for each row of $\mathbf{H}$ in (33). More precisely, in (33), we choose (positive) frequencies $\left\{k_1^s(j), \ldots, k_R^s(j)\right\} \subseteq \left\{0, \ldots, \lfloor \frac{T}{2} \rfloor + 1\right\}$ that satisfy $\left|\widehat{\mathbf{H}}_j[s, k_1^s(j)]\right| \geq \cdots \geq \left|\widehat{\mathbf{H}}_j[s, k_R^s(j)]\right| \geq \left|\widehat{\mathbf{H}}_j[s,k]\right|$ for $k \in \left\{0, \ldots, \lfloor \frac{T}{2} \rfloor + 1\right\} \setminus \left\{k_1^s(j), \ldots, k_R^s(j)\right\}$. As mentioned in Algorithm 2, if you want to prioritize removing frequencies, then you change the steps of (34) and (36) in Algorithm 4.

4 Theoretical Guarantees

4.1 Statement of Results

In Sect. 2.3, we proposed spatio-temporal matrix factorization problems, with four choices of penalization term for $\mathbf{H}$. In Sect. 3, we proposed two iterative algorithms for solving these

problems. Namely, Algorithm 1 covers SSNMF with Ridge or Lasso regularization in the time domain and soft frequency regularization and Algorithm 3 covers SSNMF with hard frequency regularization. In this section, we state theoretical convergence guarantees for these algorithms.

Theorem 1 (Convergence of algorithm with Ridge, Lasso, and soft frequency regularization) *With Algorithm 1, $\mathbf{W}_j, \mathbf{W}'_j$ and $\mathbf{H}_j$ converge to stationary points of (7) with Ridge, Lasso and soft frequency regularization on* $\mathbf{H}$.

The convergence results for Ridge or Lasso regularization can be proven in a similar way to soft frequency regularization. Hence we will omit the details for these cases and prove Theorem 1 for soft frequency regularization in the following section. Next, we also establish a similar convergence result for SSNMF with hard frequency regularization using Algorithm 3.

Theorem 2 (Convergence of algorithm with hard frequency regularization) *In Algorithm 3, if we choose Algorithm 2 in (27) and (30), then $\mathbf{W}_j, \mathbf{W}'_j$ and $\mathbf{H}_j$ converge to stationary points of (7) with hard frequency regularization.*

Algorithms 1 and 3 take the form of block coordinate descent at a high level. Because our objective function is not continuously differentiable with a convex constraint on the domain, we use the projected subgradient descent method [4].

Algorithm 5 Block Coordinate Descent (BCD)

1: **Input:** Initial point $\boldsymbol{\theta}_0 = (\theta_0^1, \ldots, \theta_0^m) \in \Theta_1 \times \cdots \times \Theta_m$; N (number of iterations)
2: **for** $j = 0, \cdots, N$ **do**
3: **for** $i = 1, \cdots, m$ **do**
4:
$$\theta_{j+1}^i = \underset{\theta \in \Theta_i}{\arg\min}\, f(\theta_{j+1}^1, \ldots, \theta_{j+1}^{i-1}, \theta, \theta_j^{i+1}, \ldots, \theta_j^m)$$
5: **end for**
6: **end for**
7: **return** $\boldsymbol{\theta}_N$

Block Coordinate Descent (BCD) is an optimization technique used for nonconvex problems where the goal is to minimize a given objective function. In each iteration, BCD updates only one block of variables while keeping the other blocks constant. This implies that each subproblem tackled by BCD is a convex problem. But BCD does not always converge, even if f is componentwise convex [10]. The challenge lies in the fact that the range of the Fourier transform is the complex domain, and $\|\cdot\|_{1,M}$ is non-differentiable.

Definition 2 (*Wirtinger derivatives* [7]) We can regard the complex manifold $\mathbb{C}^n = \{z_1, \ldots, z_n\}$ as a real manifold $\mathbb{R}^{2n} = \{x_1, \ldots, x_n, y_1, \ldots, y_n\}$ for $z_j = x_j + iy_j$. In this context, for any point $p \in \mathbb{R}^{2n}$, the tangent space is given by $T_p\mathbb{R}^{2n} = span_{\mathbb{R}} \left\{ \frac{\partial}{\partial x_j}\Big|_p, \frac{\partial}{\partial y_j}\Big|_p \right\}$ and the tensor product $T_p\mathbb{R}^{2n} \otimes \mathbb{C} = span_{\mathbb{C}} \left\{ \frac{\partial}{\partial x_j}\Big|_p, \frac{\partial}{\partial y_j}\Big|_p \right\}$. To generalize complex derivatives, we introduce the Wirtinger derivatives are defined as $\frac{\partial}{\partial z_j} = \frac{\partial}{\partial x_j} - i\frac{\partial}{\partial y_j}$ and $\frac{\partial}{\partial \bar{z}_j} = \frac{\partial}{\partial x_j} + i\frac{\partial}{\partial y_j}$. These derivatives extend the notion of differentiation in the complex

plane. It is worth noting that a $\mathbb{R}$-differentiable function $f : \mathbb{C}^n \to \mathbb{C}$ is $\mathbb{C}$-differentiable if and only if $\frac{\partial f}{\partial \bar{z}} = 0$. As an example, the conjugate $\bar{z}$ does not satisfy the Cauchy-Riemann equation, which means that $\frac{d\bar{z}}{dz}$ cannot be defined. However, we can still compute $\frac{\partial \bar{z}}{\partial z}$, which evaluates to 0. Denote $\nabla_z := \left(\frac{\partial}{\partial z_1}, \ldots, \frac{\partial}{\partial z_n}\right)$ and $\nabla_{\bar{z}} := \left(\frac{\partial}{\partial \bar{z}_1}, \ldots, \frac{\partial}{\partial \bar{z}_n}\right)$.

Proposition 1 $\frac{\partial}{\partial \bar{z}}\|z\|_{1,M} = \frac{1}{2}sign(\Re(z)) + \frac{i}{2}sign(\Im(z))$. *Here* $\Re(z)$ *is the real part of* z *and* $\Im(z)$ *is the imaginary part of* z.

Proof Let $z = x + iy$. $\frac{\partial}{\partial \bar{z}}\|z\|_{1,M} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)(|x| + |y|) = \frac{1}{2}\big(sign(x) + isign(y)\big)$. □

Proposition 2 $\nabla_{\overline{\widehat{\mathbf{H}}}}\|\widehat{\mathbf{H}}\|_{1,M} = sign(\Re(\widehat{\mathbf{H}}))$.

Proof To prove this proposition, we will consider componentwise derivatives. We have two cases of $k = 0, k = \frac{T}{2}$, and $1 \le k \le T-1$. This distinction is important because the 0-th and $\frac{T}{2}$-th Fourier coefficients of real time-series data should be real numbers.

For each entry of $\widehat{\mathbf{H}}_{sk}$,

1. if $k = 0$ or $k = \frac{T}{2}$:
   $\frac{\partial}{\partial \overline{\widehat{\mathbf{H}}}_{sk}}\|\widehat{\mathbf{H}}_{sk}\|_{1,M} = \frac{\partial}{\partial \widehat{\mathbf{H}}_{sk}}\|\widehat{\mathbf{H}}_{sk}\|_{1,M} = sign(\Re(\widehat{\mathbf{H}}_{sk}))$, since $\widehat{\mathbf{H}}_{sk} \in \mathbb{R}$.
2. if $1 \le k \le T-1$ and $k \ne \frac{T}{2}$:
   $\frac{\partial}{\partial \overline{\widehat{\mathbf{H}}}_{sk}}\|\widehat{\mathbf{H}}\|_{1,M} = \frac{\partial}{\partial \overline{\widehat{\mathbf{H}}}_{sk}}\left(\|\widehat{\mathbf{H}}_{sk}\|_{1,M} + \|\widehat{\mathbf{H}}_{s,T-k}\|_{1,M}\right)$
   ∵) The other entries of $\widehat{\mathbf{H}}$ are independent of $\overline{\widehat{\mathbf{H}}}_{sk}$.

$$
\begin{aligned}
&\overset{(a)}{=} \frac{1}{2}sign(\Re(\widehat{\mathbf{H}}_{sk})) - \frac{i}{2}sign(\Im(\widehat{\mathbf{H}}_{sk})) + \frac{1}{2}sign(\Re(\widehat{\mathbf{H}}_{s,T-k})) - \frac{i}{2}sign(\Im(\widehat{\mathbf{H}}_{s,T-k}))\\
&\overset{(b)}{=} \frac{1}{2}sign(\Re(\widehat{\mathbf{H}}_{sk})) - \frac{i}{2}sign(\Im(\widehat{\mathbf{H}}_{sk})) + \frac{1}{2}sign(\Re(\widehat{\mathbf{H}}_{sk})) + \frac{i}{2}sign(\Im(\widehat{\mathbf{H}}_{sk}))\\
&= sign(\Re(\widehat{\mathbf{H}}_{sk})),
\end{aligned}
$$

where (a) follows from Proposition 1 and (b) uses the fact that $\Re(\widehat{\mathbf{H}}_{s,T-k}) = \Re(\widehat{\mathbf{H}}_{sk})$ and $\Im(\widehat{\mathbf{H}}_{s,T-k}) = -\Im(\widehat{\mathbf{H}}_{sk})$. □

Proposition 3 $\frac{2}{T}\Re(sign(\Re(\widehat{\mathbf{H}}))\mathcal{F}_T^{-1})$ *is a subgradient of* $\mathfrak{H} := \mathfrak{G} \circ (\mathfrak{F}, \overline{\mathfrak{F}})$, *where* $\mathfrak{F}(\mathbf{H}) = \widehat{\mathbf{H}}, \overline{\mathfrak{F}}(\mathbf{H}) = \overline{\widehat{\mathbf{H}}}$ *and* $\mathfrak{G}(\widehat{\mathbf{H}}, \overline{\widehat{\mathbf{H}}}) = \|\widehat{\mathbf{H}}\|_{1,M}$. *i.e.,* $\mathfrak{H}(\mathbf{H}) = \|\widehat{\mathbf{H}}\|_{1,M}$. *We can absorb the coefficient* $\frac{2}{T}$ *into the step size in projected subgradient descent in Lemma 2.*

Proof Note that $\Delta\mathfrak{G}(\widehat{\mathbf{H}}, \overline{\widehat{\mathbf{H}}}) = \left\langle \Delta\widehat{\mathbf{H}}, \nabla_{\widehat{\mathbf{H}}}\mathfrak{G}\right\rangle + \left\langle \Delta\overline{\widehat{\mathbf{H}}}, \nabla_{\overline{\widehat{\mathbf{H}}}}\mathfrak{G}\right\rangle = 2\Re\left(\left\langle \Delta\widehat{\mathbf{H}}, \nabla_{\overline{\widehat{\mathbf{H}}}}\mathfrak{G}\right\rangle\right)$ [14].

$$\Delta\mathfrak{H}(\mathbf{H}) = 2\Re\left(\left\langle \Delta\widehat{\mathbf{H}}, \nabla_{\overline{\widehat{\mathbf{H}}}}\mathfrak{G}\right\rangle\right)$$

Since $\mathfrak{F}$ is linear,

$$\implies \Delta\mathfrak{H}(\mathbf{H}) = 2\Re\left(\left\langle \mathfrak{F}\Delta\mathbf{H}, \nabla_{\overline{\widehat{\mathbf{H}}}}\mathfrak{G}\right\rangle\right).$$

Since $\mathcal{F}_T^*\mathcal{F}_T = \frac{1}{T}\mathbf{I}$, where $\mathcal{F}_T^*$ is a Hermitian matrix of $\mathcal{F}_T$,

$$\implies \Delta\mathfrak{H}(\mathbf{H}) = 2\Re\left(\left\langle \Delta\mathbf{H}, \frac{1}{T}\mathfrak{F}^{-1}\nabla_{\overline{\widehat{\mathbf{H}}}}\mathfrak{G}\right\rangle\right).$$

$$\implies \Delta\mathfrak{H}(\mathbf{H}) = \Re\left(\left\langle \Delta\mathbf{H}, \frac{2}{T}(\nabla_{\widehat{\mathbf{H}}}\mathfrak{G})\mathcal{F}_T^{-1}\right\rangle\right)$$
$$\implies \Delta\mathfrak{H}(\mathbf{H}) = \left\langle \Delta\mathbf{H}, \frac{2}{T}\Re\left((\nabla_{\widehat{\mathbf{H}}}\mathfrak{G})\mathcal{F}_T^{-1}\right)\right\rangle$$

□

Definition 3 (*Definition 1.4*, [2]) Let $\mathcal{X}$ be a nonempty set. $f : \mathcal{X} \to [-\infty, \infty]$ is proper if $-\infty \notin f(\mathcal{X})$ and $domf = \{x \in \mathcal{X} : f(x) < \infty\} \neq \emptyset$.

Lemma 1 (Convergence of BCD, Lemma 3.1, Theorem 4.1, [22]) *Let* $f(\theta^1, \ldots, \theta^N) = f_0(\theta^1, \ldots, \theta^N) + \sum_{n=1}^{N} f_n(\theta^n)$. *Suppose that* $f_0, f_1, \ldots, f_N$ *satisfy the followings*

1. *The* $domf_0$ *is open and* f_0 *has a directional derivative in all direction on* $domf_0$.
2. *The set* $X^0 = \left\{(\theta^1, \ldots, \theta^N) : f(\theta^1, \ldots, \theta^N) \leq f(\theta_0^1, \ldots, \theta_0^N)\right\}$ *is compact, and* f *is continuous on* X^0.
3. *The mapping* $\theta^n \mapsto f(\theta^1, \ldots, \theta^N)$ *has at most one minimum for* $n \geq 2$.

Then, for $\left\{(\theta_k^1, \ldots, \theta_k^N)\right\}_{k\in\mathbb{N}}$, *every cluster point is a stationary point of* f.

Lemma 2 (Projected subgradient method, [4]) *Consider the following constraint convex optimization problem*

$$\underset{x\in\mathcal{C}}{arg\,min}\, f(x) \tag{38}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ *is convex and* $\mathcal{C} \subseteq \mathbb{R}^n$ *is convex. The projected subgradient method is given by* $x_{k+1} = P(x_k - \alpha_k g_{(k)})$, *where* P *is the projection on* $\mathcal{C}$ *and* α_k *is the* k*th step size and* $g_{(k)}$ *is a subgradient of* f *at* x_k. *Then* x_k *converges to the solution of (38).*

Note that Problem (7) is a nonconvex problem. Therefore, finding the global minimum of this problem is extremely difficult. Alternatively, to tackle this issue, we aim to find the stationary points. The Subproblems (13), (14) and (15) are all convex optimization problems. We will particularly focus on examining (13). Since $\mathbf{H} \mapsto \|\widehat{\mathbf{H}}\|_{1,M}$ is not differentiable, we use projected 'sub'gradient descent algorithm to solve (13). Proposition 3 will be used to solve it.

Proof of Theorem 1 (BCD convergence) In Lemma 1, set $\theta^1 = \mathbf{H}, \theta^2 = \mathbf{W}, \theta^3 = \mathbf{W}'$ and $f_0(\mathbf{H}, \mathbf{W}, \mathbf{W}') = \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2$, $f_1(\mathbf{H}) = \lambda\|\mathbf{H}\mathcal{F}_T\|_{1,M} + \iota_{\mathcal{G}}(\mathbf{H})$, $f_2(\mathbf{W}) = \lambda_1\|\mathbf{W}\|_F^2$ and $f_3(\mathbf{W}') = \lambda_2\|\mathbf{W}'\|_F^2$, where $\xi, \lambda, \lambda_1, \lambda_2 > 0$, $\mathcal{G} = \left\{\mathbf{H} \in \mathbb{R}^{r\times T} : \mathbf{H} \geq 0\right\}$ and $\iota_{\mathcal{G}}$ is an indicator function. Here, we introduce L_2-regularization terms for $\mathbf{W}$ and $\mathbf{W}'$ to satisfy conditions 2 and 3 in Lemma 1. Note that λ_1 and λ_2 can be selected as arbitrarily small positive numbers; thus, in Algorithm 1, we can disregard the L_2-regularization term.

Now, we check the conditions in the Lemma 1.

1. Since $domf_0 = \mathbb{R}^{r\times T} \times \mathbb{R}^{d\times r} \times \mathbb{R}^{d\times r}$ is whole space, it is open. Clearly, f_0 is continuously differentiable, it has a directional derivative in all directions.
2. Suppose we choose an initial point $\mathbf{H}_0 \in \mathcal{G}$. Then $M := f(\mathbf{H}_0, \mathbf{W}_0, \mathbf{W}_0')$ has a finite value. The set $X^0 = \left\{(\mathbf{H}, \mathbf{W}, \mathbf{W}') : f(\mathbf{H}, \mathbf{W}, \mathbf{W}') \leq M\right\} = f^{-1}([0, M])(\because f \geq 0)$. Note that f on X^0 is $\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2 + \lambda\|\mathbf{H}\mathcal{F}_T\|_{1,M} + \lambda_1\|\mathbf{W}\|_F^2 + \lambda_2\|\mathbf{W}'\|_F^2$. Since $\|\mathbf{H}\mathcal{F}_T\|_{1,M}$, $\|\mathbf{W}\|_F$ and $\|\mathbf{W}'\|_F$ are bounded, the set X^0 becomes bounded. Clearly, f is continuous on X^0, this implies that $X^0 = f^{-1}([0, M])$ is closed; therefore, it is compact.

3. Note that if the Hessian of f is positive definite, then f is strictly convex [9]. We can calculate the following (see [16])

$$\frac{\partial f}{\partial \mathbf{W}} = 2(\mathbf{W}\mathbf{H}\mathbf{H}^T - \mathbf{X}\mathbf{H}^T) + 2\lambda_1 \mathbf{W}, \frac{\partial^2 f}{\partial \mathbf{W}^2} = 2\mathbf{H}\mathbf{H}^T + 2\lambda_1 \mathbf{I}, \tag{39}$$

$$\frac{\partial f}{\partial \mathbf{W}'} = 2\xi(\mathbf{W}'\mathbf{H}\mathbf{H}^T - \mathbf{X}\mathbf{H}^T) + 2\lambda_2 \mathbf{W}', \frac{\partial^2 f}{\partial \mathbf{W}'^2} = 2\xi\mathbf{H}\mathbf{H}^T + 2\lambda_2 \mathbf{I}, \tag{40}$$

where $\mathbf{I}$ is an identity matrix. Since $\lambda_1, \lambda_2 > 0$, $\frac{\partial^2 f}{\partial \mathbf{W}^2}$ and $\frac{\partial^2 f}{\partial \mathbf{W}'^2}$ are positive definite. Therefore, $\mathbf{W} \mapsto f(\mathbf{H}, \mathbf{W}, \mathbf{W}')$ and $\mathbf{W}' \mapsto f(\mathbf{H}, \mathbf{W}, \mathbf{W}')$ are strictly convex and have at most one minimum.

(Solving subproblem via BCD): The Fourier transform $\mathbf{H} \mapsto \widehat{\mathbf{H}}$ is linear and $\|\cdot\|_{1,M}$ is a convex function. Consequently, the mapping $\mathbf{H} \mapsto \|\widehat{\mathbf{H}}\|_{1,M} = \|\mathbf{H}\mathcal{F}_T\|_{1,M}$ is also convex and it satisfies all the hypotheses of Lemma 2. We have computed its subgradient in Proposition 3. Therefore, we can solve the subproblem using the projected subgradient descent:

$$\underset{\mathbf{H}\in\mathbb{R}^{r\times T}_{\geq 0}}{\arg\min} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2 + \lambda\|\mathbf{H}\mathcal{F}_T\|_{1,M}. \tag{41}$$

The other subproblems:

$$\underset{\mathbf{W}\in\mathbb{R}^{d\times T}}{\arg\min} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2 + \lambda\|\mathbf{H}\mathcal{F}_T\|_{1,M} \tag{42}$$

and

$$\underset{\mathbf{W}'\in\mathbb{R}^{d\times T}}{\arg\min} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2 + \lambda\|\mathbf{H}\mathcal{F}_T\|_{1,M} \tag{43}$$

are usual normal equation and its solution is well-known.

Therefore $\mathbf{W}_j$, $\mathbf{W}'_j$ and $\mathbf{H}_j$ converge to stationary points in Algorithm 1. □

Remark 3 To solve the subproblem (13), we apply the following gradient-projection steps.

$$\text{(Gradient step)} \quad \mathbf{H}_j \leftarrow \mathbf{H}_j - \alpha_j \nabla f(\mathbf{H}_j), \text{ where} \tag{44}$$

$$\begin{aligned} f(\mathbf{H}_j) = &(\mathbf{W}_{j-1}^T\mathbf{W}_{j-1}\mathbf{H}_{j-1} - \mathbf{W}_{j-1}^T\mathbf{X}) \\ &+ \xi(\mathbf{W}'^T_{j-1}\mathbf{W}'_{j-1}\mathbf{H}_{j-1} - \mathbf{W}'^T_{j-1}\mathbf{Y}) \\ &+ \lambda\Re(sign(\Re(\widehat{\mathbf{H}}))\mathcal{F}_T^{-1}) \end{aligned} \tag{45}$$

$$\text{(Projection step)} \quad \mathbf{H}_j \leftarrow \max\left\{0, \mathbf{H}_j - \alpha_j \nabla f(\mathbf{H}_j)\right\}. \tag{46}$$

This can be interpreted from a frequency domain perspective.

$$\widehat{\mathbf{H}}_j \leftarrow \widehat{\mathbf{H}}_j - \alpha_j \nabla f(\widehat{\mathbf{H}}_j), \text{ where} \tag{47}$$

$$\begin{aligned} f(\widehat{\mathbf{H}}_j) = &(\mathbf{W}_{j-1}^T\mathbf{W}_{j-1}\widehat{\mathbf{H}}_{j-1} - \mathbf{W}_{j-1}^T\widehat{\mathbf{X}}) \\ &+ \xi(\mathbf{W}'^T_{j-1}\mathbf{W}'_{j-1}\widehat{\mathbf{H}}_{j-1} - \mathbf{W}'^T_{j-1}\widehat{\mathbf{Y}}) \\ &+ \lambda\Re\left(sign(\Re(\widehat{\mathbf{H}}))\right), \end{aligned} \tag{48}$$

where Proposition 2 is used in (48). By applying the inverse Fourier transform to update rule (47), we obtain the update rule in the time domain.

Now, consider Theorem 2.

Proposition 4 (Corollary 1 and Theorem 4, [23]) *If $f : \mathbb{R}^{r\times T} \to \mathbb{R}$ and $g = \iota_{\mathcal{G}}, h_R = \iota_{\mathcal{H}} : \mathbb{R}^{r\times T} \to \mathbb{R}\cup\{\infty\}$ are proper, lower semi-continuous and convex, where $\mathcal{G}$ and $\mathcal{H} \subseteq \mathbb{R}^n$ are convex and $\iota_{\mathcal{G}}, \iota_{\mathcal{H}}$ are indicator functions. Let $\overline{\mathbf{H}}_N$ be output of Algorithm 2, then*

$$f(\overline{\mathbf{H}}_N) - \min f = \tilde{\mathcal{O}}\left(\frac{1}{N}\right), \tag{49}$$

$$\operatorname{dist}(\overline{\mathbf{H}}_N, \mathcal{H}) = \tilde{\mathcal{O}}\left(\frac{1}{N}\right). \tag{50}$$

We will show that the matrix factorization with two constraints, i.e. 1) the nonnegativity constraint of $\mathbf{H}$ and 2) the constraint that $\mathbf{H}$ has no specific frequencies and also satisfies the hypothesis of Proposition 4.

Lemma 3 (Example 1.25, [2]) *Let $\mathcal{X}$ be a Hausdorff space. The indicator function of a set $C \subset \mathcal{X}$, i.e. $\iota_C : \mathcal{X} \to [-\infty, \infty] : x \mapsto \begin{cases} 0, & \text{if } x \in C \\ \infty, & \text{Otherwise} \end{cases}$ is lower semi-continuous if and only if C is closed.*

Proof of Theorem 2 This proof is similar to the proof of Theorem 1. (BCD convergence) Let $\mathcal{G} = \left\{\mathbf{H} \in \mathbb{R}^{r\times T} : \mathbf{H} \geq 0\right\}$ and $\mathcal{H} = \mathcal{F}_T^{-1}(\mathbb{C}^{r\times R} \times \{0\}^{T-R})$. In Lemma 1, set $\theta^1 = \mathbf{H}, \theta^2 = \mathbf{W}, \theta^3 = \mathbf{W}'$ and $f_0(\mathbf{H}, \mathbf{W}, \mathbf{W}') = \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2$, $f_1(\mathbf{H}) = \lambda\|\mathbf{H}\|_F^2 + \iota_{\mathcal{G}\bigcap\mathcal{H}}(\mathbf{H})$, $f_2(\mathbf{W}) = \lambda_1\|\mathbf{W}\|_F^2$ and $f_3(\mathbf{W}') = \lambda_2\|\mathbf{W}'\|_F^2$, where $\xi, \lambda, \lambda_1, \lambda_2 > 0$. If we choose an initial point $\mathbf{H}_0 \in \mathcal{G}\bigcap\mathcal{H}$, then f on X^0 is $\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \xi\|\mathbf{Y} - \mathbf{W}'\mathbf{H}\|_F^2 + \lambda\|\mathbf{H}\|_F^2 + \lambda_1\|\mathbf{W}\|_F^2 + \lambda_2\|\mathbf{W}'\|_F^2$. The remainder of the proof is similar to that of Theorem 1.

(How to solve subproblem in BCD) Set f, g and h_R defined in (20)–(22). Since $\mathcal{G}$ and $\mathcal{H}$ are closed convex sets in $\mathbb{R}^{r\times T}$, by Lemma 3, g and h_R are lower semi-continuous. Since the indicator function on convex set is convex map, g and h_R are convex maps. Clearly, g and h_R are proper maps. Therefore f, g and h_R satisfy the all hypothesis of Proposition 4. Therefore $\mathbf{W}_j$, $\mathbf{W}'_j$ and $\mathbf{H}_j$ converge to stationary points in Algorithm 3. □

5 Analysis on Synthetic Data

The following proposition states that the presence of a mismatch term between $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ can disrupt the inference of the periodic pattern of $\mathbf{X}$. Here S is the number of auxiliary data.

Proposition 5 *Let $\overline{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\xi}\mathbf{Y} \end{bmatrix}$ and $\overline{\mathbf{W}} = \begin{bmatrix} \mathbf{W} \\ \sqrt{\xi}\mathbf{W}' \end{bmatrix}$. If we let $\mathbf{X}_{mn} = \sum_{l=0}^{T-1} \widehat{\mathbf{X}}_{ml} e^{\frac{2\pi i l n}{T}}$ and $\mathbf{Y}_{mn} = \sum_{l=0}^{T-1} \widehat{\mathbf{Y}}_{ml} e^{\frac{2\pi i l n}{T}}$, then the solution of $\min_{\mathbf{H}} \|\overline{\mathbf{X}} - \overline{\mathbf{W}}\mathbf{H}\|_F^2$ is given by the following.*

$$\begin{aligned} \mathbf{H}_{kj} = &\sum_{s=0}^{d-1} \left[\left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T \right)_{ks} + \sum_{p=0}^{S-1} \sqrt{\xi} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T \right)_{k,s+pd} \right] \mathbf{X}_{sj} \\ &+ \sum_{s=0}^{d-1}\sum_{p=0}^{S-1}\sum_{l=0}^{T-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T \right)_{k,s+pd} \sqrt{\xi} \left((\widehat{\mathbf{Y}}_p)_{sl} - \widehat{\mathbf{X}}_{sl} \right) e^{\frac{2\pi i l j}{T}}. \end{aligned} \tag{51}$$

Proof The solution to $\min_{\mathbf{H}} \|\overline{\mathbf{X}} - \overline{\mathbf{W}}\mathbf{H}\|_F^2$ is well-known as the normal equation, and its solution is $\mathbf{H} = (\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\overline{\mathbf{X}}$.

$$\begin{aligned}
\mathbf{H}_{kj} &= \sum_{s=0}^{d+dS-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{ks} \overline{\mathbf{X}}_{sj} \\
&= \sum_{s=0}^{d-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{ks} \mathbf{X}_{sj} + \sum_{s=0}^{dS-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+d} \sqrt{\xi}\mathbf{Y}_{sj} \\
&= \sum_{s=0}^{d-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{ks} \mathbf{X}_{sj} + \sum_{s=0}^{d-1}\sum_{p=0}^{S-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd} \sqrt{\xi}(\mathbf{Y}_p)_{sj}.
\end{aligned}$$

Using the relations $\mathbf{X}_{sj} = \sum_{l=0}^{T-1} \widehat{\mathbf{X}}_{sl} e^{\frac{2\pi i l j}{T}}$ and $(\mathbf{Y}_p)_{sj} = \sum_{l=0}^{T-1} (\widehat{\mathbf{Y}_p})_{sl} e^{\frac{2\pi i l j}{T}}$, we have

$$\begin{aligned}
\mathbf{H}_{kj} &= \sum_{s=0}^{d-1}\sum_{l=0}^{T-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{ks} \widehat{\mathbf{X}}_{sl} e^{\frac{2\pi i l j}{T}} \\
&\quad + \sum_{s=0}^{d-1}\sum_{p=0}^{S-1}\sum_{l=0}^{T-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd} \sqrt{\xi}(\widehat{\mathbf{Y}_p})_{sl} e^{\frac{2\pi i l j}{T}}.
\end{aligned}$$

To see the disparity term between $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$, we rearrange the above as the following,

$$\begin{aligned}
\mathbf{H}_{kj} &= \sum_{s=0}^{d-1}\sum_{l=0}^{T-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{ks} \widehat{\mathbf{X}}_{sl} e^{\frac{2\pi i l j}{T}} \\
&\quad + \sum_{s=0}^{d-1}\sum_{p=0}^{S-1}\sum_{l=0}^{T-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd} \sqrt{\xi}\widehat{\mathbf{X}}_{sl} e^{\frac{2\pi i l j}{T}} \\
&\quad + \sum_{s=0}^{d-1}\sum_{p=0}^{S-1}\sum_{l=0}^{T-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd} \sqrt{\xi}\left((\widehat{\mathbf{Y}_p})_{sl} - \widehat{\mathbf{X}}_{sl}\right) e^{\frac{2\pi i l j}{T}} \\
&= \sum_{s=0}^{d-1}\left[\left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{ks} + \sum_{p=0}^{S-1}\sqrt{\xi}\left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd}\right] \\
&\quad \times \left(\sum_{l=0}^{T-1} \widehat{\mathbf{X}}_{sl} e^{\frac{2\pi i l j}{T}}\right) \\
&\quad + \sum_{s=0}^{d-1}\sum_{p=0}^{S-1}\sum_{l=0}^{T-1} \left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd} \sqrt{\xi}\left((\widehat{\mathbf{Y}_p})_{sl} - \widehat{\mathbf{X}}_{sl}\right) e^{\frac{2\pi i l j}{T}}.
\end{aligned}$$

Now the inverse Fourier transform yields

$$\mathbf{H}_{kj} = \sum_{s=0}^{d-1}\left[\left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{ks} + \sum_{p=0}^{S-1}\sqrt{\xi}\left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd}\right]\mathbf{X}_{sj}$$

$$+\sum_{s=0}^{d-1}\sum_{p=0}^{S-1}\sum_{l=0}^{T-1}\left((\overline{\mathbf{W}}^T\overline{\mathbf{W}})^{-1}\overline{\mathbf{W}}^T\right)_{k,s+pd}\sqrt{\xi}\left((\widehat{\mathbf{Y}_p})_{sl}-\widehat{\mathbf{X}}_{sl}\right)e^{\frac{2\pi ilj}{T}}.$$

□

Example 2 Set $\mathbf{X}_{ij} = \mathbf{X}_{i0}\cdot\left[\cos\left(\frac{2\pi\cdot 14j}{163}\right)+\cos\left(\frac{2\pi\cdot 6j}{163}\right)\right]+\epsilon(0,1)$, $(\mathbf{Y}_0)_{ij} = (\mathbf{Y}_0)_{i0}\times\cos\left(\frac{2\pi\cdot 14j}{163}\right)+\epsilon(0,\sigma)$ and $(\mathbf{Y}_1)_{ij} = (\mathbf{Y}_1)_{i0}\cos\left(\frac{2\pi\cdot 6j}{163}\right)+\epsilon(0,\sigma)$ where $\sigma\in\{0,1,2,3,4,5\}$ and $\epsilon(m,\sigma)$ is a Gaussian noise with the mean m and the standard deviation σ. We apply the matrix factorization based on the following three cases i) with the constraint on $\mathbf{H}$, ii) without any constraint on $\mathbf{H}$, iii) with the Principal Component Analysis(PCA). Figure 2 shows the spatial (top rows) and temporal (middle rows) patterns and the corresponding spectrums (bottom rows). The left, middle, and right columns correspond to the matrix factorization method with the nonnegativity constraint enforced, without the nonnegativity constraint enforced, and with the PCA approach, respectively. In this example, we use synthetic data for $\mathbf{X}$ that has two temporal patterns with frequencies of 6 and 14. As shown in Fig. 2, when the nonnegativity constraint is enforced to $\mathbf{H}$, we obtain clear spatial patterns, and each row of $\mathbf{H}$ captures different temporal patterns. Here note that the result may depend on the initial values assigned to $\mathbf{H}$ in BCD algorithm. On the other hand, when we do not impose the nonnegative constraint on $\mathbf{H}$ (or when using PCA), their temporal patterns cannot separate those two temporal frequency patterns, 6 and 14. In the time domain, we can observe that $\mathbf{H}$ has nonnegative values in the first column.

Example 3 (Comparison of Frobenius norm, L1-norm, and Minkowski 1-norm using projected (sub)gradient descent) In this example, we compare the behaviors of $\mathbf{H}$ in both the time and frequency domains when applying projected (sub)gradient descent method with the Frobenius norm, L1-norm, and Minkowski 1-norm of $\mathbf{H}$. Consider a time-series data $\mathbf{H}_0[k] = \cos(10\pi\cdot\frac{k}{30})+\epsilon_k$, where $\epsilon_k\in[0,1]$ is uniform noise for $k=0,1,\ldots,29$. In order to minimize the Frobenius, L1-, and Minkowski 1-norms, $i.e.\|\mathbf{H}\|_F^2$, $\|\mathbf{H}\|_1$, and $\|\mathbf{H}\|_{1,M}$, respectively, we apply the update rules with the projected subgradient descent method to each norm as follows:

Frobenius norm: $\mathbf{H}_{j+\frac{1}{2}} = \mathbf{H}_j - \alpha_j\mathbf{H}_j$ and $\mathbf{H}_{j+1} = \max\{0, \mathbf{H}_{j+\frac{1}{2}}\}$.

L1-norm: $\mathbf{H}_{j+\frac{1}{2}} = \mathbf{H}_j - \beta_j(sign(\mathbf{H}_j))$ and $\mathbf{H}_{j+1} = \max\{0, \mathbf{H}_{j+\frac{1}{2}}\}$.

Minkowski 1-norm: $\mathbf{H}_{j+\frac{1}{2}} = \mathbf{H}_j - \gamma_j(sign(\Re(\hat{\mathbf{H}}_j))\mathcal{F}_T^{-1})$ and $\mathbf{H}_{j+1} = \max\left\{0, \mathbf{H}_{j+\frac{1}{2}}\right\}$.

In this experiment, we set $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 0.5$. For a fair comparison, we iterate the projected (sub)gradient method until $\|\mathbf{H}\|_F < 3$. Figure 3 displays $\mathbf{H}$ after projected (sub)gradient descent, with the Frobenius norm (blue), the L1-norm (green), and the Minkowski 1-norm (red). The left figure in Fig. 3 shows the data in the time domain and the right figure in the frequency domain. In this example, the solution is simply $\mathbf{H} = 0$. We want to observe how $\mathbf{H}_j$ approaches 0 during the gradient descent process. Note that the iteration goes until $\|\mathbf{H}\|_F < 3$ and according to Parseval's identity, $\|\widehat{\mathbf{H}}\|_F$ are same for three cases. The difference among these three cases in the time and frequency domains is the amplitude distribution. In the case of the L1-norm, as expected, it exhibits significant sparsity in the time domain. Although the Minkowski 1-norm does not achieve the desired sparsity in the frequency domain, it shows a tendency to yield smaller amplitudes of frequencies compared to other norms. In this experiment, we expect the 5th frequency to be dominant, while the other frequencies are considered noise originating from uniform noise. The Minkowski 1-norm shows an effect of increasing the amplitude of the 0th frequency while reducing the

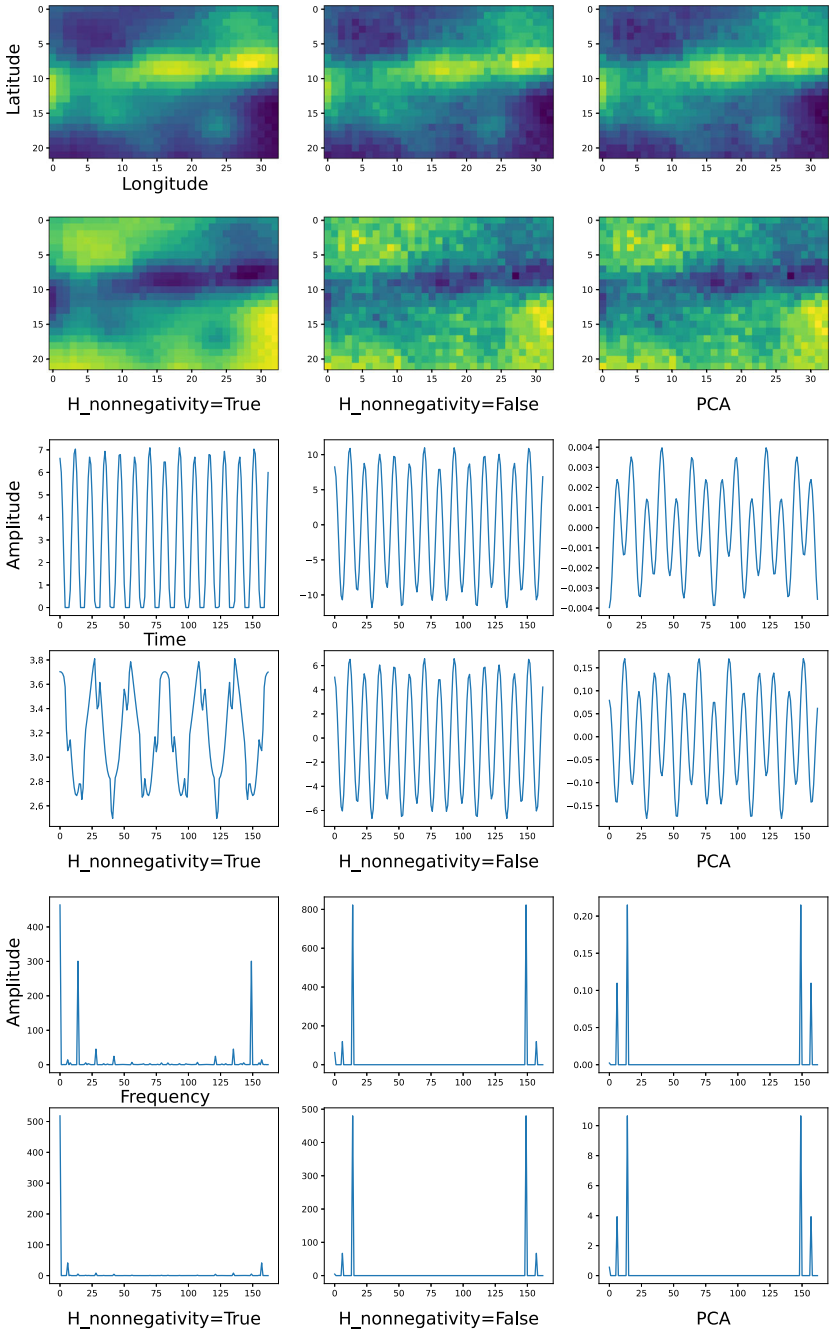

Fig. 2 Top: spatial patterns of $\mathbf{X}$ in Example 2. First(second) row shows the first(second) column of $\mathbf{W}$. Middle: temporal patterns of $\mathbf{H}$. First(second) row is the first(second) row of $\mathbf{H}$. Bottom: spectrums of temporal patterns of $\mathbf{H}$. First(second) row is the Fourier transform of the first(second) row of $\mathbf{H}$. The left column shows the results when the nonnegativity constraint is enforced to $\mathbf{H}$, the middle column without the nonnegativity constraint and the right column with PCA

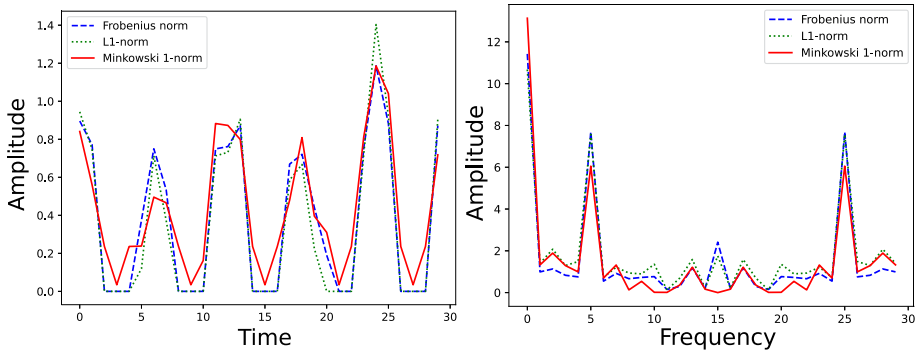

Fig. 3 Behaviors under the projected (sub)gradient descent in Example 3. $\mathbf{H}$ with the Frobenius norm (blue), the L1-norm (green), and the Minkowski 1-norm (red). Left: $\mathbf{H}$ in the time domain. Right: $|\widehat{\mathbf{H}}|$ in the frequency domain

amplitudes of undominated (noise) frequencies. This effect remains consistent even when experimenting with different random seeds for uniform noise. The experimental results show that the Minkowski 1-norm reduces the amplitude of the undominated frequencies more effectively than the Frobenius and L1-norms. That is, the soft frequency regularization achieves our goal partially as stated in Sect. 3.2.

6 Experimental Results

6.1 GRACE Data

Gravity Recovery and Climate Experiment (GRACE) satellites measure the change of the gravitational field on Earth. From this measurement, we can obtain the total water storage anomaly. Our dataset $\mathcal{X}$ represented as a 3-tensor, and each entry value $\mathcal{X}[m, n, t]$ is the variation in total water storage compared to the temporal average during Jan. 2004 - Dec. 2009 at a latitude of m, longitude of n, and time (months) of t.[2] The measurement unit is centimeters (cm). By considering variations in total water storage, we can remove static mass and infer the redistribution of water. Matrix factorization is one method used to extract latent patterns from the data. In our research, matrix factorization is used to separate spatio-temporal data into spatial latent figures and temporal latent figures. This approach enables us to analyze temporal patterns. The nonnegative constraint helps us to obtain more interpretable spatial atoms. To analyze the temporal patterns, we consider a soft constraint in the frequency domain. For the reconstruction of the total water storage anomaly (TWSA),[3] we use auxiliary data that was also utilized in [21]: precipitation,[4] temperature[4], and Noah TWS.[5] Noah TWS refers to the traditional method of restoring TWSA before the GRACE era. We divide the dataset into the

[2] https://edo.jrc.ec.europa.eu/documents/factsheets/factsheet_grace_tws_anomaly.pdf.

[3] CSR GRACE/GRACE-FO RL06 Mascon Solutions, https://www2.csr.utexas.edu/grace/RL06_mascons.html [19, 20].

[4] ERA5 monthly averaged data on single levels, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview [12].

[5] GLDAS Noah Land Surface Model, https://disc.gsfc.nasa.gov/datasets/GLDAS_NOAH025_M_2.1/summary?keywords=gldas [3, 18].

training set (132 months, Apr. 2002–Jan. 2014) and the test set (31 months, Feb. 2014–Jun. 2017). The train set consists of the data for the 142 months, and the test set for the remaining 41 months. However, due to technical issues, there are missing data in both sets, occurring intermittently every 10 months (06/2002, 07/2002, 06/2003, 01/2011, 06/2011, 05/2012, 10/2012, 03/2013, 08/2013, 09/2013, 02/2014, 07/2014, 12/2014, 06/2015, 10/2015, 11/2015, 04/2016, 09/2016, 10/2016, 02/2017). These missing data have been excluded as described in [21]. We chose 17 river basin data to assess our proposed method, that is, the areas of Amazon, Congo, Ganges, Indus, Huanghe, Volga, Parana, Lena, Mackenzie, Ob, Yenisei, Nile, Yangtze, Indigirka, Zambezi, Mississippi, Murray.[6]

We focus on the temporal patterns present in spatio-temporal data. In our experiments with the GRACE data, we expect to find a spatial pattern that exhibits an annual cycle with a period of 12 months. In this study, we utilize the discrete Fourier transform on the time-series data obtained through SSNMF to uncover the temporal patterns in GRACE data. By preserving specific frequencies selectively in the time-series corresponding to each spatial basis, we can enhance the interpretability of the periodic patterns. To achieve this, we use the method proposed in the pervious section, i.e. the soft/hard constraints in the frequency domain and compare the results with the results by the Lasso/Ridge regression methods. Furthermore, we compare the methodology with the existing methods such as Deep Neural Network(DNN), Multiple Linear Regression(MLR), Seasonal ARIMA with eXogenous variables(SARIMAX) [6], and assess its forecasting performance, demonstrating comparable results.

6.2 Comparison Among Lasso, Soft and Hard Frequency Regularization

Definition 4 (*Inverse usage ratio of frequencies*) In Example 3, we focused on reducing the amplitude of undominated frequencies. To measure the decrease in amplitude for the frequencies that are not dominant, we define the inverse usage ratio of frequencies as follows:

$$\mu(\mathbf{H}) = \left[\left(\frac{|\widehat{\mathbf{H}}[s,k]|}{\sum_l |\widehat{\mathbf{H}}[s,l]|} \right)^{-1} \right]_{\substack{0\le s\le r-1,\\ 0\le k\le T-1}} \in \mathbb{R}^{r\times T}, \tag{52}$$

where $\mathbf{H} \in \mathbb{R}^{r\times T}$. A large value of $\mu(\mathbf{H})[s_0, k_0]$ indicates that the k_0th Fourier coefficient of the s_0th temporal pattern is small.

Example 4 For the comparison of the Lasso and soft frequency regularization methods we consider the Yangtze dataset. Figure 4 shows that in soft frequency regularization, $\mu(\mathbf{H})$ has larger values than Lasso. Similar to Example 3, we observe that soft frequency regularization reduces the usage of undominated (noise) frequencies more effectively than Lasso regularization, but it does not achieve the desired sparsity in the frequency domain.

Example 5 In this example, we compare soft and hard frequency regularization methods for the Yangtze dataset. It can be observed that the elements of $\mu(\mathbf{H})$ are higher with hard frequency regularization than soft frequency regularization. In this experiment, we can verify that hard frequency regularization is more effective than soft frequency regularization in

[6] The latitude and longitude information for the river basin was obtained from https://hydro.iis.u-tokyo.ac.jp/~taikan/TRIPDATA/Data/rivers.idx [15].

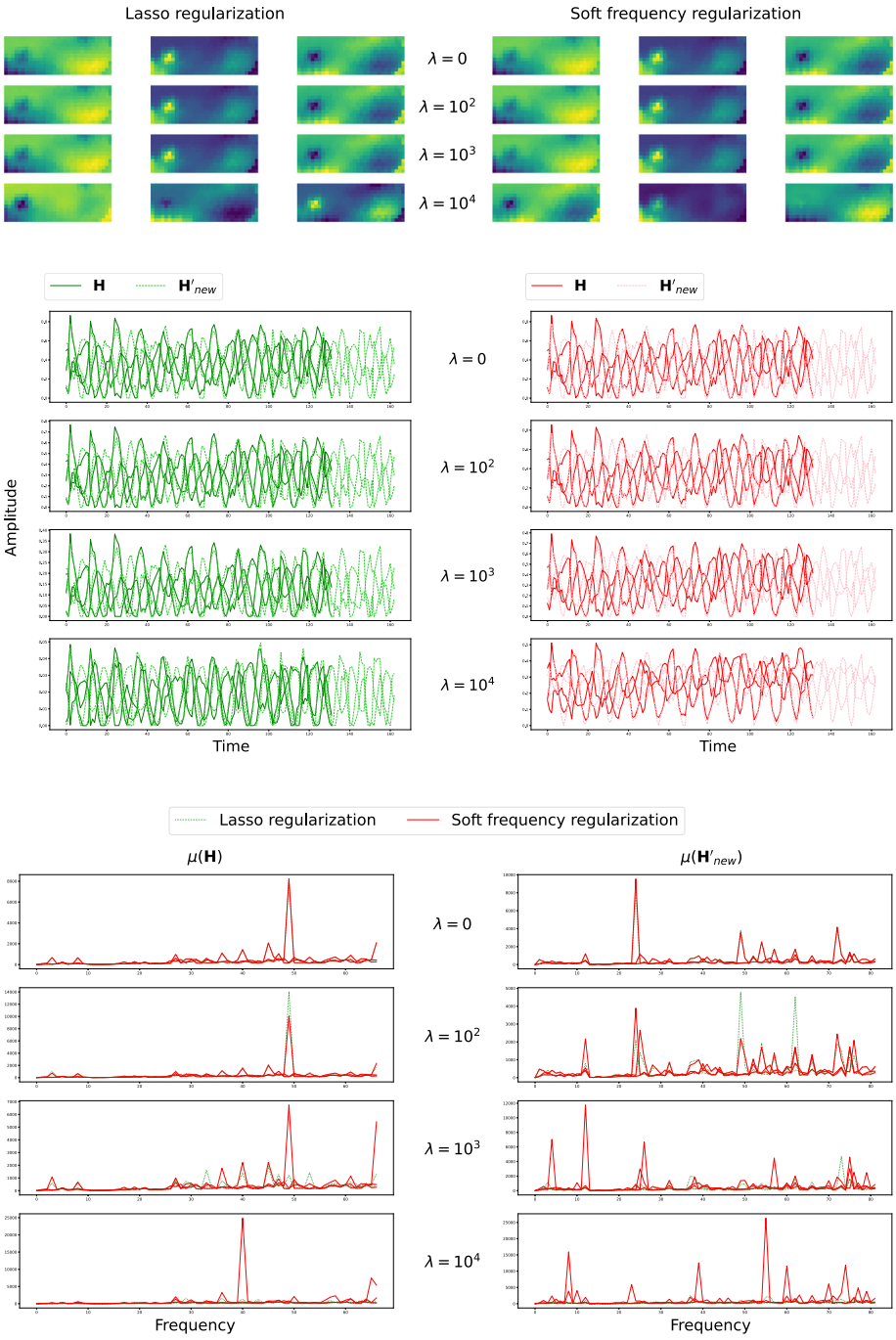

Fig. 4 Yangtze data: $r = 3$, $\xi = 100$. Each row corresponds to a choice of regularization parameters $\lambda \in \left\{0, 10^2, 10^3, 10^4\right\}$. First row: spatial pattern of $\mathbf{W}$, Second row: temporal pattern of $\mathbf{H}$. Third row: $\mu(\mathbf{H})$ and $\mu(\mathbf{H}'_{\text{new}})$, respectively. This shows only positive frequencies (frequencies up to half of the length of the time-series) since they have the same value as the negative frequencies

minimizing the utilization of particular frequencies. Figure 5 verifies these results. Note that the reason why there is no always value of $\mu(\mathbf{H})$ that diverges to ∞ with the hard constraint is explained in Algorithm 4, where we prioritize the nonnegativity condition over the removal of specific frequencies in $\mathbf{H}$.

6.3 Accuracy of Spatio-temporal Prediction

Now for the evaluation of the proposed method, consider the following metric.

Definition 5 (Nash–Sutcliffe efficiency (*NSE*), [21]) Nash–Sutcliffe efficiency (NSE) is defined as follows:

$$NSE = 1 - \frac{\sum_{i=0}^{T-1} \left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_{rec,i} \right)}{\sum_{i=0}^{T-1} \left(\overline{\mathbf{X}}_i - \langle \overline{\mathbf{X}} \rangle \right)},$$

where $\mathbf{X}_{rec}$ is the output of the model, $\overline{\mathbf{X}}$ is the spatial average of $\mathbf{X}$ and $\langle \cdot \rangle$ denotes the average in time. NSE ranges between $-\infty$ and 1. If NSE is close to 1, it is indicated that the model predicts perfectly.

We use the number of spatial-temporal patterns, denoted as r, as a hyperparameter, ranging from 2 to 20. Additionally, we chose additional parameters for $\xi \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5\}$, $\lambda \in \{10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$, and $R \in \{10, 20, 30, 40\}$. For each hyperparameter, we conduct 10 experiments and calculate the median value of the NSEs obtained from these 10 experiments. Table 1 summarizes the result.

In these experimental results, we can observe that the soft frequency, Lasso, and Ridge regularization methods generally demonstrate the best predictive performance among all, as shown in Table 1. As confirmed in Example 5, hard frequency regularization is most effective in removing undominated frequencies compared to other regularization mehtods, but it appears less effective in predictive performance. This experiment shows that we can select each form of regularization depending on whether our analysis primarily centers on the time domain or the frequency domain. Our choice can also depend on our priority for interpretability and predictive accuracy.

6.4 Atom Removal

The term 'atom' refers to a spatial latent pattern of the given data. Note that some atoms could represent noise. Therefore, by removing such noisy atoms learned during the coding process, we may achieve better results. This process is similar to PCA, where we select the principal components that capture the most variation in the data. However, the key difference between PCA and the proposed method is the utilization of supervised information to identify the appropriate atoms. If the NSE value significantly increases after removing a particular atom, that atom can be considered as noise. Given the possibilities for discarding multiple atoms, we only consider removing a single atom. All hyperparameter settings are the same as those in above. As shown in Tables 1, 2 and 3, our method is comparable to previous research in the geophysical research community. The sharp increase in NSE in the Ob region is noteworthy.

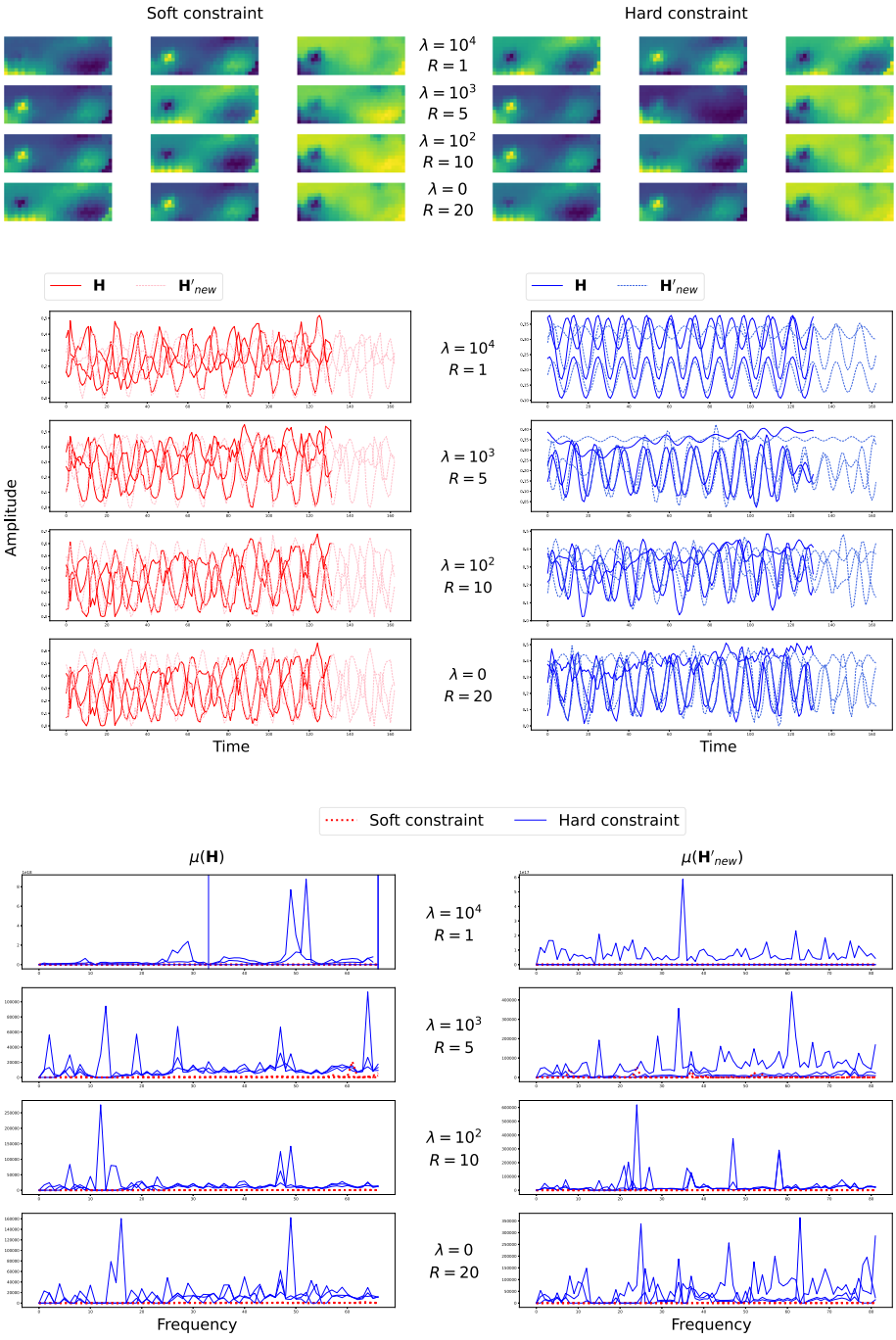

Fig. 5 Yangtze data, $r = 3$, $\xi = 100$. Each row corresponds to regularization parameters $\lambda \in \left\{0, 10^2, 10^3, 10^4\right\}$ and remaining frequencies, specifically $R \in \{1, 5, 10, 20\}$. First row: spatial pattern of $\mathbf{W}$, Second row: temporal pattern of $\mathbf{H}$. Third row: $\mu(\mathbf{H})$ and $\mu(\mathbf{H}'_{\text{new}})$ respectively. This shows only positive frequencies (frequencies up to half of the length of the time-series) since they have the same value as the negative frequencies. Vertical lines in plots signify an infinite value

Table 1 Median value from 10 experiments

Region	Method Without any regularization	Soft constraint on frequencies	Hard constraint on frequencies	Lasso	Ridge
Amazon	**0.87**	0.86	0.83	**0.87**	**0.87**
Congo	0.47	0.47	**0.49**	0.47	**0.49**
Ganges	0.1	0.35	0.21	**0.46**	0.4
Indus	0.41	0.44	0.1	**0.59**	0.45
Huanghe	−0.49	−0.39	−0.59	0.04	**0.18**
Volga	**0.73**	**0.73**	0.71	0.71	0.68
Parana	0	**0.03**	−0.01	0	−0.02
Lena	0.89	0.89	0.88	**0.9**	**0.9**
Mackenzie	0.87	**0.88**	0.61	0.87	0.87
Ob	0.63	0.63	0.53	0.65	**0.67**
Yenisei	0.9	**0.91**	0.88	**0.91**	**0.91**
Nile	0.6	**0.67**	0.5	0.66	**0.67**
Yangtze	0.14	0.45	0.12	**0.58**	0.57
Indigirka	0.11	**0.16**	0.11	0.12	0.14
Zambezi	0.78	**0.79**	0.72	**0.79**	**0.79**
Mississippi	0.58	0.62	0.53	0.61	**0.66**
Murray	0.43	**0.54**	0.33	0.42	0.48

The numbers in boldface represent the highest NSE values in the methods

Table 2 Median of NSE in previous research in the geophysical science [21]

Region	Method DNN	MLR	SARIMAX
Amazon	0.85	0.51	0.8
Congo	0.41	−1.49	**0.6**
Ganges	**0.83**	−0.44	0.76
Indus	0.62	−2.74	0.45
Huanghe	**0.35**	−9.53	0.29
Volga	**0.95**	0.87	0.93
Parana	−0.23	−1.21	−0.16
Lena	0.8	0.66	0.73
Mackenzie	0.68	0.09	0.65
Ob	0.71	0.35	0.75
Yenisei	0.8	0.55	0.78
Nile	0.71	−0.77	0.61
Yangtze	**0.68**	0.28	0.66
Indigirka	0.38	0.1	**0.5**
Zambezi	0.8	0.61	0.67
Mississippi	0.65	0.28	0.73
Murray	**0.72**	−0.95	0.7

The boldface numbers indicate the highest NSE values in both the current table and Table 3

Prior to discarding the atoms under the hard constraint, the NSE value in the Ob region was 0.18. However, after discarding the atoms, it dramatically increased to 0.77. See Fig. 6.

7 Conclusion

In this paper, we considered forecasting problems, specifically those arising in applications characterized by time periodicity, such as geophysical problems. In geophysical scenarios, spatial patterns exhibit changes with distinct periodicities, and frequency information becomes pivotal for accurate forecasting. To address this, we introduced novel methods based on Supervised low-rank Semi-Nonnegative Matrix Factorization (SSNMF), incorporating both soft and hard regularization in the frequency domain. These proposed methods aim to extract accurate temporal patterns from spatio-temporal data by leveraging essential periodicity information.

For the soft constraint approach, we proposed to use the Minkowski 1-norm for feature selection in the frequency domain, similar to Lasso. Although the amplitudes of undominated (noise) frequencies are reduced with soft regularization, those noisy frequencies are not completely eliminated. Consequently, we introduced the hard constraint in the frequency domain, employing three operator splitting techniques. However, for the hard constraint it is required to provide prior knowledge about the frequencies to be removed. For this, we introduced the heuristic approach, which allows for the removal of specific frequencies in the frequency domain without requiring any prior knowledge. Consequently, the forecasting of spatio-temporal data through the soft frequency, Lasso, and Ridge regularization methods comparably align with prior research in geophysical science, while hard frequency regular-

Table 3 Median of NSE the final change is the best score before and after atom removal

NSE	Without any regularization			Soft constraint on frequencies			Hard constraint on frequencies			Lasso			Ridge		
Amazon	0.87	→	0.87	0.86	→	**0.89**	0.83	→	0.86	0.87	→	0.88	0.87	→	0.88
Congo	0.47	→	0.58	0.47	→	0.58	0.49	→	0.58	0.47	→	0.56	0.49	→	0.58
Ganges	0.1	→	0.48	0.35	→	0.56	0.21	→	0.62	0.46	→	0.54	0.4	→	0.52
Indus	0.41	→	0.69	0.44	→	0.68	0.1	→	0.54	0.59	→	**0.74**	0.45	→	0.69
Huanghe	−0.49	→	0.24	−0.39	→	0.28	−0.59	→	0.13	0.04	→	0.3	0.18	→	0.3
Volga	0.73	→	0.72	0.73	→	0.71	0.71	→	0.76	0.71	→	0.71	0.68	→	0.71
Parana	0	→	0.29	0.03	→	**0.38**	−0.01	→	0.35	0	→	0.32	−0.02	→	0.34
Lena	0.89	→	0.9	0.89	→	**0.91**	0.88	→	0.9	0.9	→	0.9	0.9	→	**0.91**
Mackenzie	0.87	→	0.88	0.88	→	**0.89**	0.61	→	0.83	0.87	→	**0.89**	0.87	→	**0.89**
Ob	0.63	→	0.77	0.63	→	**0.81**	0.53	→	0.76	0.65	→	0.79	0.67	→	0.8
Yenisei	0.9	→	0.92	0.91	→	**0.93**	0.88	→	0.91	0.91	→	**0.93**	0.91	→	**0.93**
Nile	0.6	→	0.79	0.67	→	**0.8**	0.5	→	0.79	0.66	→	0.79	0.67	→	0.79
Yangtze	0.14	→	0.55	0.45	→	0.58	0.12	→	0.6	0.58	→	0.59	0.57	→	0.57
Indigirka	0.11	→	0.27	0.16	→	0.28	0.11	→	0.31	0.12	→	0.33	0.14	→	0.29
Zambezi	0.78	→	0.79	0.79	→	0.81	0.72	→	**0.82**	0.79	→	0.8	0.79	→	0.8
Mississippi	0.58	→	0.78	0.62	→	**0.8**	0.53	→	**0.8**	0.61	→	0.79	0.66	→	0.79
Murray	0.43	→	0.53	0.54	→	0.62	0.33	→	0.57	0.42	→	0.57	0.48	→	0.57

The boldface numbers indicate the highest NSE values in both the current table and Table 2. Each arrow represents the change in NSE before and after discarding an atom

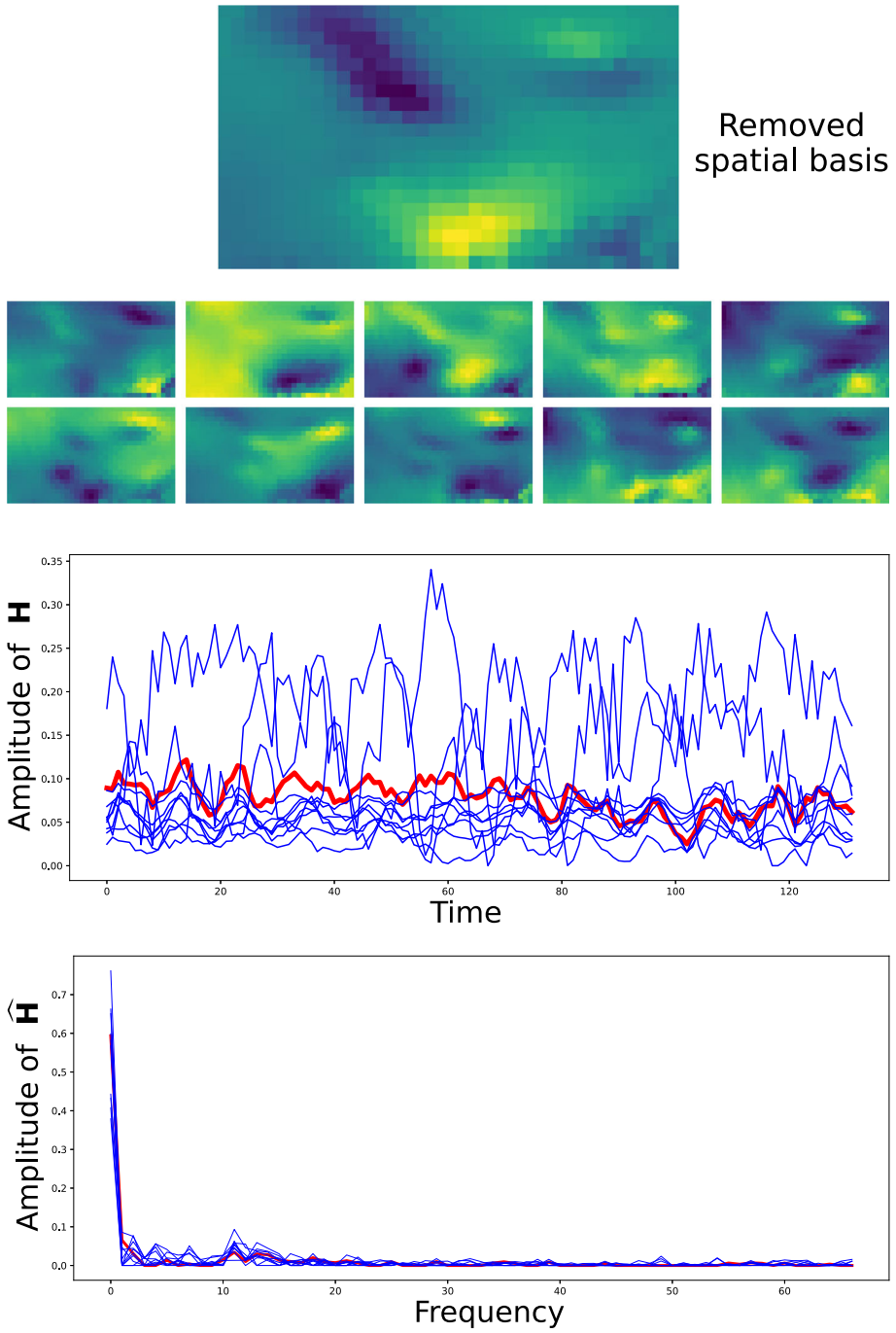

Fig. 6 In the Mississippi region with 11 atoms, $\xi = 10^4$, and 30 remaining frequencies, removing one spatial and the corresponding temporal atom (bold red line in the middle and bottom figures) causes the NSE to change from 0.1 to 0.77. This suggests that the first atom contains noisy components

ization falls behind in this regard. Nonetheless, hard frequency regularization contributes to more concise interpretation by enforcing sparsity in the frequency domain. This affords us the flexibility to choose the most suitable regularization method, depending on whether our primary emphasis lies in the time domain or the frequency domain, as well as our priority between interpretability and predictive capacity. We provided theoretical convergence results for the proposed methods. We also presented numerical results demonstrating the effectiveness of the proposed methods using GRACE data from geophysical applications. These results show the benefits of taking frequency information into account for forecasting problems, enhancing both accuracy and interpretability.

As explained in this paper, the proposed method proves effective for problems characterized by time periodicity, particularly enhancing the interpretability of the provided data. However, there are instances when non-periodic events, such as anomalous occurrences, become significant in time-series data. Our current method might overlook such frequency information. In future work, we will attempt to further develop the proposed method to address situations where non-periodic characteristics also play a crucial role in the observed phenomena within the given data. Additionally, for this study, we did not conduct a comprehensive analysis of geophysical data, specifically GRACE data, which falls outside the scope of this paper. Instead, our emphasis lies on developing the methodology and providing proofs of performance. In our future research, we will also attempt to apply the proposed method to domain-specific problems, such as geophysical problems, and perform a thorough analysis, emphasizing interpretability.

8 Notation

1. T, T_{tot}: Train time period and test period
2. $\mathcal{X}$, $\mathcal{Y}_i \in \mathbb{R}^{A\times B\times T}$: Main spatio-temporal data and supervision datas
3. r: The number of spatial/temporal patterns of $\mathcal{X}$
4. $\mathbf{H}$: Temporal pattern on train period of $\mathcal{X}$
5. S: The number of data for supervision
6. $\mathbf{H}'_{\text{new}}$: Temporal pattern on train + test period of $\mathcal{X}$
7. $\mathbf{H}_{\text{new}}$: Test period part in $\mathbf{H}'_{\text{new}}$
8. $\widetilde{\mathbf{H}}$: Concatenation of $\mathbf{H}$ and $\mathbf{H}_{\text{new}}$
9. $\|\cdot\|_F$: Frobenius norm
10. $\mathbf{X}$, $\mathbf{Y}_i \in \mathbb{R}^{d\times T}$: Matricization of $\mathcal{X}$ and $\mathcal{Y}_i$
11. $\mathbf{Y} \in \mathbb{R}^{dS\times T}$: Matrix obtained by stacking $\mathbf{Y}_1, \ldots, \mathbf{Y}_S$ vertically
12. $\mathcal{W}$, $\mathcal{W}'$, $\mathbf{W}$, $\mathbf{W}'$: Spatial pattern of $\mathcal{X}$, $\mathcal{Y}$ and its matricization
13. $\widehat{\mathbf{H}}$, $\mathcal{F}_T$: Fourier transform of $\mathbf{H}$ and Fourier transform matrix
14. ξ, λ: Regularization parameter of supervision data and frequency
15. $\|\cdot\|_{1,M}$: Minkowski 1-norm
16. R: The number of remaining frequencies
17. $\frac{\partial}{\partial z}$, $\frac{\partial}{\partial \bar{z}}$, ∇_z, $\nabla_{\bar{z}}$: Wirtinger derivatives
18. $\Re(z)$, $\Im(z)$: Real and imaginary part of z
19. $\mu(\mathbf{H})$: Inverse usage ratio of frequencies of $\mathbf{H}$

Acknowledgements We thank Jae-Seung Kim and Ki-Weon Seo at Seoul National University for their helpful discussions on the data used in this research.

Funding The second author was partially supported by the NSF through grants DMS-2206296 and DMS-2010035. The third author was supported by Samsung Electronics Co., Ltd (IO230407-05812-01), National

Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00219980, No. 2022R1C1C1008491 and No. 2021R1A6A1A10042944), and POSCO HOLDINGS research fund (2022Q019). The fourth author was supported by the NRF under the Grant Number 2021R1A2C3009648 and POSTECH Basic Science Research Institute under the NRF 2021R1A6A1A1004294412.

Data Availability The datasets used in this work are publicly accessible through the following sources. The CSR GRACE/GRACE-FO RL06 data are obtained from https://www2.csr.utexas.edu/grace/RL06_mascons.html, ERA5 monthly averaged data on single levels data from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview, GLDAS Noah Land Surface Model data from https://disc.gsfc.nasa.gov/datasets/GLDAS_NOAH025_M_2.1/summary?keywords=gldas, and the latitude and longitude information for the river basin from https://hydro.iis.u-tokyo.ac.jp/~taikan/TRIPDATA/Data/rivers.idx. The data availability of the proposed methods should be directed to the authors.

Declarations

Conflict of interest The authors confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. The authors further confirm that the order of authors listed in the manuscript has been approved by all of us. The authors also hereby certify that there is not any actual or potential conflict of interest.

References

1. Austin, W., Anderson, D., Ghosh, J.: Fully supervised non-negative matrix factorization for feature extraction. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 5772–5775. IEEE (2018)
2. Bauschke, H.H., Combettes, P.L., et al.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, vol. 408. Springer, Berlin (2011)
3. Beaudoing, H., Rodell, M., Getirana, A., Li, B.: Nasa/gsfc/hsl. GLDAS Noah Land Surface Model L4 monthly 0.25 x 0.25 degree V2, vol. 1 (2016)
4. Boyd, S., Xiao, L., Mutapcic, A.: Subgradient methods. Lecture notes of EE392o. Stanford University, Autumn Quarter **2004**, 2004–2005 (2003)
5. Brockwell, P.J., Davis, R.A.: Introduction to time series and forecasting (2016)
6. Cools, M., Moons, E., Wets, G.: Investigating the variability in daily traffic counts through use of arimax and sarimax models: assessing the effect of holidays on two site locations. Transp. Res. Rec. **2136**(1), 57–66 (2009)
7. Da Silva, A.C., Da Salva, A.C.: Lectures on Symplectic Geometry, vol. 3575. Springer, Berlin (2001)
8. Ding, C.H., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 45–55 (2008)
9. Grasmair, M.: Basic Properties of Convex Functions. Department of Mathematics, Norwegian University of Science and Technology, Trondheim (2016)
10. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. Oper. Res. Lett. **26**(3), 127–136 (2000)
11. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2. Springer, Berlin (2009)
12. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al.: Era5 monthly averaged data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) **10**, 252–266 (2019)
13. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
14. Li, H., Adalı, T.: Complex-valued adaptive signal processing using nonlinear functions. EURASIP J. Adv. Signal Process. **2008**, 1–9 (2008)
15. Oki, T., Sud, Y.: Design of total runoff integrating pathways (trip)-a global river channel network. Earth Interact. **2**(1), 1–37 (1998)
16. Petersen, K.B., Pedersen, M.S., et al.: The Matrix Cookbook, vol. 7(15), p. 510. Technical University of Denmark, Kongens Lyngby (2008)
17. Ramirez, C., Kreinovich, V., Argaez, M.: Why $\ell 1$ is a good approximation to $\ell 0$: a geometric explanation (2013)

18. Rodell, M., Houser, P., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., et al.: The global land data assimilation system. Bull. Am. Meteorol. Soc. **85**(3), 381–394 (2004)
19. Save, H.: Csr grace and grace-fo rl06 mascon solutions v02. Mascon Solut. **12**, 24 (2020)
20. Save, H., Bettadpur, S., Tapley, B.D.: High-resolution CSR GRACE RL05 mascons. J. Geophys. Res. Solid Earth **121**(10), 7547–7569 (2016)
21. Sun, Z., Long, D., Yang, W., Li, X., Pan, Y.: Reconstruction of grace data on changes in total water storage over the global land surface and 60 basins. Water Resour. Res. **56**(4), e2019WR026250 (2020)
22. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**, 475–494 (2001)
23. Yurtsever, A., Gu, A., Sra, S.: Three operator splitting with subgradients, stochastic gradients, and adaptive learning rates. Adv. Neural Inf. Process. Syst. **34**, 19743–19756 (2021)