IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 May 2023, accepted 8 July 2023, date of publication 25 July 2023, date of current version 16 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298597

RESEARCH ARTICLE

Modeling Hierarchical Seasonality Through Low-Rank Tensor Decompositions in Time Series Analysis

MELİH BARSBEY AND ALİ TAYLAN CEMGİL
Department of Computer Engineering, Boğaziçi University, 34342 İstanbul, Turkey
Corresponding author: Melih Barsbey (melih.barsbey@boun.edu.tr)

ABSTRACT Accurately representing periodic behavior is a frequently encountered challenge in modeling time series. This is especially true for observations where multiple, nested seasonalities are present, which is often encountered in data that pertain to collective human activity. In this work, we propose a new method that models seasonality through the multilinear representations that characterize low-rank tensor decompositions. We show that the tensor formalism accurately describes multiple nested periodic patterns, and well-known tensor decompositions can be used to parametrize cyclical patterns, leading to superior generalization and parameter efficiency. Furthermore, we develop a Bayesian variant of our approach which facilitates extraction of these seasonal patterns in an interpretable fashion from large-scale datasets, providing insight into the underlying dynamics that create such emergent behavior. We lastly test our method in missing data imputation, where the results show that our method couples interpretability with accuracy in time series analysis.

INDEX TERMS Time series analysis, nonnegative tensor factorization, seasonality, tensor decomposition, Bayesian model selection.

I. INTRODUCTION

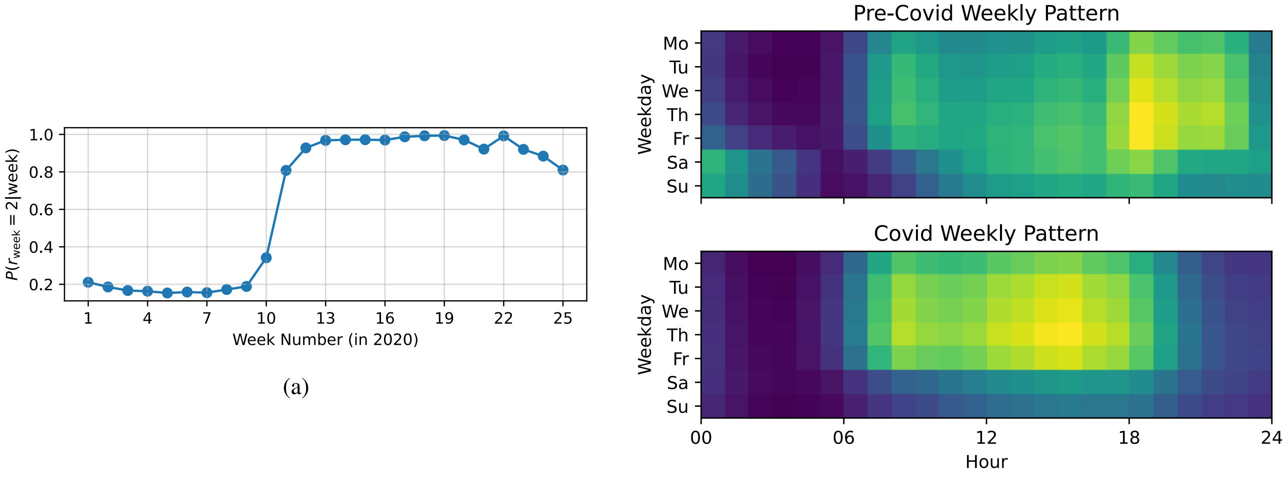
Dealing with seasonality, or periodic behavior, is a common and challenging task in time series analysis [1], [2] [3, Ch. 6]. This becomes even more challenging in contexts where multiple, nested seasonality patterns are present. Given the rhythms that determine human behavior, this is all too common in datasets that record aspects of collective human activity. For example, electricity demand in a city or total foot traffic in a university campus would involve multiple seasonal patterns, ranging from hourly to yearly cycles. In this work, we investigate a simple yet effective way of modeling seasonality in time series with arbitrarily many hierarchical seasonal components. Namely, we propose *low-rank hierarchical seasonality (LRHS)* that prescribes casting the problem as tensor decomposition, which allows us to utilize the mature arsenal for such methods to be used directly for this task. Moreover, to facilitate efficient extraction of interpretable seasonal patterns, we introduce a Bayesian version of our method (BLRHS), that not only helps make sense of very large data sets by distilling useful multi-seasonal patterns, but also demonstrates high accuracy in missing data imputation tasks. Our work complements and improves upon previous work that utilizes matrix/tensor decompositions in time series research. As a preview of our results, Figure 1 illustrates an analysis on New York City Yellow Taxi dataset, where by examining our model's latent factors we discover how complex weekly travel patterns change with the start of the Covid-19 pandemic.

Let $y_t \in \mathbb{R}$ represent a uniformly sampled univariate time series indexed by $t \in \mathbb{N}$, such that $\mathbf{y} = (y_t)_{t=1}^{T}$. Borrowing the notation of the classical time series decomposition [4, Ch. 6], one can model y_t via the additive decomposition

$$y_t = \ell_t + s_t + \varepsilon_t,$$

or a multiplicative decomposition $y_t = \ell_t \cdot s_t \cdot \varepsilon_t$; where ℓ_t is a *trend-cycle* component, s_t is the seasonal component, and ε_t are residual terms. The seasonal component s_t

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.



(a)

(b)

FIGURE 1. Changing dynamics of New York City Yellow Taxi rides through the start of Covid-19. See Section V-B for more details. Figure a: Analysis of the latent factors clearly show the transition around 11th week of 2020, just as pandemic measures came into place. The weeks before and after 11th week load onto different latent factors according to our model. Figure b: Pre-Covid and Covid travel patterns as implied by our model's latent factors (lighter color implies busier). With the pandemic measures, travel patterns are strongly restricted to daytime activities, most dramatically observed through the disappearance of late night, entertainment-related travel on weekends.

is modeled via a repeating fixed-length pattern, $\mathbf{m} \in \mathbb{R}^P$, where $P \in \mathbb{N}$ is the period of the data known a priori. One then sets $s_t = m_{t \bmod P}$, where mod denotes the modulo operator. Estimation of the classical decomposition is simple. In the additive case, this would involve first detrending the data via ℓ_t, the suitably chosen moving average, and then taking seasonal deviations m_i, $i \in \{1, \cdots, P\}$ as simple averages of deviations for season i. Within this general framework, STL and Holt-Winters methods can be thought of as variations where the trend-cycle and seasonal component estimates take different parametric forms and are updated through time. We refer the reader to [4] for a review of these methods.

Our focus here will be on the dimensionality of the seasonal component $\mathbf{m}$. In many common modeling tasks, the time series or signal contains multiple nested periodic patterns. For example, in data sampled at minutely frequency, there could be hourly, daily, and weekly periodic behavior all at once. Representing such seasonal patterns with methods above requires dimensionality of $\mathbf{m} = [m_i]_{i=1}^P$ to be $60 \times 24 \times 7 = 10080$. To tackle this problem of very long seasonal patterns with nested structure, so-called *multiple-seasonal* heuristics have been proposed [5]. For example, given a short period of P_1 and a longer period $P_1 \times P_2$, one could write $s_t = s_t^{(1)} s_t^{(2)}$ defining $s_t^{(1)} = m^{(1)}_{t \bmod P_1}$ and $s_t^{(2)} = m^{(2)}_{\lfloor t/P_1 \rfloor \bmod P_2}$ where $\mathbf{m}^{(1)} \in \mathbb{R}^{P_1}$ and $\mathbf{m}^{(2)} \in \mathbb{R}^{P_2}$. Concretely, if we are interested in hourly sampled data with both daily and weekly patterns, our notation suggests setting $P_1 = 24$ and $P_2 = 7$. By viewing seasonality as the multiplication of an hour-of-day and day-of-week patterns, this model assumption conveniently reduces the dimensionality of seasonality parameters $\mathbf{m}$ from 24×7 to $24 + 7$.

In this work, we extend such multiple seasonal patterns in two directions. First, we introduce N-many nested periodicities $P_1, P_2, \ldots, P_N$. Second, we assume that an array of individual multiseasonal patterns can be combined, *e.g.*,

$$s_t = \sum_r s_t^{(1,r)} s_t^{(2,r)}, \tag{1}$$

where $s^{(1,r)}, s^{(2,r)}$ are defined analogously to above and r indexes a set of distinct seasonal patterns.

We demonstrate that such multilinear combinations of seasonal patterns yield flexible and accurate representations of many cyclical patterns that occur in real-world data. Moreover, we demonstrate that components $s^{(n,r)}$ can be recovered easily by casting estimation as an instance of low-rank tensor decomposition. The data representation implied by our approach can be seen in Figure 2. Through various experiments, we show that our method accurately and efficiently captures hierarchical seasonality in time series. Before moving on to describing related literature and our method, we summarize our contributions:

- We propose a novel method for representing seasonality through low-rank multilinear decompositions, called *low-rank hierarchical seasonality (LRHS)*.
- Through experiments on various datasets we show that our approach is both accurate and parameter-efficient compared to other canonical methods.
- We propose a Bayesian variant of our method *(BLRHS)*, allowing interpretable and scaleable knowledge extraction from multivariate datasets, uniquely combining probabilistic inference with nonnegative factorization in the analysis of complex seasonality.
- We show that our method efficiently extracts useful representations from large-scale time series data, e.g.

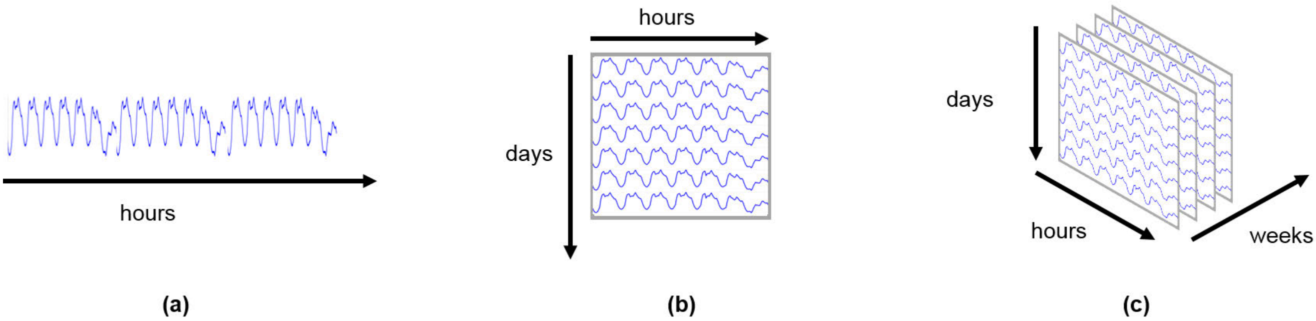


FIGURE 2. Folding time series into a tensor. In (a), the time series is represented as a single vector. In (b), each hour of day is presented in a new mode. In (c), each week is a slice as days of the week are folded into the third mode. Extension to lower or higher seasonality components is straightforward, and explored in our numerical experiments.

12 years of *hourly* ridership history of a prominent rail network, with over 1 billion trips spread among 2500 different routes.

- We show that our method does not sacrifice performance to achieve interpretability, and demonstrate that BLRHS is highly accurate in multivariate missing data imputation tasks.

II. RELATED WORK

Matrix/tensor based analysis of time series has been attracting considerable interest from various fields. The scope of such research ranges from theoretical and algorithmic developments [6], [7] to numerous applications, the latter including energy management and load monitoring [8], [9], intelligent traffic systems [12], [13], mobility and urban planning [14], logistics [15], and structural monitoring [16]. These work often represent time as one of the modes of a multiway data representation. For example, Dunlavy et al. [17] factorize a three-way tensor where one of the modes is time, and use Holt-Winters method on the *temporal* factor matrix to extrapolate to new time points. Similarly, de Araujo et al. [18] apply exponential smoothing to the temporal factor for forecasting, within a coupled factorization framework. A Bayesian approach to the same problem is studied by Xiong et al. [19]. Matsubara et al. [20] apply a *multi-scale* type autoregressive model for extrapolating the temporal dimension.

In a line of work that pertains to epidemiology applications, Matsubara et al. [21] introduce FUNNEL, a model for coevolving epidemics inspired by the well-known SIR model. FUNNEL incorporates diseases, different locations, and time into a data tensor while accounting for seasonal patterns by including a sinusoidal seasonal component. Rogers et al. [22] propose a multilinear extension of the well known linear dynamical system. Yu et al. [23] propose a matrix factorization approach where the autoregressive time series model likelihood appears as a regularization term. An extension of this idea into the spatio-temporal domain is proposed by Takeuchi et al. [24], and Sun and Chen [25] provide a Bayesian version of the same approach. More recently, Kawabata et al. [26] propose an online inference scheme for multivariate data where the model can detect the appearance of a new seasonality regime.

Methods surveyed above use the tensor factorization formalism with only one mode of the tensor of interest representing time, reserving other modes for different entities (*e.g.*, users, items, locations, etc.) An idea that follows immediately is to let one of the modes represent seasonal behavior, *i.e.*, to "fold" the time series into a 2-way array, as in Figure 2b. In a more general tensor factorization framework, this means that each latent dimension has a representation in the entity domains, as well as a certain seasonal pattern. This idea appears early in the beginning of time series analysis with factorizations, e.g. in an atmospheric science application by Xie et al. [27]. Takahashi et al. [28]. apply this idea to various time series, with a specific focus on folding periods—or *cyclic(al) folding* in their terms which we reuse here. Cyclical folding also appears in Matsubara et al. [29], [30], which are models of competing entities inspired by population modeling. Chen and Sun [31] augment the work of [23] by adding another temporal dimension of seasonality for forecasting, and Chen and Lei [32] extend this approach to missing data imputation.

Though less common, some studies extend the cyclical folding idea to higher temporal dimensions, modeling two seasonalities simultaneously. Notably, Tan et al. [33] uses a sliding temporal tensor construction to impute future values for multivariate time series, and Chen and He [34] compare the performance of having cyclical folding to obtain zero, one, or two seasonality dimensions while utilizing a CP (CANDECOMP/PARAFAC) decomposition for missing data imputation. Given [34] provides the only quasi-systematic exploration of multiple cyclical folding, we compare our method to theirs, as well as some other baselines in Section V-D to provide an informative test of our approach.

Also less common is the use of nonnegative matrix/tensor factorization to model time series data, and especially with consideration of hierarchical seasonality. Espin-Noboa et al. [14] apply nonnegative tensor factorization to decompose the NYC taxi dataset into coherent patterns to test hypotheses about urban mobility. Wang et al. [35]

apply nonnegative matrix factorization techniques for the analysis of traffic patterns, and Figueiredo et al. [8] provide an analysis of power consumption through nonnegative tensor decompositions by including two temporal dimensions, however neither studies exploit hierarchical patterns in seasonality. The relative neglect of nonnegative methods is especially important from a knowledge discovery point of view. This is because nonnegative decompositions are known to lead to interpretable, "by-parts" representations [36], [37], which could help extract the simpler multi-seasonal components that produce the emergent complex behavior of such time series. To our knowledge, the Bayesian variant of our approach is the first application of probabilistic nonnegative tensor decomposition to the analysis of complex seasonality.

Our method is based on the hypothesis that low-rank multilinear representations can accurately and efficiently capture nested, complex seasonalities. It directly targets to exploit the low-rank decomposability of such intricate patterns, improving on the sporadic use of cyclical folding as a data representation heuristic for specific multivariate time series problems in the literature. Our proposal's generality allows us to apply and test this idea with different decomposition methods (CP vs. Tucker), in various tasks (prediction, imputation, knowledge extraction), in different experimental settings (univariate vs. multivariate), with arbitrary number of seasonal hierarchies (e.g. hourly, daily, weekly, yearly), and with previously unexplored model capabilities (likelihood-based model selection). As models and data get larger in modern machine learning, the need for such parsimonious representations increase. As such, our results not only contribute to tensor-based time series analysis, but also offers valuable insight for classical and deep learning based approaches, given the transferability of this inductive bias.

III. PRELIMINARIES

In the following, we use lowercase italics for scalars (a, γ), uppercase italics for integers (I, J), lowercase bold letters for vectors ($\mathbf{a}, \mathbf{b}$), uppercase bold letters for matrices ($\mathbf{A}, \mathbf{B}$), and uppercase cursive letters for tensors of higher order ($\mathcal{X}, \mathcal{A}$). a_i refers to the i'th element of a vector $\mathbf{a}$, and $B_{i,j}$ refers to the i, j'th element of a matrix $\mathbf{B}$. For convenience, we refer to the elements of tensors using parantheses notation, such that the element of $\mathcal{X}$ with the index i, j, k is denoted with $\mathcal{X}(i, j, k)$. For any positive integer L, we let $[L] := \{1, \dots, L\}$. We use $i_{[L]} := (i_1, \dots, i_L)$ to denote an L-tuple of indices, which we can use to refer to tensor elements, as in $\mathcal{X}(i_{[L]})$.

A. TENSOR DECOMPOSITIONS

Tensors are N-way arrays, where 1-way and 2-way tensors are commonly known as vectors and matrices respectively [38]. Tensor *decompositions* are approximations of tensors as multilinear functions of factor matrices or tensors. For the purposes of this text, we use tensor *decomposition* and *factorization* interchangeably. A frequently used tensor decomposition is the *CP decomposition*, also known as CANDECOMP/PARAFAC or canonical polyadic [39], [40]. In CP, the original N-way tensor is expressed as the sum of a finite number of *rank-one N-way tensors*, where a rank-one N-way tensor is the outer product of N vectors. The CP decomposition of a 3-way tensor would be:

$$\mathcal{X} \approx \widehat{\mathcal{X}} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r,$$

where $\mathcal{X}, \widehat{\mathcal{X}} \in \mathbb{R}^{I \times J \times K}$, $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$, $\mathbf{c}_r \in \mathbb{R}^K$ for $r \in \{1, \dots, R\}$, and $\circ$ denotes outer product. Then, *factor matrices* $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ refer to the matrices constituted by collecting these vectors as columns of separate matrices, as in $\mathbf{A} \in \mathbb{R}^{I \times R} = [\mathbf{a}_1, \dots, \mathbf{a}_R]$. It is also common to assume factor matrices to have unit columns and to have an additional weight vector $\mathbf{w} \in \mathbb{R}^R$, such that $\mathcal{X} \approx \widehat{\mathcal{X}} = \sum_{r=1}^{R} w_r (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$.

Another frequently used decomposition is the *Tucker decomposition* [41], [42]. Tucker decomposition approximates the target N-way tensor with an N-way *core tensor* $\mathcal{G}$ and N factor matrices. Specifically, in Tucker decomposition the target tensor is approximated by the sum of rank-one tensors weighted by the components of the core tensor. For example, for $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ we would have:

$$\mathcal{X} \approx \widehat{\mathcal{X}} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} (\mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r),$$

where g_{pqr} are the entries of the core tensor $\mathcal{G} \in \mathbb{R}^{P \times Q \times R}$ and $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$ correspond to the factor matrices constructed as above. A succinct notation for Tucker decomposition is $\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$ [38], which we will use below.

Computing a CP or Tucker decomposition involves *learning* the factors as part of an optimization scheme

$$\underset{\mathbf{A},\mathbf{B},\mathbf{C}}{\arg\min}\, D(\mathcal{X} \| \widehat{\mathcal{X}}),$$

with an appropriately chosen divergence function D. Most popular choices include the Euclidean distance ($\|\mathcal{X} - \widehat{\mathcal{X}}\|$) and Kullback-Leibler (KL) divergence. Commonly used optimization algorithms for CP and Tucker decompositions include alternating least squares (ALS) [39], [40] and higher-order orthogonal iteration (HOOI) [43], respectively [38]. Tensor decompositions are utilized in a wide variety of scientific disciplines ranging from signal processing [44], [45], [46] to social science [47]. For other kinds of decompositions, problems, and theoretical results see [38], [48], [49], and [50].

1) NONNEGATIVE MATRIX/TENSOR FACTORIZATIONS

An important variant of matrix/tensor decompositions is nonnegative matrix/tensor factorization (NMF/NTF), which targets nonnegative data, and also constrains factor matrices to have nonnegative entries [51], [52], [53]. Nonnegative variants of the aforementioned CP and Tucker decompositions are two of the most prominent NTF methods [38], with similar choices for D. Usually utilized optimization methods for NMF/NTF include projected gradient descent and multiplicative updates [37]. Research on NMF/NTF produced numerous approaches throughout the years; see [37], [54], and [55] for informative reviews.

2) PROBABILISTIC NTF METHODS

Probabilistic formulations of NMF and NTF propose a generative model for the observed tensor $\mathcal{X}$, which allows posterior inference regarding the model parameters, i.e. the latent factors [36]. These include point estimates such as maximum likelihood and maximum a posteriori (MAP), as well as full posterior inference through sampling-based or variational procedures. The formulation of NTF as a probabilistic modeling task allows the wide range of inference methods developed in Bayesian statistics to be used for NMF/NTF [56]. Such models prove to be capable of extracting information from high-dimensional, sparse, relational data, combining the representative power of nonnegative decompositions with well-quantified uncertainties allowed by novel inference and model scoring capabilities [57], [58], [59].

IV. METHOD

We now present our proposed approach for modeling hierarchical seasonality, which we call *low-rank hierarchical seasonality* (LRHS). Afterwards, we will present its probabilistic variant.

A. LOW-RANK HIERARCHICAL SEASONALITY (LRHS)

Recall the definition of a *multiple seasonal* s_t, and how it was defined as the *multiplication* of *factors* $s_t = s_t^{(1)} s_t^{(2)}$ to account for two "nested" periodic patterns, such as days-of-week and hours-of-day. We now formally define our modeling approach by generalizing this idea to N seasonalities and multiple such nested patterns, and we will rely on tensor formalism when doing so. This implies a particular representation of the data, where we "fold" an observed seasonality vector $\mathbf{s}$ based on its seasonalities to create a tensor:

Definition 1 (Multiple Cyclical Folding; MCF): Let $\mathbf{s} \in \mathbb{R}^T$ *denote a vector that contains observed values of the seasonal component of a time series, and let* $P_1, \ldots, P_N \in \mathbb{Z}^+$ *be the periods for the* N *nested seasonalities observed in the data. Assume the data includes* $K \in \mathbb{Z}$ *full cycles of the longest seasonality, such that* $T = K \cdot \prod_{i=1}^{N} P_i$. *Multiple cyclical folding creates the tensor* $\mathcal{M} \in \mathbb{R}^{P_1 \times \cdots \times P_N \times K}$:

$$\begin{aligned} &\mathcal{M} \\ &= MCF(\mathbf{s}), \ \text{such that}\ \mathcal{M}(t \bmod P_1, \\ &\quad \lfloor t/\bar{P}_1 \rfloor \bmod P_2, \ldots, \lfloor t/\bar{P}_{N-1} \rfloor \bmod P_N, \lfloor t/\bar{P}_N \rfloor) := s_t, \end{aligned}$$

where $\bar{P}_n = \prod_{i=1}^{n} P_i$.

A visualization of this data representation is presented in Figure 2. For example, if we were to model 10 weeks of hourly data with daily and weekly seasonalities, this would imply that $T = 24 \times 7 \times 10 = 1680$ and $\mathbf{s} \in \mathbb{R}^{1680}$, with $P_1 = 24, P_2 = 7$, and $K = 10$. After MCF we would have $\mathcal{M} \in \mathbb{R}^{24 \times 7 \times 10} = \text{MCF}(\mathbf{s})$. As defined above, the last dimension of $\mathcal{M}$ indexes the *full cycles* of the data, such that $k = 1$ for week 1, $k = 2$ for week 2 and so forth. For brevity we will call this dimension *cycle index*. With the exception of Definition 1 that uses 0-indexing for convenience, we use 1-indexing convention in tensors for a more natural exposition.

As we will see below, our modeling approach naturally extends to multivariate time series. With multivariate data, time series index becomes another dimension of the representation (or dimension*s*, should there exist multiple non-temporal indices). That is, if the example dataset above had measurements from $I = 100$ stations, MCF would produce $\mathcal{M} \in \mathbb{R}^{100 \times 24 \times 7 \times 10}$. Having described our data representation, we now introduce our modeling approach.

Definition 2 (Low-Rank Hierarchical Seasonality; LRHS): Let $\mathcal{M} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_N \times K}$ *be the multiple cyclical folding of an observed seasonality* $\mathbf{s} \in \mathbb{R}^T$, *such that* $\mathcal{M} = MCF(\mathbf{s})$. *Low-rank hierarchical seasonality (LRHS) proposes the following approximation for modeling seasonality*:

$$\mathcal{M} \approx \widehat{\mathcal{M}} = \mathcal{G} \times_1 \mathbf{M}^{(1)} \ldots \mathbf{M}^{(N)} \times_{N+1} \mathbf{K}, \qquad (2)$$

where $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_N \times R_K}$ *is the latent core tensor,* $\mathbf{M}^{(n)} \in \mathbb{R}^{P_n \times R_n}$, $\forall n \in [N]$ *and* $\mathbf{K} \in \mathbb{K}^{K \times R_K}$ *are the factor matrices for the seasonal components and the cycle index respectively. Factors can be found by minimizing* $D(\mathcal{M} \| \widehat{\mathcal{M}})$ *with a suitably chosen divergence function* D. *A multivariate extension is straightforward with an additional mode of cardinality* I *and the corresponding factor matrix* $\mathbf{I} \in \mathbb{R}^{I \times R_I}$.

We follow our definition with some remarks. First, note that the presence of the cycle index, and the corresponding factor $\mathbf{K}$ enables LRHS to detect and model *changing* multi-seasonal patterns through K full cycles. For example, in Figure 1 LRHS captures the fact that although weekday daytime taxi trips existed in the weeks before Covid ($k \leq 10$), their relative importance tangibly increased with Covid ($k > 10$) as other seasonal patterns withered. We will see more examples of LRHS capturing such dynamics in Section V.

Another point to note is that for full generality, we introduced LRHS with a Tucker decomposition in Definition 2, where all factor matrices have their own latent factor with different cardinalities $R_1, \ldots, R_N, R_K$, and core tensor $\mathcal{G}$ determines how these latent factors interact. However, if we set $R = R_1 = \cdots = R_N = R_K$ and let $\mathcal{G}$ be superdiagonal, this formulation would correspond to a CP decomposition. Also importantly, other constraints on factors can also be easily integrated to this framework: e.g. a *nonnegative* variant would assume nonnegative observations and would constrain all factors to be nonnegative.

The *main hypothesis* of this paper is that LRHS is an accurate and efficient approach for modeling hierarchical seasonality. LRHS is designed for contexts where the periods of these seasonalities are known a priori: this is most importantly the case for datasets that characterize various aspects of collective human behavior. For example, the number of customers arriving at a restaurant can be driven by two distinct segments of customers, those who prefer lunches on weekdays, and others who come to Friday and Saturday dinners. Our results in Section V indeed confirm the abundance of such compositions, in complex and meaningful interactions with non-temporal (*i.e.* spatial) dimensions. LRHS allows the rich arsenal of tensor decomposition methodology to be used in the accurate and efficient recovery of such patterns.

1) UTILIZING LEARNED LRHS COMPONENTS

LRHS allows the specification of a decomposition model and an optimization scheme. Once we recover the component factors based on our modeling assumptions, these learned factors can be utilized for various inferential tasks. First, the learned factors can directly be investigated to discover the underlying dynamics that produce the temporal patterns in question (as in [14]). Second, in the case of *missing data*, given a set of non-missing indices Ω and a projector $\Pi : \mathbb{R}^{P_1\times\cdots\times P_N\times K} \to \mathbb{R}^{P_1\times\cdots\times P_N\times K}$ that sets cells with missing observations to 0, learning could take place as

$$\underset{\mathcal{G},\mathbf{M}^{(1)},\dots\mathbf{M}^{(N)},\mathbf{K}}{\arg\min} \quad D\left(\Pi(\mathcal{M})\|\Pi\left(\mathcal{G}\times_1\cdots\times_{N+1}\mathbf{K}\right)\right),$$

after which the reconstruction $\widehat{\mathcal{M}} = \mathcal{G}\times_1\mathbf{M}^{(1)}\dots\mathbf{M}^{(N)}\times_{N+1}\mathbf{K}$ can be used to fill in the missing data (as in [32]).

Lastly, for predicting future observations, imagine without loss of generality that we have a *forecast horizon* of one full cycle, i.e. $1\cdot\prod_{i=1}^{N} P_i$. Then, an appropriately chosen cycle index row vector $\mathbf{k}_{K+1} \in \mathbb{R}^{1\times R_K}$ can be used to compute the prediction for the new cycle:

$$\widehat{\mathcal{M}}_{K+1} = \mathcal{G}\times_1\mathbf{M}^{(1)}\dots\mathbf{M}^{(N)}\times_{N+1}\mathbf{k}_{K+1},$$

where $\widehat{\mathcal{M}}_{K+1}$ hold the predictions for the forecast horizon. $\mathbf{k}_{K+1}$ can be chosen in various ways, such as setting it equal to the last cycle index of the model $\mathbf{k}_{K+1} := \mathbf{k}_K$ (as in [26]), or computing it through the exponential smoothing of past cycle indices: $\mathbf{k}_{K+1} := \widehat{\mathbf{k}}_K$, where $\widehat{\mathbf{k}}_\kappa = \alpha\mathbf{k}_\kappa + (1-\alpha)\widehat{\mathbf{k}}_{\kappa-1}$ and $\widehat{\mathbf{k}}_0 = \mathbf{k}_1$ by convention (as in [18]). If detrending was conducted beforehand, both imputation and prediction would require re-adding the trend component (or its extrapolation) before finalizing inference.

B. BAYESIAN LOW-RANK HIERARCHICAL SEASONALITY (BLRHS)

We now define a Bayesian variant of LRHS, for robust and interpretable posterior inference and convenient model selection. Here, the fact that the time series data of interest usually are nonnegative (e.g. counts of trips, occupancy of routes, amount of energy consumption) presents a unique opportunity to integrate NTF methods into our formulation. This allows us to exploit the additive, sparse, by-parts representations NTF methods are known to produce [37] and combine it with convenient inference and model selection methods probabilistic models enable. As discussed above, probabilistic NTF is a vibrant research area with various modeling approaches adopting different choices for model structure and conditional distributions [57], [58], [60]. Here, we emulate the choices by [58], whose modeling approach, called Bayesian Allocation Model (BAM), allows expedited inference and model selection. We now introduce BLRHS with a CP decomposition for brevity, and present its Tucker variant in the supplementary material.

Definition 3 (Bayesian Low-Rank Hierarchical Seasonality; BLRHS): Let $\mathcal{M} \in \mathbb{R}_{\geq 0}^{P_1\times\cdots\times P_N\times K}$ *be the multiple cyclical folding of an observed seasonality. Let* $\mathbf{m}_r^{(n)}$ *(resp.* $\mathbf{k}_r$*) be the* r'*th column of the random factor matrix* $\mathbf{M}^{(n)}$*, (resp.* $\mathbf{K}$*). Bayesian low-rank hierarchical seasonality (BLRHS) proposes the following generative model for seasonality*:

$$P(\mathcal{M}|\widehat{\mathcal{M}}) = \prod_{i_{[N+1]}} \text{Poisson}(\mathcal{M}(i_{[N+1]})|\widehat{\mathcal{M}}(i_{[N+1]})),$$

$$\widehat{\mathcal{M}} = \lambda\sum_{r=1}^{R} w_r\left(\mathbf{m}_r^{(1)}\circ\cdots\circ\mathbf{m}_r^{(N)}\circ\mathbf{k}_r\right),$$

$$\begin{aligned}\lambda &\sim \text{Gamma}(a, b),\\ \mathbf{w} &\sim \text{Dirichlet}(\mathbf{1}\cdot\alpha(R)),\\ \mathbf{m}_r^{(n)} &\sim \text{Dirichlet}(\mathbf{1}\cdot\alpha(P_n\cdot R)),\ \forall r\in[R], \forall n\in[N]\\ \mathbf{k}_r &\sim \text{Dirichlet}(\mathbf{1}\cdot\alpha(K\cdot R)),\ \forall r\in[R].\end{aligned}$$

Here the Dirichlet distributions have flat priors with the concentration parameter $\alpha(L) := a/L$*, and* a, b *are hyperparameters of the model. A multivariate extension of this construction is straightforward with an additional mode of cardinality* I *equal to the number of time series and the corresponding factor matrix* $\mathbf{I} = [\mathbf{i}_1,\dots,\mathbf{i}_R]$*, with* $\mathbf{i}_r \sim \text{Dirichlet}(\mathbf{1}\cdot\alpha(I\cdot R))$.

See Figure 3 for graphical models implied by BLRHS, with CP and Tucker decompositions. We now describe inference with BLRHS, and then detail how the results can be used in *interpreting* seasonality components.

1) INFERENCE WITH BLRHS

Two inferential tasks are important for utilizing the model defined above. One is *posterior inference*, that is obtaining or approximating the posterior of the model parameters:

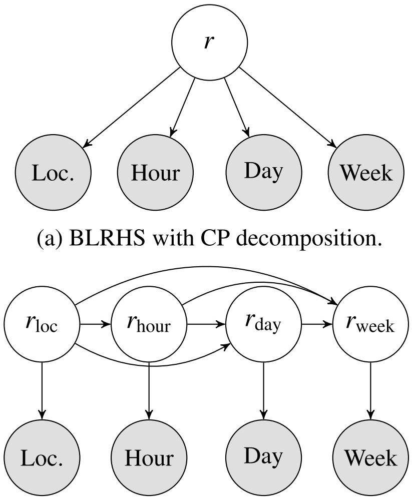
$$P(\lambda, \mathbf{M}^{(1)},\dots,\mathbf{M}^{(N)},\mathbf{K},\mathbf{w}|\mathcal{M}, R, a, b), \tag{3}$$

where R is the latent rank and a, b are the hyperparameters. Posterior inference is the probabilistic counterpart of learning the latent factors. Another important inferential task would be computing the *marginal likelihood* of the data $\mathcal{M}$:

$$P(\mathcal{M}|R, a, b), \tag{4}$$

where the model parameters, $\lambda, \mathbf{M}^{(1)},\dots,\mathbf{M}^{(N)},\mathbf{K},\mathbf{w}$, are integrated out. Marginal likelihood is a crucial quantity, as it can be used to score the overall model, thus be used for model selection. However, it is a notoriously difficult quantity to estimate in such latent variable models [61].

A frequently used inference algorithm for latent variable models is mean-field variational inference (MFVI), which we



(a) BLRHS with CP decomposition.

(b) BLRHS with Tucker decomposition.

FIGURE 3. BLRHS utilizing (a) CP and (b) Tucker decompositions, modeling multivariate time series with temporal observations at I different locations, and with $P_1 = 24, P_2 = 7$, for K weeks of data. Higher or lower seasonality terms can be added if desired. Hyperparameters are omitted for brevity.

utilize for inference with BLRHS. MFVI proposes to approximate the model posterior (3) with a factorized *variational* distribution, the parameters of which can be optimized to minimize the KL-divergence between the actual and variational posteriors, $D_{KL}(Q\|P)$. Ignoring hyperparameters for brevity, regarding the model proposed in Definition 3 MFVI would define the following variational distribution:

$$\begin{aligned} &P(\lambda, \mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(N)}, \mathbf{K}, \mathbf{w}|\mathcal{M}) \\ &\approx Q(\lambda, \mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(N)}, \mathbf{K}, \mathbf{w}) \\ &= q(\lambda)q(\mathbf{M}^{(1)})\ldots q(\mathbf{M}^{(N)})q(\mathbf{K})q(\mathbf{w}). \end{aligned}$$

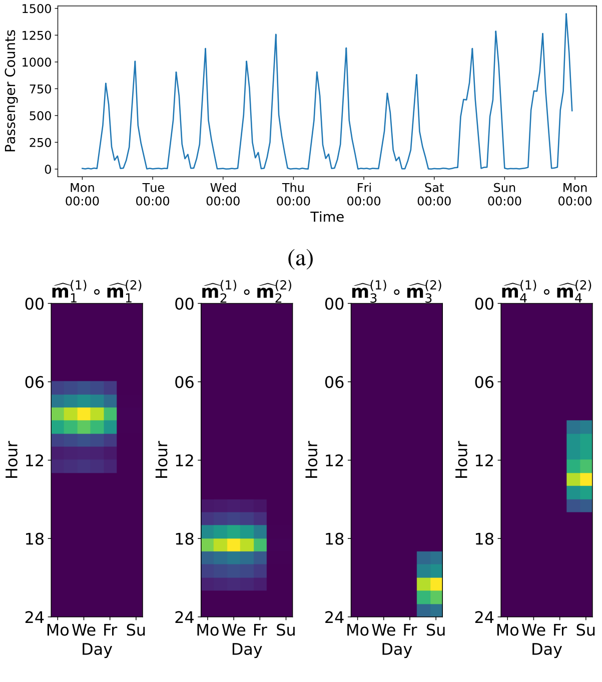
MFVI allows each q to be updated in an alternating fashion to minimize the $D_{KL}(Q\|P)$, in a procedure sometimes called coordinate ascent variational inference (CAVI). In BLRHS, with the use of an *auxiliary latent tensor* prescribed by the BAM framework (as detailed in the supplement), these updates remain tractable even for data with very large dimensions.

A closer look at $D_{KL}(Q\|P)$ reveals another relation:

$$\log P(\mathcal{M}) = D_{KL}(Q\|P) - \mathbb{E}_Q[\log Q(\zeta) - \log P(\zeta, \mathcal{M})],$$

where we let $\zeta = (\lambda, \mathbf{M}^{(1)}, \ldots)$ collect all parameters for brevity. Since marginal log-likelihood $\log P(\mathcal{M})$ is *constant* given the model, decreasing D_{KL} implies increasing $-\mathbb{E}_Q[\log Q(\zeta) - \log P(\zeta, \mathcal{M})]$. Since D_{KL} is always nonnegative, this also means that this latter term is a lower bound to marginal log-likelihood. A well known quantity in probabilistic inference, this term is called *evidence lower bound (ELBO)* and it is frequently used for model selection as an approximation of marginal (log-)likelihood [62].

Thus the MFVI procedure for BLRHS allows not only expedient posterior inference, but also convenient model selection. For example, after MFVI we can compute the ELBO with BLRHS's CP and Tucker variants with different ranks, and select the highest-scoring model. The previous work we reviewed in Section II rarely, if ever, provide a method for principled model selection, which BLRHS allows naturally, as seen in our results in Section V.



(a)

(b)

FIGURE 4. Synthetic example demonstrating that once BLRHS is trained on a time series (shown in a), its estimated factors can be utilized to explore the latent patterns captured by it (shown in b, lighter implies busier). Remember that $\widehat{P}$(hour, day$|r = i) := \widehat{\mathbf{m}}_i^{(1)} \circ \widehat{\mathbf{m}}_i^{(2)}$, so the heatmaps for $i = \{1, 2, 3, 4\}$ correspond to four disparate weekly patterns that are captured by BLRHS.

The exact form of the CAVI updates for BLRHS, as well as a more in-depth discussion of inference with this modeling approach can be found in the supplementary material. Importantly, our particular implementation of this procedure (`https://github.com/mbarsbey/lrhs`) exploits the parallel processing capabilities of modern hardware and computational frameworks, allowing us to conduct inference on and make sense of very large temporal datasets.

2) INTERPRETABLE SEASONALITY WITH BLRHS

To demonstrate how BLRHS helps extract interpretable representations from hierarchically seasonal data, let us start with a toy univariate dataset, pertaining to the ridership counts for a subway station. We model daily and weekly seasonalities ($P_1 = 24$, $P_2 = 7$), and observe this station for K full cycles (weeks). Notice that the parameters of this decomposition according to BLRHS would correspond to conditional probability tables such that:

$$P(\text{hour} = j|r = i) := M_{j,i}^{(1)}, \quad \text{or} \quad P(\text{hour}|r = i) := \mathbf{m}_i^{(1)}.$$

Although this is promising, by definition we do not have direct access to these factors, as they are latent variables. However, we can use our *estimates* of these through

MFVI, i.e. $\widehat{\mathbf{M}}^{(1)} = \mathbb{E}_Q(\mathbf{M}^{(1)})$ to examine and utilize these latent factors:

$$\widehat{P}(\text{hour} = j|r = i) := \widehat{M}^{(1)}_{j,i}, \quad \text{or} \quad \widehat{P}(\text{hour}|r = i) := \widehat{\mathbf{m}}^{(1)}_i.$$

This interpretability of parameters (or their estimates) as probabilities enables various interesting queries to be made through a trained BLRHS model. For example, we could examine the joint daily-hourly seasonality patterns that are captured by our latent factor:

$$\widehat{P}(\text{hour}, \text{day}|r = i) := \widehat{\mathbf{m}}^{(1)}_i \circ \widehat{\mathbf{m}}^{(2)}_i.$$

We present idealized examples of such queries in Figure 4. Conducting inference with BLRHS on the synthetic time series seen in Figure 4a with $R = 4$ results in the factors seen in Figure 4b. By examining $\widehat{P}(\text{hour}, \text{day}|r = 1) := \widehat{\mathbf{m}}^{(1)}_1 \circ \widehat{\mathbf{m}}^{(2)}_1$ we can see that $r = 1$ has captured morning commute in this time series. We could also infer that $r = 2, 3, 4$ correspond to evening commute, weekend night entertainment, and day trips, respectively. We could also look at $\widehat{P}(r = i) := \hat{w}_i$ to see the strength of these patterns in the overall dataset. We could track their strengths *throughout* weeks, $\widehat{P}(r|\text{week})$, to explore the dynamics of these patterns through time, as in weekend night taxi rides disappearing with Covid in Figure 1. In Section V, we investigate these complex interactive patterns and examine how they interact with other variables (e.g. location, yearly seasonality) on large real life datasets. The public repository for this paper includes trained models of BLRHS for the reader to query and explore these spatiotemporal patterns as desired.

V. RESULTS

We now conduct various experiments to test whether LRHS/BLRHS accurately captures seasonal patterns in various datasets and problem contexts. Source code for reproducing our experiments and experimenting with the proposed modeling approach is provided in our public GitHub repository.[1]

A. TESTING LOW-RANK HIERARCHICAL SEASONALITY in UNIVARIATE TIME SERIES

As the first test of our idea, we conduct numerical experiments on a range of time series prediction tasks on univariate data, comparing LRHS with canonical methods for handling seasonality from signal processing and time series analysis.

1) DATASETS

We use the Bay Area Rapid Transport (BART) ridership data provided by its administration,[2] Hourly Energy Consumption[3] dataset from Kaggle, and Electricity and Traffic datasets from GluonTS [63]. Electricity and Traffic datasets include 321 and 862 individual time series, and a total of 125 and 83 weeks of hourly data respectively. Given LRHS is developed for modeling seasonality, series that do not have pronounced seasonal components are of little interest in evaluating LRHS. So, we take time series in which the seasonal component accounts for 75 percent of the variance ($R^2 > 0.75$) after subtracting the trend-cycle component. To make sure that our conclusions are not based on a specific choice of cut-off point, we also repeat this procedure with a threshold of 85 percent. This gives us two different versions of these datasets, which we denote by adding -75 and -85 as suffixes to their names. The resulting number of time series are 236 and 146 for Electricity-75 and Electricity-85, and 508 and 133 for Traffic-75 and Traffic-85 respectively. Kaggle Energy and BART datasets provide a different challenge as they have much longer histories. From Kaggle Energy dataset we take the 6 series that have more than 12 years of data, and from BART dataset we use the 10 mostly occupied routes between 2011-2019.[4]

For all datasets, we take $P_1 = 24, P_2 = 7$ and use the last 10 cycles of the data (1680 hours) as the held out set, *i.e.*, the forecast horizon. Additionally, since Energy and BART datasets are considerably longer than the others, we use this as an opportunity to create an additional experiment setting: we set $P_1 = 24, P_2 = 7, P_3 = 52$ and keep the final year of each time series as the held out sample (8736 hours). These variants are denoted Energy-1Y and BART-1Y, respectively. Notice that this representation results in four temporal dimensions (hour, day, week, year), speaking to the ability of LRHS to take modeling hierarchical seasonality to previously unexplored extents.

2) EXPERIMENTAL SETTING, BASELINES, METRICS

We first compute and subtract the trend-cycle component l_t via a simple moving average of twice the cycle length (i.e. $2 \times \bar{P}_N$), assuming an additive decomposition. We then attempt to forecast seasonality for the held out periods via eight methods. The baselines include discrete cosine (DCT) and Fourier transforms (DFT), Fourier basis regression (FB) (described fully in the supplemental material), and Holt-Winters method (HW) [64]. We use LRHS with CP and Tucker decompositions, denoted LRHS-CP and LRHS-T, respectively. We also utilize a variant of either where we conduct smoothing of the temporal factor as described in Section IV-A1. Therefore, an additional -S suffix for LRHS methods denotes the version where smoothing adjustment is applied.

We compare the methods using mean absolute error (MAE) and root mean squared error (RMSE). Given a ground truth vector and an estimator thereof, $\mathbf{y}, \widehat{\mathbf{y}} \in \mathbb{R}^T$, these metrics are defined as:

$$\begin{aligned} \text{MAE}(\mathbf{y}, \widehat{\mathbf{y}}) &= \tfrac{1}{T} \textstyle\sum_{t=1}^{T} |y_t - \widehat{y}_t| \\ \text{RMSE}(\mathbf{y}, \widehat{\mathbf{y}}) &= \left(\tfrac{1}{T} \textstyle\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \right)^{\frac{1}{2}} \end{aligned}$$

[1] https://github.com/mbarsbey/lrhs

[2] https://www.bart.gov/about/reports/ridership

[3] https://www.kaggle.com/robikscube/hourly-energy-consumption

[4] For BART experiments, we limit ourselves to pre-Covid era due to dramatic trend changes in Spring 2020. However, see Section V-C for a full, multivariate analysis of the BART data.

TABLE 1. Mean absolute errors (MAE) and root mean squared errors (RMSE) of forecasts for held out time series. The two best performing algorithms are presented in bold.

	MAE								**RMSE**							
Dataset	DCT	DFT	FB	HW	LRHS-CP	LRHS-CP-S	LRHS-T	LRHS-T-S	DCT	DFT	FB	HW	LRHS-CP	LRHS-CP-S	LRHS-T	LRHS-T-S
BART	0.0486	0.0258	0.0269	**0.0213**	0.0217	**0.0215**	0.0235	0.0234	0.0782	**0.0468**	0.0492	0.0496	0.0493	**0.0488**	0.0496	0.0493
BART-1Y	0.0342	0.0255	0.0251	**0.0171**	**0.0158**	0.0172	0.0191	0.0199	0.0539	0.0422	0.0413	0.0333	**0.0318**	**0.0326**	0.0357	0.0362
Electricity-75	0.0598	0.0460	0.0413	0.0409	0.0410	0.0405	**0.0381**	**0.0386**	0.0770	0.0602	0.0545	0.0557	0.0552	0.0545	**0.0518**	**0.0524**
Electricity-85	0.0556	0.0415	0.0381	0.0358	0.0363	0.0354	**0.0350**	**0.0350**	0.0714	0.0544	0.0503	0.0492	0.0492	0.0482	**0.0473**	**0.0474**
Kaggle Energy	0.0770	**0.0464**	0.0565	0.0533	**0.0505**	0.0527	0.0508	0.0519	0.0939	**0.0601**	0.0717	0.0661	**0.0628**	0.0655	0.0651	0.0663
Kaggle Energy-1Y	0.0564	0.0468	0.0518	0.0488	0.0485	0.0475	**0.0454**	**0.0456**	0.0754	0.0618	0.0680	0.0660	0.0642	0.0637	**0.0610**	**0.0611**
Traffic-75	0.0420	0.0326	0.0278	**0.0233**	0.0247	**0.0232**	0.0267	0.0263	0.0620	0.0516	0.0481	0.0469	**0.0465**	**0.0452**	0.0478	0.0476
Traffic-85	0.0410	0.0314	0.0260	**0.0203**	**0.0210**	**0.0203**	0.0244	0.0244	0.0573	0.0468	0.0426	0.0391	**0.0388**	**0.0382**	0.0419	0.0419

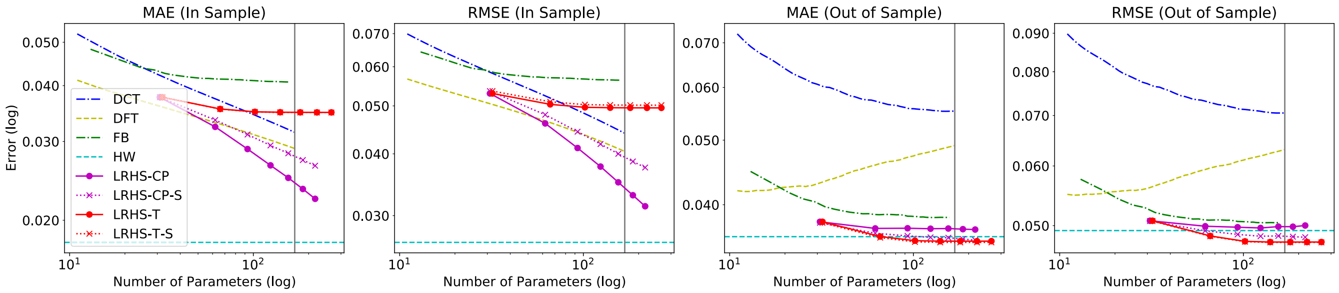


FIGURE 5. Comparison of errors vs. the number of parameters required for each model in the Electricity-75 dataset. The vertical dashed line denotes the number of parameters required by Holt-Winters and other classical decompositions. Both axes are logarithmic for ease of presentation.

For this task, we use Tensorly [65] for tensor decompositions and SciPy [66] for discrete transforms with Python programming language, without any explicit regularization.

3) EXPERIMENT RESULTS

For all baselines and LRHS variants, we run each algorithm for a variety of parameter dimensionality settings, and report the best generalization performance in Table 1. Our approaches generally yield favorable results, often matching and outperforming the Holt-Winters approach which appears to be the best performer among baselines. While short-term forecasting results on BART and Energy datasets are comparable to other methods, LRHS clearly outperforms baselines in one year forecasts. The latter result is especially important since these experiment settings include the longest forecast horizon (a year), and *three* hierarchical seasonality patterns (daily, weekly, yearly), implying that LRHS becomes more useful as seasonal patterns get more intricate.

In Figure 5, we plot the performance of different algorithms with varying number of parameters, corresponding to the number of components in discrete transform and FB methods, and varying tensor rank in our methods. We present results for the Electricity-75 dataset, where we can observe that tensor-based methods not only outperform baselines out of sample, but also do so with lower number of parameters, speaking to the appropriateness of tensor formalism in compressing/representing hierarchical seasonality. Similar figures for all seven other datasets can be found in the supplementary material.

It is especially significant that these favorable results are obtained in a univariate setting, without exploiting the advantage of LRHS being easily generalizable to multivariate data. Accordingly, we now move on to multivariate experiments, where we first focus on LRHS/BLRHS's ability to extract meaningful patterns from large multivariate datasets, and then compare its accuracy in imputation tasks to its alternatives.

B. EXPERIMENTS WITH NEW YORK CITY YELLOW TAXI DATA

We now use BLRHS to analyze New York City Yellow Taxi dataset,[5] which is a record of the yellow taxi rides within New York City, including start and end locations (265 each), as well as time of the trip. We use the first six months (or 25 weeks) of data from 2020 and create a 5-mode count tensor of dimensions $265 \times 265 \times 24 \times 7 \times 25$ where the dimensions correspond to start and end locations, hour of the day, day of the week, and week of the year, respectively.

We model this dataset using BLRHS with a Tucker decomposition.[6] We conduct model selection among different ranks and hyperparameters using the evidence lower bound (ELBO). Our procedure selects the model with latent ranks $R = (4, 4, 5, 2, 2)$, with each rank corresponding to the dimensions in the order enumerated above. In this section and the next, we do not explicitly detrend the data and allow the cycle index dimension account for the long term changes in the dataset. Further details on our model selection procedure can be found in the supplementary material. All multivariate experiments have been implemented with the JAX framework in Python programming language [67].

[5] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[6] See supplementary material for experiments that conduct model selection across decomposition models, CP vs. Tucker.

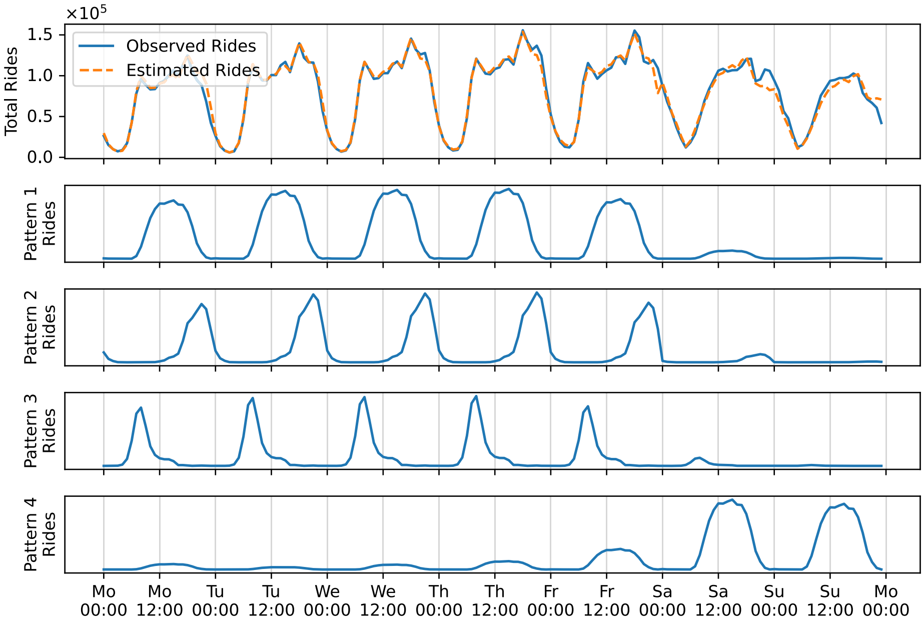


FIGURE 6. BLRHS breaks down the seasonalities that underlie observed patterns. Top figure includes total hourly pre-Covid rides in the NYC YT system and BLRHS's approximation. The four bottom figures plot the top weekly patterns that contribute to the overall number of rides. *y*-axes are not in scale for improved visualization.

1) EXTRACTED SEASONAL PATTERNS

Our results reveal a good reason for our model selection to have selected $R_{\text{week}} = 2$: this latent factor is used to characterize pre- and post-Covid patterns as shown in Figure 1. More specifically we see that

$$\widehat{P}(r_{\text{week}} = 1|\text{week}) \gg \widehat{P}(r_{\text{week}} = 2|\text{week}), \quad \forall \text{week} \in \{1, \ldots, 10\},$$

while this trend dramatically reverses after week 10. Note that the model also picks up on the slight, relative return to normal in Summer 2020, as $\widehat{P}(r_{\text{week}} = 2|\text{week})$ starts to fall back down towards the end of our time window, week 25. Having established that $r_{\text{week}} = 1$ and 2 roughly characterizes pre- and post-Covid patterns, we further investigate $\widehat{P}(\text{hour}, \text{day}|r_{\text{week}})$. This reveals the pre- and post-pandemic travel patterns as seen in Figure 1b, and discussed previously. Our results are in perfect alignment with those obtained by [26], who also infer Spring 2020 to be a landmark for seasonality regime change in NYC-YT data. That we can corroborate their results with a likelihood-based model selection and posterior inference further confirms the usefulness of our approach.[7]

We now further challenge BLRHS to break down the overall observed temporal dynamics to simpler multi-seasonal interaction patterns. For this, we first investigate the latent factor space, and find out the most dominant patterns by examining $\hat{P}(r_{\text{hour}} = i, r_{\text{day}} = j)$ for $i \in [R_{\text{hour}}], j \in [R_{\text{day}}]$. We then extract the daily-weekly patterns these latent dynamics correspond to by observing $\widehat{P}(\text{hour}, \text{day}|r_{\text{hour}}, r_{\text{day}})$. In Figure 6, we plot the four most dominant patterns extracted by BLRHS against total taxi ridership pre-Covid (week $<$ 10). BLRHS detects that overall ridership patterns are most strongly contributed by weekday afternoon rides (Pattern 1), weekday night rides (Pattern 2), morning commute rides (Pattern 3), and weekend day trips (Pattern 4) in order. Our results shows the various ways in which BLRHS can be used to make sense of large-scale temporal observations.

C. EXPERIMENTS WITH BART RIDERSHIP DATA

We now move on to an experiment that shows BLRHS can extract useful information from even larger datasets. For this experiment we use the full BART ridership dataset, which records the hourly number of passengers in the San Francisco Bay Area rail transportation system since 2011, in 2500 different pairwise routes among 50 stations. We model the data between 2011-2022 corresponding to 12 years, and we ignore the data from 2023 as it is incomplete at the time of writing of this article. Given the length of data, we add another seasonality component (yearly), thanks to the flexibility of BLRHS. The resulting tensor is of size $50 \times 50 \times 24 \times 7 \times 53 \times 12$, and contains a total ~ 1.18 billion rides. Even more challenging from a computational perspective, it is not sparse, with only $\sim$ 60% of its cells empty (as opposed to NYC YT data's $\sim 98.6\%$).

1) EXTRACTED SEASONAL PATTERNS

We again start with likelihood-based model selection among models with different ranks, and select a model with $R = (4, 4, 4, 2, 4, 2)$. Examining the results of inference

[7] BLRHS can ideally utilize an online procedure like [26], which we leave as an exciting future direction.

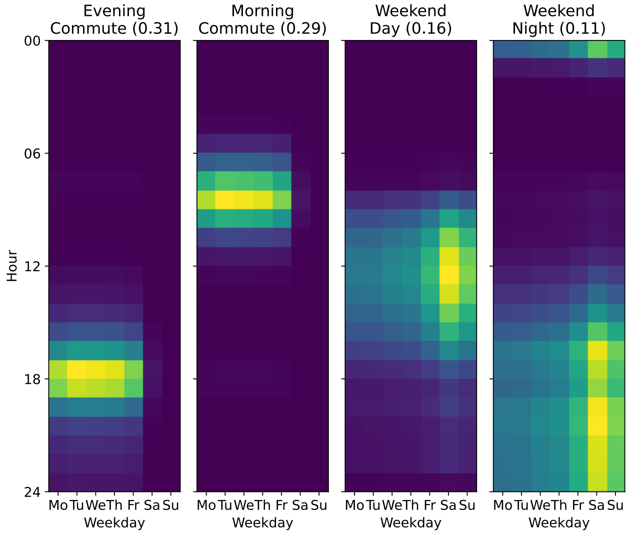


FIGURE 7. BLRHS extracts the daily-weekly latent patterns that contribute most to overall BART ridership. The four strongest are evening commute, morning commute, weekend day travel, and weekend night/entertainment travel. These constitute 0.31, 0.29, 0.16, and 0.11 of all travel respectively in expectation, computed through $\widehat{P}(r_{\text{hour}}, r_{\text{day}})$. For all remaining patterns this ratio is <0.1, lighter implies busier.

with this model exposes a number of interesting phenomena. We repeat our analysis about dominant daily-weekly patterns and present them in Figure 7, visualizing $\widehat{P}(\text{hour}, \text{day}|r_{\text{hour}}, r_{\text{day}})$ along with their strengths $\widehat{P}(r_{\text{hour}}, r_{\text{day}})$. The patterns are similar to those in NYC-YT, however take place slightly earlier in the day, given that these are records of mass transportation rides as opposed to private commercial trips. See the supplementary material for more details on model selection.

More interestingly, now that we have extracted them, BLRHS allows us to examine these latent patterns' evolution through time. In Figure 8, we do exactly this, and see that with Covid-19 the relative prevalence of these different travel tendencies drastically change (while total ridership falls overall). Weekend night travel almost goes extinct by Spring and Summer 2020, not only in absolute terms but also in relation to other types of travel. In relative terms, the reduction of weekend night travel is not counteracted by the increase of weekday commute, but by increased weekend daytime travel. This somewhat counterintuitive pattern is possibly due to essential workers and travels by residents to parks within the city, while work-from-home measures keep weekday commute at bay. Notice that this is an examination of the *latent* multi-seasonal interactions and is thus not easily accessible through marginal statistics of the data.

Being able to map out these high level interactions further demonstrates the usefulness of BLRHS for making sense of very large temporal data, making full use of the inductive bias that lies at the heart of our approach. One last example of this is presented in Figure 9, which relates these latent patterns to multiple spatial components.

D. MISSING DATA IMPUTATION in MULTIVARIATE TIME SERIES USING BLRHS

We have seen that BLRHS allows scalable and interpretable inference as well as convenient model selection. In this section, we investigate whether it eschews accuracy for interpretability. Our results show that it does not: BLRHS either surpasses or performs on par with other models that target multiple seasonality. As described in Section II, one of the most relevant preceding work is by [34], who experiment with one seasonality vs. two nested seasonalities in a given spatiotemporal model, using Bayesian CP decomposition. Using their setting, we compare BLRHS with their results, as well as a more recent method by [68] who utilize a similar approach yet add a sparse tensor term for modeling outliers. See the supplementary material for the details of the variational procedure and how missing data are handled within it.

1) DATASET

The dataset used is the Guangzhou Traffic Data [69]: a multivariate time series dataset that includes two months of observations of traffic speeds at 214 different road segments in Guangzhou, China, measured with a frequency of 10 minutes, with data dimensions 214 × 144 × 61. We follow the authors' methodology by testing the imputation performance of our method when 10%, 20%, 30%, 40%, and 50% of the data are missing. The entries can be missing either randomly or in a time-correlated fashion (*i.e.* missing for a day at a time). The dataset has ~ 1.87 million entries, and the 1.29% of the entries are missing in the original dataset.

2) EXPERIMENTAL SETTING, METRICS, MODEL SELECTION

We compare our results with Bayesian Gaussian CP (BGCP) from [34] and with Bayesian Robust CP (BRCP) from [68]. Moreover, since we use the exact experimental setting of [34] including randomization,[8] we compare our results with baseline model performances included therein as well, using numerical results reported by the authors. We do this because these baselines can surpass BGCP and BRCP, albeit rarely. These alternative baselines include high accuracy low-rank tensor completion (HaLRTC) [70], SVD-combined tensor decomposition (STD) [71], and two simpler baselines daily average (DA; filling in missing data with daily averages over all observations), and a kNN algorithm (using the average of k nearest roads for the missing values). As performance metrics, we use MAPE and RMSE to be able to compare our approach with previous work. RMSE is as defined above, and MAPE is defined as:

$$\text{MAPE}(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{T}\sum_{t=1}^{T}\frac{|y_t - \widehat{y}_t|}{y_t}$$

We use a BLRHS model with CP decomposition. Both [34] and [68] provide their models' results across different hyperparameter settings and/or different data representations.

[8] https://github.com/mcgill-smart-transport/bgcp_imputation

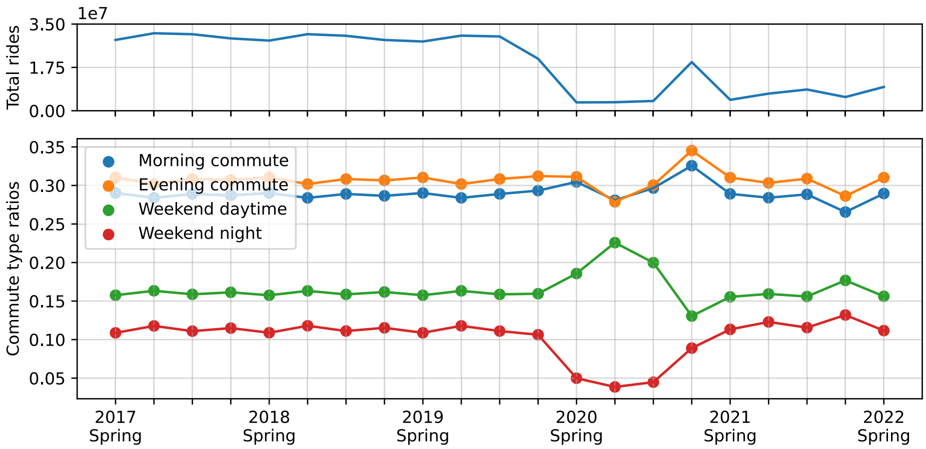


FIGURE 8. Total ridership dramatically decreases overall due to the Covid-19 pandemic (top). The relative prevalence of major leading patterns extracted in Figure 7 can be tracked through the years (bottom). In relative terms, our results show that during Covid the disappearing weekend night travel is replaced by weekend daytime trips, and not weekday commute.

TABLE 2. The comparison of BLRHS with other methods for missing data imputation. Best results in each column are in boldface. BLRHS is ours; BRCP is from [68], the rest is from [34].

Random Missing										
	10%		**20%**		**30%**		**40%**		**50%**	
Model	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
DA	0.1213	5.1778	0.1218	5.1905	0.1217	5.1977	0.1217	5.1993	0.1221	5.2113
kNN(10)	0.1303	5.1101	0.1314	5.1486	0.1322	5.1966	0.1333	5.2565	0.1356	5.3573
HaLRTC	0.0776	**3.1716**	0.0817	3.3231	0.0850	3.4748	0.0887	3.6143	0.0931	3.7730
BGCP	0.0795	3.4521	0.0798	3.4531	0.0799	3.4655	0.0801	3.4756	0.807	**3.5042**
BRCP	0.0832	3.5918	0.0836	3.6001	0.0838	3.6160	0.0837	3.6165	0.0844	3.6429
BLRHS (Ours)	**0.0731**	3.2320	**0.0742**	**3.2716**	**0.0765**	**3.3525**	**0.0780**	**3.4625**	**0.0801**	3.5373
Correlated Missing										
	10%		**20%**		**30%**		**40%**		**50%**	
Model	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
DA	0.1208	5.1128	0.1207	5.1983	0.1346	5.2591	0.1388	5.4405	0.1445	5.6516
kNN(13)	0.1342	5.1714	0.1340	5.1486	0.1322	5.1966	0.1333	5.2565	0.1356	5.3573
HaLRTC	0.1015	4.1322	0.1022	4.1716	0.1035	4.2372	0.1057	4.3232	0.1090	4.4472
BGCP	0.0980	4.1413	0.0984	4.1477	**0.0980**	**4.1857**	**0.1001**	**4.2881**	0.1030	4.4199
BRCP	0.0982	4.1352	**0.0980**	**4.1400**	0.0989	4.2197	0.1003	4.2909	**0.1022**	4.4497
BLRHS (Ours)	**0.0976**	**4.1314**	0.0984	4.2090	0.0998	4.2508	0.1021	4.3086	0.1045	**4.4191**

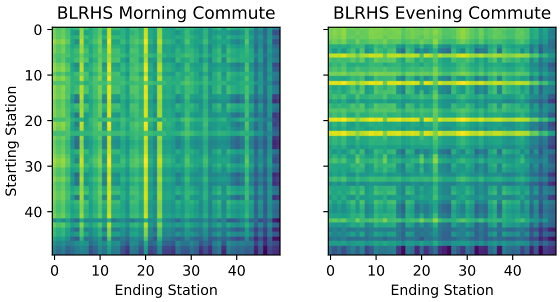


FIGURE 9. Examining how latent temporal patterns extracted by BLRHS can be related to spatial dimensions: A number of urban and commercial hubs (e.g. 6: Civic Center, 23: Powell Street) witness a large influx from more residential and suburban areas (e.g. 29: Balboa Park, 33: Walnut Creek), and the trend reverses at evening commute time. Lighter implies busier. Station names for all indices can be found in the supplementary material.

To provide a more rigorous challenge to our method, we use 1% of the data (selected from the uncensored cells) as validation set for model selection, and we compare our selected model with the *best-in-hindsight* versions of the results by [34] and [68].[9]

During model selection we search among $R = \{5, 25, 50, 150, 300, 450\}$, corresponding to a range of 1000-fold to 10-fold parameter decrease in the representation of the data. Since this is not a dataset of counts or frequencies but traffic speeds, the observed entries are almost always larger than 0. Thus, we apply a simple detrending scheme where we subtract the minimum of each time series from all observations before conducting inference with BLRHS, and add this to BLRHS's reconstructed output after inference. We present our results using a 3-order tensor data representation, discussion of BLRHS experiments up to 4 temporal dimensions

[9] Given that we use MAPE and RMSE for comparison, ELBO is of less utility here for model selection, given that it assumes a Poisson observation likelihood.

(5-order tensor), as well as further details on model selection can be found in the supplementary material.

3) EXPERIMENT RESULTS

We present our results in Table 2. For randomly missing observations the results are almost unanimous: save for a few exceptions our approach outperforms the best-in-hindsight alternative methods. This is very encouraging, and implies that BLRHS not only provides an interpretable analysis of seasonal patterns, but also an accurate one.

For correlated missing data, the results are more equivocal, in that the three methods, BLRHS, BGCP, and BRCP almost equally share the first place in different tasks. An important point to note is that since the metrics in question are RMSE and MAPE, the Gaussian likelihood used by [34] and [68] is arguably advantageous compared to BLRHS's Poisson likelihood. Given this point, and given that BLRHS is losing a portion of the data for model selection instead of providing results for all hyperparameters, we also take these results to be an encouraging sign, and consider improving accuracy of BLRHS in MAPE and RMSE further as an exciting future direction.

VI. CONCLUSION

We present a general framework for representing multiple periodic patterns in time series as an additive combination of underlying patterns. We connect our approach to the powerful formalism of low-rank tensor decompositions, which allows us to propose generally applicable tensor decomposition algorithms for estimating parsimonious representations of cyclical patterns in time series data. Moreover, our approach naturally extends to multivariate time series, and with the probabilistic version of our model we facilitate knowledge discovery not only through scaleable and accurate posterior inference, but also through likelihood-based model scoring. Although we have examined various applications of our work, these only represent a small selection of potential avenues that our approach can be extended to.

Exciting future directions include utilizing other observation likelihoods like Gaussian [34], using other models and/or inference methods for BLRHS [47], [58], more closely integrating the modeling of trend to our framework [72], using an additive sparse tensor term for outliers [68], involving ordinal dimensions [59], and capacity for handling seasonalities of statistically different nature such as those with heavy-tailed distributions [73]. Systematic application of our approach in fields such as energy management, logistics, and demand forecasting is another important future direction [8], [10], [11], [15]. As expressed before, our results have implications not only for matrix/tensor based methods, but also for methods based on classical time series decompositions, which are still widely used [74], [75], as well as deep learning based methods [76], [77]. Therefore, the integration of LRHS within such work constitutes another important future research direction. As our models and data get increasingly larger, the need for such parsimonious representations are only likely to increase. We hope that our work encourages further research in these exciting avenues.

ACKNOWLEDGMENT

The authors would like to thank A. Caner Türkmen for his generous contributions to the conceptualization, methodology, and writing of this article.

REFERENCES

[1] E. Ghysels and D. R. Osborn, *The Econometric Analysis of Seasonal Time Series*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[2] P. H. Franses and R. Paap, *Periodic Time Series Models*. Oxford, U.K.: Oxford Univ. Press, 2004.

[3] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1994.

[4] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Melbourne, VIC, Australia: OTexts, 2018.

[5] P. G. Gould, A. B. Koehler, J. K. Ord, R. D. Snyder, R. J. Hyndman, and F. Vahid-Araghi, "Forecasting time series with multiple seasonal patterns," *Eur. J. Oper. Res.*, vol. 191, no. 1, pp. 207–222, Nov. 2008.

[6] X. Bi, X. Tang, Y. Yuan, Y. Zhang, and A. Qu, "Tensors in statistics," *Annu. Rev. Statist. Appl.*, vol. 8, no. 1, pp. 345–368, 2021.

[7] J. Chang, J. He, L. Yang, and Q. Yao, "Modelling matrix time series via a tensor CP-decomposition," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 85, no. 1, pp. 127–148, Feb. 2023.

[8] M. Figueiredo, B. Ribeiro, and A. D. Almeida, "Analysis of trends in seasonal electrical energy consumption via non-negative tensor factorization," *Neurocomputing*, vol. 170, pp. 318–327, Dec. 2015.

[9] T. Ji, Y. Jiang, M. Li, and Q. Wu, "Ultra-short-term wind speed and wind power forecast via selective Hankelization and low-rank tensor learning-based predictor," *Int. J. Electr. Power Energy Syst.*, vol. 140, Sep. 2022, Art. no. 107994.

[10] M. Seeger, S. Rangapuram, Y. Wang, D. Salinas, J. Gasthaus, T. Januschowski, and V. Flunkert, "Approximate Bayesian inference in linear state space models for intermittent demand forecasting at scale," 2017, *arXiv:1709.07638*.

[11] S. Nejad, "Data-driven analysis of time of day pricing for residential consumers," M.S. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, May 2022.

[12] X. Chen, C. Zhang, X.-L. Zhao, N. Saunier, and L. Sun, "Nonstationary temporal matrix factorization for sparse traffic time series forecasting," 2022, *arXiv:2203.10651*.

[13] R. K. C. Chan, J. M. Lim, and R. Parthiban, "Missing traffic data imputation for artificial intelligence in intelligent transportation systems: Review of methods, limitations, and challenges," *IEEE Access*, vol. 11, pp. 34080–34093, 2023.

[14] L. Espín Noboa, F. Lemmerich, P. Singer, and M. Strohmaier, "Discovering and characterizing mobility patterns in urban spaces: A study of Manhattan taxi data," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 537–542.

[15] Y. Gao, L. T. Yang, J. Yang, D. Zheng, and Y. Zhao, "Jointly low-rank tensor completion for estimating missing spatiotemporal values in logistics systems," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1814–1822, Feb. 2023.

[16] P. Zhang, P. Ren, Y. Liu, and H. Sun, "Autoregressive matrix factorization for imputation and forecasting of spatiotemporal structural monitoring time series," *Mech. Syst. Signal Process.*, vol. 169, Apr. 2022, Art. no. 108718.

[17] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Temporal link prediction using matrix and tensor factorizations," *ACM Trans. Knowl. Discovery Data (TKDD)*, vol. 5, no. 2, pp. 1–27, Feb. 2011.

[18] M. R. D. Araujo, P. M. P. Ribeiro, and C. Faloutsos, "TensorCast: Forecasting with context using coupled tensors," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 71–80.

[19] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with Bayesian probabilistic tensor factorization," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2010, pp. 211–222.

[20] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa, ''Fast mining and forecasting of complex time-stamped events,'' in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2012, pp. 271–279.

[21] Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos, ''FUNNEL: Automatic mining of spatially coevolving epidemics,'' in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 105–114.

[22] M. Rogers, L. Li, and S. J. Russell, ''Multilinear dynamical systems for tensor time series,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2634–2642.

[23] H.-F. Yu, N. Rao, and I. S. Dhillon, ''Temporal regularized matrix factorization for high-dimensional time series prediction,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 847–855.

[24] K. Takeuchi, H. Kashima, and N. Ueda, ''Autoregressive tensor factorization for spatio-temporal predictions,'' in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1105–1110.

[25] X. Chen and L. Sun, ''Bayesian temporal factorization for multidimensional time series prediction,'' 2019, *arXiv:1910.06366*.

[26] K. Kawabata, S. Bhatia, R. Liu, M. Wadhwa, and B. Hooi, ''SSMF: Shifting Seasonal Matrix Factorization,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34. Red Hook, NY, USA: Curran Associates, 2021, pp. 3863–3873.

[27] Y.-L. Xie, P. K. Hopke, P. Paatero, L. A. Barrie, and S.-M. Li, ''Identification of source nature and seasonal variations of Arctic aerosol by positive matrix factorization,'' *J. Atmos. Sci.*, vol. 56, no. 2, pp. 249–260, Jan. 1999.

[28] T. Takahashi, B. Hooi, and C. Faloutsos, ''AutoCyclone: Automatic mining of cyclic online activities with robust tensor factorization,'' in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 213–221.

[29] Y. Matsubara, Y. Sakurai, and C. Faloutsos, ''The web as a jungle: Nonlinear dynamical systems for co-evolving online activities,'' in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 721–731.

[30] Y. Matsubara, Y. Sakurai, and C. Faloutsos, ''Non-linear mining of competing local activities,'' in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 737–747.

[31] X. Chen and L. Sun, ''Low-rank autoregressive tensor completion for multivariate time series forecasting,'' 2020, *arXiv:2006.10436*.

[32] X. Chen, M. Lei, N. Saunier, and L. Sun, ''Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12301–12310, Aug. 2022.

[33] H. Tan, Y. Wu, B. Shen, P. J. Jin, and B. Ran, ''Short-term traffic prediction based on dynamic tensor completion,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2123–2133, Aug. 2016.

[34] X. Chen, Z. He, and L. Sun, ''A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation,'' *Transp. Res. C, Emerg. Technol.*, vol. 98, pp. 73–84, Jan. 2019.

[35] Y. Wang, Y. Zhang, L. Wang, Y. Hu, and B. Yin, ''Urban traffic pattern analysis and applications based on spatio-temporal non-negative matrix factorization,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12752–12765, Aug. 2022.

[36] A. Mnih and R. R. Salakhutdinov, ''Probabilistic matrix factorization,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20. Red Hook, NY, USA: Curran Associates, 2007, pp. 1–8.

[37] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, 1st ed. Chichester, U.K: Wiley, 2009.

[38] T. G. Kolda and B. W. Bader, ''Tensor decompositions and applications,'' *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/07070111X

[39] R. A. Harshman, ''Foundations of the parafac procedure: Models and conditions for an 'exploratory' multimodal factor analysis,'' in *Proc. UCLA Work. Papers Phonetics*, Jan. 1970, pp. 1–84.

[40] J. D. Carroll and J.-J. Chang, ''Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart–Young' decomposition,'' *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[41] L. R. Tucker, ''The extension of factor analysis to three-dimensional matrices,'' in *Contributions to Mathematical Psychology*, N. Frederiksen and H. Gulliksen, Eds. New York, NY, USA: Holt, Rinehart and Winston, 1964, pp. 110–127.

[42] L. R. Tucker, ''Some mathematical notes on three-mode factor analysis,'' *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sep. 1966.

[43] L. De Lathauwer, B. De Moor, and J. Vandewalle, ''On the best rank-1 and rank-$(R1R2\ldots, R_N)$ approximation of higher-order tensors,'' *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1324–1342, 2000.

[44] J. Cohen, R. C. Farias, and P. Comon, ''Fast decomposition of large nonnegative tensors,'' *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 862–866, Jul. 2015.

[45] D. Nion and N. D. Sidiropoulos, ''Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar,'' *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5693–5705, Nov. 2010.

[46] M. Haardt, F. Roemer, and G. Del Galdo, ''Higher-order SVD-based subspace estimation to improve the parameter estimation accuracy in multidimensional harmonic retrieval problems,'' *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3198–3213, Jul. 2008.

[47] A. Schein, J. Paisley, D. M. Blei, and H. Wallach, ''Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts,'' in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2015, pp. 1045–1054, doi: 10.1145/2783258.2783414.

[48] S. Rabanser, O. Shchur, and S. Günnemann, ''Introduction to tensor decompositions and their applications in machine learning,'' 2017, *arXiv:1711.10781*.

[49] M. Ashraphijuo and X. Wang, ''Fundamental conditions for low-CP-rank tensor completion,'' *J. Mach. Learn. Res.*, vol. 18, no. 63, pp. 1–29, Jan. 2017.

[50] M. Ashraphijuo, V. Aggarwal, and X. Wang, ''Deterministic and probabilistic conditions for finite completability of low-Tucker-rank tensor,'' *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5380–5400, Sep. 2019.

[51] J. E. Cohen and U. G. Rothblum, ''Nonnegative ranks, decompositions, and factorizations of nonnegative matrices,'' *Linear Algebra Appl.*, vol. 190, pp. 149–168, Sep. 1993.

[52] P. Paatero and U. Tapper, ''Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,'' *Environmetrics*, vol. 5, no. 2, pp. 111–126, Jun. 1994.

[53] D. D. Lee and H. S. Seung, ''Algorithms for non-negative matrix factorization,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[54] Y.-X. Wang and Y.-J. Zhang, ''Nonnegative matrix factorization: A comprehensive review,'' *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

[55] W.-S. Chen, Q. Zeng, and B. Pan, ''A survey of deep nonnegative matrix factorization,'' *Neurocomputing*, vol. 491, pp. 305–320, Jun. 2022.

[56] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL, USA: CRC Press, Nov. 2013.

[57] A. Schein, M. Zhou, D. M. Blei, and H. Wallach, ''Bayesian Poisson Tucker decomposition for learning the structure of international relations,'' in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1–10.

[58] S. Yildirim, M. B. Kurutmaz, M. Barsbey, U. Simsekli, and A. T. Cemgil, ''Bayesian allocation model: Marginal likelihood-based model selection for count tensors,'' *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 560–573, Apr. 2021.

[59] N. Stoehr, B. J. Radford, R. Cotterell, and A. Schein, ''The Ordered Matrix Dirichlet for modeling ordinal dynamics,'' in *Proc. 26th Int. Conf. Artif. Intell. Statist.*, Dec. 2022, pp. 1–16.

[60] C. Hu, P. Rai, C. Chen, M. Harding, and L. Carin, ''Scalable Bayesian non-negative tensor factorization for massive count data,'' in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Aug. 2015, pp. 53–70.

[61] M. J. Beal and Z. Ghahramani, ''Variational Bayesian learning of directed graphical models with hidden variables,'' *Bayesian Anal.*, vol. 1, no. 1, pp. 1–4, Dec. 2006.

[62] M. J. Wainwright and M. I. Jordan, ''Graphical models, exponential families, and variational inference,'' *Found. Trends Mach. Learn.*, vol. 1, nos. 1–2, pp. 1–305, 2008.

[63] A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, L. Stella, A. Caner Türkmen, and Y. Wang, ''GluonTS: Probabilistic time series models in Python,'' 2019, *arXiv:1906.05264*.

[64] P. R. Winters, "Forecasting sales by exponentially weighted moving averages," *Manage. Sci.*, vol. 6, no. 3, pp. 324–342, Apr. 1960.

[65] J. Kossaifi, Y. Panagakis, A. Anandkumar, and M. Pantic, "TensorLy: Tensor learning in Python," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 925–930, 2019.

[66] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[67] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. (2018). *JAX: Composable transformations of Python+NumPy programs*. [Online]. Available: http://github.com/google/jax

[68] Y. Zhu, W. Wang, G. Yu, J. Wang, and L. Tang, "A Bayesian robust CP decomposition approach for missing traffic data imputation," *Multimedia Tools Appl.*, vol. 81, no. 23, pp. 33171–33184, Sep. 2022.

[69] X. Chen, Y. Chen, and Z. He, "Urban traffic speed dataset of Guangzhou, China," 2018. Accessed: Jul. 25, 2023. [Online]. Available: https://zenodo.org/record/1205229

[70] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.

[71] X. Chen, Z. He, and J. Wang, "Spatial–temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 59–77, Jan. 2018.

[72] C. Gong and Y. Zhang, "Urban traffic data imputation with detrending and tensor decomposition," *IEEE Access*, vol. 8, pp. 11124–11137, 2020.

[73] U. Simsekli, A. Liutkus, and A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2289–2293, Dec. 2015.

[74] S. J. Taylor and B. Letham, "Forecasting at scale," PeerJ, Corte Madera, CA, USA, Tech. Rep. e3190v2, Sep. 2017.

[75] K. Bandara, R. Hyndman, and C. Bergmeir, "MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns," 2021, *arXiv:2107.13462*.

[76] R. Sen, H.-F. Yu, and I. S. Dhillon, "Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 1–10.

[77] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021.