Supplementary Material for *Modeling Hierarchical Seasonality Through Low-Rank Tensor Decompositions in Time Series Analysis*

Melih Barsbey and Taylan Cemgil

This document constitutes the supplementary material for the paper *Modeling Hierarchical Seasonality Through Low-Rank Tensor Decompositions in Time Series Analysis*. Here, we provide further details regarding Bayesian low-rank hierarchical seasonality (BLRHS) in Section 1, and include additional details and results regarding our univariate (Section 2) and multivariate experiments (Section 3).

1 Further Details on Bayesian Low Rank Hierarchical Seasonality

1.1 Defining BLRHS with a Tucker Decomposition

We have defined BLRHS with a CP decomposition in the main paper. Here we define BLRHS with a Tucker decomposition.

Definition 1 (Bayesian low-rank hierarchical seasonality with a Tucker decomposition)**.** *Let the tensor* $\mathcal{M} \in \mathbb{R}_{\geq 0}^{P_1 \times \cdots \times P_N \times K}$ *be the multiple cyclical folding of an observed seasonality. Let* $\mathbf{m}_{r_n}^{(n)}$ *(resp.* $\mathbf{k}_{r_K}$*) be* r_n*'th (resp.* r_K*'th) column of the random factor matrix* $\mathbf{M}^{(n)}$*, (resp.* $\mathbf{K}$*). Bayesian low-rank hierarchical seasonality (BLRHS) utilizing a Tucker decomposition proposes the following generative model:*

$$
\begin{aligned}
P(\mathcal{M}|\widehat{\mathcal{M}}) &= \prod_{i_{[N+1]}} \text{Poisson}(\mathcal{M}(i_{[N+1]})|\widehat{\mathcal{M}}(i_{[N+1]})), \\
\widehat{\mathcal{M}} &= \lambda \sum_{r_1,\ldots,r_N,r_K} \mathcal{G}(r_1,\ldots,r_N,r_K) \left(\mathbf{m}_{r_1}^{(1)} \circ \cdots \circ \mathbf{m}_{r_N}^{(N)} \circ \mathbf{k}_{r_K}\right), \\
\lambda &\sim \text{Gamma}(a,b), \\
\mathbf{m}_{r_n}^{(n)} &\sim \text{Dirichlet}(\mathbf{1} \cdot \alpha(P_n \cdot R_n)),\ \forall r_n \in [R_n], \forall n \in [N] \\
\mathbf{k}_{r_K} &\sim \text{Dirichlet}(\mathbf{1} \cdot \alpha(K \cdot R_K)),\ \forall r_K \in [R_K] \\
\text{vec}(\mathcal{G}) &\sim \text{Dirichlet}(\mathbf{1} \cdot \alpha(R_1 \cdot \cdots \cdot R_N \cdot R_K)),
\end{aligned}
$$

where vec *is the tensor vectorization operator, Dirichlet distributions have flat priors with a 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 *and the corresponding random factor matrix* $\mathbf{I}$ *with columns* $\mathbf{i}_{r_I} \sim \text{Dirichlet}(\mathbf{1} \cdot \alpha(I \cdot R_I))$*.*

1.2 Variational Inference with BLRHS

We now detail the variational inference (VI) procedure we utilize in BLRHS as sketched out in Section IV-B1 of the main paper. We described *posterior inference* with respect to the parameters/latent variables of the model as a central inferential task.[1] Ignoring the hyperparameters for convenience, in the case of BLRHS

[1] In certain probabilistic modeling contexts, it might make sense to distinguish between model *parameters* and *latent variables*. This is not the case in this paper, and we use these terms interchangeably.

with CP decomposition, this corresponds to characterizing or approximating $P(\lambda, \mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(N)}, \mathbf{K}, \mathbf{w}|\mathcal{M})$. Mean-field VI (MFVI) [1] proposes to achieve this approximation with a factorized variational distribution:

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

In order to approximate the posterior, we minimize the KL-divergence between the actual and variational posteriors $D_{KL}(Q\|P)$. Collecting all latent variables under the term ζ, this implies the following equality:

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

with the rightmost term, evidence lower bound (ELBO) being a lower bound to the log-likelihood due to the nonnegativity of KL-divergence. ELBO is frequently used for model selection as an approximation of the marginal log-likelihood. For ease of presentation we assign the following notation to ELBO:

$$\mathcal{B}_P[Q] := -\mathbb{E}_Q[\log Q(\zeta) - \log P(\zeta, \mathcal{M})].$$

Under the mean-field assumption, a common procedure for maximizing the ELBO, sometimes referred to as coordinate ascent variational inference (CAVI), involves updating the parameters of variational distributions in an alternating fashion. This update scheme has convergence guarantee to a local optimum, and for a random variable Z the updates can be derived as

$$q(Z) \propto \exp \mathbb{E}_{\zeta \setminus Z}[P(\mathcal{M}, \zeta)], \tag{1}$$

where the expectation is with respect to the variational distributions of all latent variables except Z. The updates dictate the forms of the variational posteriors. Models with appropriately chosen conditional distributions lead to variational posteriors that have recognizable forms, making it possible to derive the updates analytically [1, 2].

As described in the main paper, among different probabilistic formulations of nonnegative tensor factorization, we emulate the approach by Bayesian Allocation Model (BAM) framework as presented in [3]. A central modeling choice by the authors is important for inference with BLRHS, which is augmenting the generative model with the auxiliary latent tensor $\mathcal{S}$ called the *allocation tensor*. As we will see below, this is very useful for obtaining tractable updates, and is therefore utilized in this paper as well. In describing inference with this extended model we will emulate some of [3]'s notation for convenience. Their formulation's generality will allow the procedure derived below to be equally applicable to the CP and Tucker versions of BLRHS we presented in the main paper and above. Our implementation of both, as well as source code for replicating our experiments and further exploring our modeling approach can be found in the paper's GitHub repository: `https://github.com/mbarsbey/lrhs`.

We let $V \subseteq [F]$ denote the set of all observed variables, and $\bar{V}$ the set of all unobserved (latent) variables. We let the total number of random variables equal $F = |V| + |\bar{V}|$. For example in univariate time series models, $|V| = N + 1$ for both CP and Tucker decompositions, while $|\bar{V}|$ is 1 and $N + 1$ for these two models respectively. We will apply this notation when indexing as well, e.g. $i_V := \{i_v : v \in V\}$, alongside our existing notation for indexing, e.g. $i_{[F]}$. Note that our use of N and F differ from [3]. For deriving MFVI in this augmented model, instead of $\mathcal{M}$ and $\widehat{\mathcal{M}}$, we will be working with their extended versions, $\mathcal{S}$ and $\widehat{\mathcal{S}}$. $\mathcal{S}$ is a latent tensor defined such that it includes all F random variables, observed and latent, as its modes. Crucially, $\mathcal{M} = \mathcal{S}_V$, where a subscript on $\mathcal{S}_A$ refers to the sum of $\mathcal{S}$ across all variables that are not in the variable set A. We now write BLRHS according to [3]'s extended model structure:

$$\mathcal{M} = \mathcal{S}_V \tag{2}$$

$$P(\mathcal{S}|\widehat{\mathcal{S}}) = \prod_{i_{[F]}} \text{Poisson}(\mathcal{S}(i_{[F]})|\widehat{\mathcal{S}}(i_{[F]})), \tag{3}$$

$$\widehat{\mathcal{S}}(i_{[F]}) = \lambda\theta(i_{[F]}), \tag{4}$$

$$\lambda \sim \text{Gamma}(a, b), \tag{5}$$

$$\theta(i_{[F]}) = \prod_{f=1}^{F} \theta_{f|\text{pa}(f)}(f|\text{pa}(f)), \tag{6}$$

$$\theta_{f|\text{pa}(f)} \sim \text{Dirichlet}(\mathbf{1} \cdot \alpha(f, \text{pa}(f))), \ \forall f \in [F]. \tag{7}$$

Note that in this definition, instead of fixing a particular conditional independence structure, as in $\sum_{r=1}^{R} w_r(\mathbf{m}_r^{(1)} \circ \cdots \circ \mathbf{m}_r^{(N)} \circ \mathbf{k}_r)$ in the original definition of BLRHS with CP, we explicitly invoke the factorization implied by the Bayesian network that characterizes the assumed decomposition, $\Theta := \{\theta_{f|\text{pa}(f)} : f \in [F]\}$, where $\text{pa}(f)$ refers to the parents of the variable f in the graphical model. This means that the derivation here applies for both CP and Tucker variants of BLRHS as well as other custom dependence structures [3]. The graphical models for both variants can be seen in Figure 1, which we repeat from the main paper for convenience. Also note that here we redefine $\alpha(f, \text{pa}(f)) := a/L(f, \text{pa}(f))$ such that $L(\cdot)$ is a function that takes the product of the cardinalities of all variables in its input, and $f, \text{pa}(f) := \{f\} \cup \text{pa}(f)$.

In this setting, $\theta_{f|\text{pa}(f)}$ are (latent) conditional probability tables. To reemphasize the connection between the notation here and in the main paper, in a BLRHS model with a CP decomposition and with daily and weekly seasonalities ($P_1 = 24, P_2 = 7$), observe that $\theta_{\text{hour}|r} = \mathbf{M}^{(1)}$, and $\theta_{\text{hour}|r=i}(\text{hour}|r = i) = P(\text{hour}|r = i) = \mathbf{m}_i^{(1)}$. Similarly, $\theta_{\text{day}|r} = \mathbf{M}^{(2)}$, and $\theta_{\text{day}|r=i}(\text{day}|r = i) = P(\text{day}|r = i) = \mathbf{m}_i^{(2)}$.

From this point onwards, our exposition closely follows [3] with notational modifications. We will provide the forms of the updates for the variational distributions, as well as the expression for the ELBO. For a detailed derivation of these updates we refer the reader to [3]. Having restructured our generative model, we state ELBO as below:

$$\begin{aligned}
\mathcal{B}_P[Q] &= -\mathbb{E}_Q \left[\log Q(\zeta) - \log P(\zeta, \mathcal{M})\right] && (8)\\
&= \mathbb{E}_Q \left[\log \frac{P(\mathcal{S}, \lambda, \Theta, \mathcal{M})}{Q(\mathcal{S}, \lambda, \Theta)}\right] && (9)\\
&= \mathbb{E}_Q \left[\log \frac{P(\mathcal{S}, \lambda, \Theta)\mathbb{I}(\mathcal{S}_V = \mathcal{M})}{Q(\mathcal{S}, \lambda, \Theta)}\right] && (10)
\end{aligned}$$

where $\mathbb{I}(\cdot)$ is 1 if its argument is true, 0 otherwise, and (10) is due to $\mathcal{M} = \mathcal{S}_V$. In [3] the following factorization scheme is assumed for the variational posterior[2]: $Q(\mathcal{S}, \lambda, \Theta) = q(\mathcal{S})q(\lambda, \Theta)$. Based on (1), the CAVI updates in this construction correspond to:

$$\begin{aligned}
q(\mathcal{S}) &\propto \exp\left(\mathbb{E}_{q(\lambda,\Theta)}\left[\log P(\mathcal{S}, \lambda, \Theta) + \mathbb{I}(\mathcal{S}_V = \mathcal{M})\right]\right), && (11)\\
q(\lambda, \Theta) &\propto \exp\left(\mathbb{E}_{q(\mathcal{S})}\left[\log P(\mathcal{S}, \lambda, \Theta) + \mathbb{I}(\mathcal{S}_V = \mathcal{M})\right]\right). && (12)
\end{aligned}$$

Computing these expectations imply the following forms for the variational distributions, where we write the parameters of the variational distributions explicitly as subscripts when relevant:

$$\begin{aligned}
q_\sigma(\mathcal{S}) &= \prod_{i_V} \text{Multinomial}\big(\mathcal{S}(:, i_V); \mathcal{M}(i_V), \sigma(:, i_V)\big), && (13)\\
q(\lambda, \Theta) &= q(\lambda)q(\Theta) = q(\lambda) \prod_{f=1}^{F} q(\theta_{f|\text{pa}(f)}), && (14)\\
q_\ell(\lambda) &= \text{Gamma}(\lambda; \ell, b + 1), && (15)\\
q_\tau(\theta_{f|\text{pa}(f)}) &= \prod_{i_{\text{pa}(f)}} \text{Dirichlet}(\theta_{f|\text{pa}(f)}(: |\text{pa}(f)); \tau_{f,\text{pa}(f)}(:, \text{pa}(f))), \ \forall f \in [F] && (16)\\
& && (17)
\end{aligned}$$

where σ is a tensor that has same dimensionality with $\mathcal{S}$ and τ is a set of vectors/matrices/tensors with the dimensions identical to those in Θ. Given a set of variables $A \subset [F]$, $\mathcal{S}(:, i_A)$, refers to the fiber/slice/subtensor that extends along the indices $f \in [F] : f \notin A$ for the given i_A. During inference, the following variational

[2] Although the assumption $Q(\mathcal{S}, \lambda, \Theta) = q(\mathcal{S})q(\lambda, \Theta)$ can be more accurately termed *structured mean-field*, as seen below $q(\mathcal{S})$ and $q(\lambda, \Theta)$ further factorize to produce a fully factorized variational posterior.

parameter updates are made in an alternating fashion:

$$\sigma(i_{[F]}) \leftarrow \frac{\exp\left(\sum_{f=1}^{F} \mathbb{E}_Q\left[\log \theta_{f|\mathrm{pa}(f)}(i_f|i_{\mathrm{pa}(f)})\right]\right)}{\sum_{i_{\bar{V}}} \exp\left(\sum_{f=1}^{F} \mathbb{E}_Q\left[\log \theta_{f|\mathrm{pa}(f)}(i_f|i_{\mathrm{pa}(f)})\right]\right)}, \ \forall i_{[F]} \tag{18}$$

$$\ell \leftarrow a + \mathbb{E}_Q\left[\mathcal{S}_+\right] \tag{19}$$

$$\tau_{f,\mathrm{pa}(f)} \leftarrow \bar{\alpha}(f, \mathrm{pa}(f)) + \mathbb{E}_Q\left[\mathcal{S}_{f,\mathrm{pa}(f)}\right], \tag{20}$$

where $\mathcal{S}_+ = \sum_{i_{[F]}} \mathcal{S}(i_{[F]})$, $\mathcal{S}_{f,\mathrm{pa}(f)}$ corresponds to the sum of $\mathcal{S}$ over $[F]\backslash(f, \mathrm{pa}(f))$, and $\mathcal{S}_{f,\mathrm{pa}(f)}(:, i_{\mathrm{pa}(f)})$ corresponds to the slice/fiber/subtensor of this sum selected by $i_{\mathrm{pa}(f)}$. For a set of variables A, $\bar{\alpha}_A$ is tensor that have the dimensionality of the members of A, and equals $\mathbf{1} \cdot \alpha(A)$ with α (re)defined as in Definition 1 in Appendix 1.2. The expectations included above can be written as

$$\mathbb{E}_Q\left[\lambda\right] = \ell/(b+1), \tag{21}$$

$$\mathbb{E}_Q\left[\log \lambda\right] = \psi(a + \mathbb{E}_Q\left[\mathcal{S}_+\right]) - \log(b+1), \tag{22}$$

$$\mathbb{E}_Q\left[\theta_{f|\mathrm{pa}(f)}(i_f|i_{\mathrm{pa}(f)})\right] = \psi(\tau_{f,\mathrm{pa}(f)}(i_{f,\mathrm{pa}(f)})) - \psi(\tau_{\mathrm{pa}(f)}(i_{\mathrm{pa}(f)})), \tag{23}$$

$$\mathbb{E}_Q\left[\mathcal{S}_{f,\mathrm{pa}(f)}\right] = \left(\mathbb{E}_Q\left[\mathcal{S}\right]\right)_{f,\mathrm{pa}(f)}, \tag{24}$$

$$\mathbb{E}_Q\left[\mathcal{S}(i_{[F]})\right] = \mathcal{M}(i_V)\sigma(i_{[F]}), \tag{25}$$

with ψ corresponding to the *digamma* function. The update procedure described above requires $\mathbb{E}_Q\left[\mathcal{S}_{f,\mathrm{pa}(f)}\right]$, i.e. *marginal* statistics of $\mathbb{E}_Q\left[\mathcal{S}\right]$. This in turn requires access to $\sigma(i_{[F]}), \forall i_{[F]}$. While explicitly storing σ might be potentially prohibitive, since $\sigma(i_{[F]})$ respects the factorization implied by the graphical model in question, we can compute the marginal statistics $\mathbb{E}_Q\left[\mathcal{S}_{f,\mathrm{pa}(f)}\right]$ with the junction tree algorithm, which drastically reduces the memory requirements for the overall procedure [3].

Any missing entries are treated as Poisson-distributed latent variables, whose variational distributions $q_{\widehat{\mathcal{M}}}$ are also updated at each iteration of MFVI. Given the set of all missing entry indices Ω, we have:

$$\widehat{\mathcal{M}}(i'_V) \leftarrow \sum_{i'_{\bar{V}}} \exp\left(\mathbb{E}_Q\left[\log \lambda\right] + \sum_{f=1}^{F} \mathbb{E}_Q\left[\log \theta_{f|\mathrm{pa}(f)}(i'_f|i'_{\mathrm{pa}(f)})\right]\right), \forall i'_V \in \Omega. \tag{26}$$

Accordingly, (13) and (25) are modified such that $\widehat{\mathcal{M}}(i'_V)$ replace $\mathcal{M}(i'_V)$ for missing entries. After training, the variational posterior can be used to infer the missing entries, $\mathcal{M}(i'_V) \approx \mathbb{E}_Q\left[\mathcal{M}(i'_V)\right] = \widehat{\mathcal{M}}(i'_V), \forall i'_V \in \Omega$.

The updates can be stopped as the ELBO converges to a local optimum. Given that it has no guarantees to converge to a *global* minimum, it is a good practice to start this procedure with different random initializations and pick the solution with highest ELBO. The resulting ELBO can be and is frequently used for model selection as described above. As [3] show, expanding (10) reveals the following expression for ELBO:

$$\begin{aligned}
\mathcal{B}_P[Q] =& a\log b - (a + \mathbb{E}_Q\left[\mathcal{S}_+\right])\log(b+1) + \log\Gamma(a + \mathbb{E}_Q\left[\mathcal{S}_+\right]) - \log\Gamma(a) \\
&+ \sum_{f=1}^{F} \sum_{i_{\mathrm{pa}}(f)} \log\Gamma(\bar{\alpha}_{\mathrm{pa}(f)}(i_{\mathrm{pa}(f)})) - \sum_{f=1}^{F} \sum_{i_{f,\mathrm{pa}(f)}} \log\Gamma(\bar{\alpha}_{f,\mathrm{pa}(f)}(i_{f,\mathrm{pa}(f)})) \\
&- \sum_{f=1}^{F} \sum_{i_{\mathrm{pa}}(f)} \log\Gamma(\bar{\alpha}_{\mathrm{pa}(f)}(i_{\mathrm{pa}(f)}) + \mathbb{E}_Q\left[\mathcal{S}_{\mathrm{pa}(f)}\right](i_{\mathrm{pa}(f)})) \\
&+ \sum_{f=1}^{F} \sum_{i_{f,\mathrm{pa}(f)}} \log\Gamma(\bar{\alpha}_{f,\mathrm{pa}(f)}(i_{f,\mathrm{pa}(f)}) + \mathbb{E}_Q\left[\mathcal{S}_{f,\mathrm{pa}(f)}\right](i_{f,\mathrm{pa}(f)})) \\
&- \sum_{i_V \notin \Omega} \log\Gamma(\mathcal{M}(i_V) + 1) - \sum_{i_{[F]}} \mathbb{E}_Q\left[\mathcal{S}(i_{[F]})\right] \log\sigma(i_{[F]}).
\end{aligned} \tag{27}$$

As expressed above, the exposition above is mostly a restatement of the findings of [3] adapted for our notation. See [3] for a more in-depth treatment of variational inference under these modeling assumptions, as well as a sequential Monte Carlo based inference scheme that relies on a sequential interpretation of the same model.

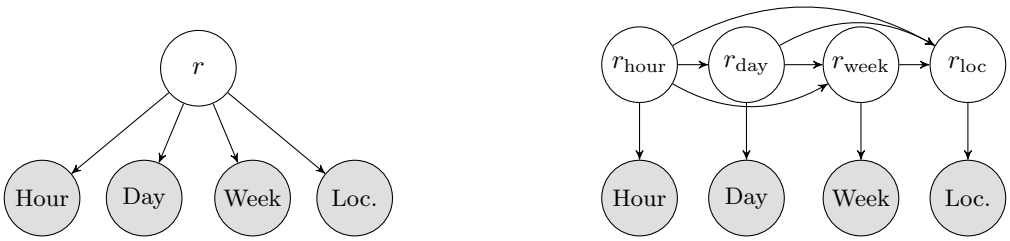


Figure 1: BLRHS utilizing CP and Tucker decompositions, modeling multivariate time series with temporal observations taken at different locations. Figure repeated from the main paper for convenience.

1.3 BLRHS with Continuous Nonnegative Data

A potential limitation of utilizing BAM for BLRHS is the former's assumption of Poisson likelihood, limiting the model to *discrete* nonnegative data. In practice this is not necessarily an important hindrance for continuous nonnegative datasets with large observations since the data can be interpreted as the sum of discrete values and element-wise independent uniform noise. So, given $\mathcal{M} \in \mathbb{R}_{\geq 0}^{P_1 \times \cdots \times P_N \times K}$, we can let $\mathcal{M} = \widetilde{\mathcal{M}} + \mathcal{E}$, where $\widetilde{\mathcal{M}} \in \mathbb{Z}_{\geq 0}^{P_1 \times \cdots \times P_N \times K}$, and $\mathcal{E}(i_{[F]}) \sim \mathcal{U}(0,1), \forall i_{[F]}$. This allows recovering $\widetilde{\mathcal{M}}$ with an element-wise flooring operation.

1.4 Hyperparameters of BLRHS

Before providing additional background and results regarding our experiments, we detail model selection for model hyperparameters, a and b. We set b using an empirical Bayesian approach following [3], where we let $b = a/\mathcal{S}_+ = a/\mathcal{M}_+$, leading to the expectation of Gamma distribution to be the sum of all data. In the case of missing observations in $\mathcal{M}$, we set $b = a/\Pi(\mathcal{M})_+$, where Π is a projector that sets missing entries to 0. Note that as in [3], the α also depends on a, where the resulting Bayesian Dirichlet (likelihood) Equivalent Uniform prior assigns equal likelihoods to models within the same Markov equivalence class. This leaves a as the only free hyperparameter. For all experiments where model selection was conducted, the hyperparameter a was scanned for between $1e2$ and $1e6$. We specify the a hyperparameters we ended up selecting for each experiment separately below, along with model selection results on ranks and/or decompositions.

2 Additional Details and Results for Univariate Experiments

2.1 Discrete Transforms and Fourier Basis Regression

Discrete Fourier transform (DFT) [4, Ch. 4] and discrete cosine transform (DCT) [5, 6] are frequently used to recover periodic patterns in signals in signal processing and time-series analysis. This is especially useful when one has no prior knowledge of the periods of the seasonalities present in the data, since one can use only the frequencies that have the most power to reconstruct the periodic signal $\mathbf{s}$.

In contrast, when the periods in data are known beforehand, one can regress the time series on a set of known orthogonal Fourier components [7]. For example, for periods $P_1, P_2, \cdots, P_N$, and defining $\bar{P}_n = \prod_{i=1}^{n} P_i$ one can assume

$$s_t = \sum_{n=1}^{N} \sum_{r=1}^{R} \left(a_{(n,r)} \sin\left(\frac{2\pi r t}{\bar{P}_n}\right) + b_{(n,r)} \cos\left(\frac{2\pi r t}{\bar{P}_n}\right) \right).$$

Then, $a_{(n,r)}, b_{(n,r)}$ can be inferred in a straightforward manner *e.g.*, via ordinary least squares. This is an approach taken within many industrial forecasting systems including Prophet [8], usually in combination with other methods for handling trend-cycle and outlier observations.

2.2 Additional Parameter and Performance Analyses

In the main paper, we compared the parameter count and performance of the algorithms tested on the Electricity-75 dataset. In Figure 2 we present the same analysis for the remaining datasets. For all, the vertical

dashed line denotes the number of parameters required by Holt-Winters and other classical decompositions. The x-axis denotes the number of free parameters in each method, and y-axis denotes error. Both axes are log-scaled for ease of presentation.

3 Additional Details and Results for Multivariate Experiments

3.1 Multivariate Experiments on Traffic-75 Dataset

We now conduct an analysis of the Traffic-75 dataset with BLRHS. As mentioned in the main paper, Traffic dataset records the occupancy rates of road segments in California, United States as the percentage of total capacity [9]. For our multivariate experiments, we model one year of this dataset, which corresponds to a tensor with dimensions $508 \times 24 \times 7 \times 52$, with the indices corresponding to the time series, hour of the day, day of the week, and week of the year. For this experiment we scan through various CP and Tucker decomposition structures. For CP decompositions we scan through $R = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and for Tucker decompositions through $R = (R_i, R_i, R_i, R_j)$, where $R_i \in \{2, 3, 4\}$ and $R_j \in \{2, 3, 4, 5, 6, 7, 8\}$. As opposed to missing data imputation experiments, we replace missing data with segment-based averages as a preprocessing step in order to avoid the additional computational cost of conducting MFVI with missing data, since the missing data only accounts for < 0.01 of the cells. We scale all observations by multiplying with 10 and apply the discretization described in Appendix 1.3 to emulate counts. The selected model is a CP decomposition with $R = 7$ and $a = 2000$.

3.1.1 Results

To examine what different seasonality patterns are captured by different latent variable levels, we visualize the model's estimates $\widehat{P}(\text{day}, \text{hour}|r)$ in Figure 3. Note that the different patterns that are picked up by separate latent dimensions in the form of conditional probabilities correspond to qualitatively different seasonal behavior in vehicle use. For example, the latent variable values $r = 1, 2, 3, 4$ correspond to morning and evening rush hours, where vehicle use increases due to commuters. Interestingly, the model allocates two latent dimensions to the morning and evening rush hours each. This turns out to be appropriate, as the model picks up on the changes in the highway occupancy due to *daylight saving time*. $r = 5$ seems to capture early commute at the end of the week, and the remaining $r = 6, r = 7$ capture weekend traffic, in addition to some of the remaining weekday traffic. These results support those in Section V of the main paper in demonstrating that BLRHS can be used to extract interpretable information from large temporal datasets.

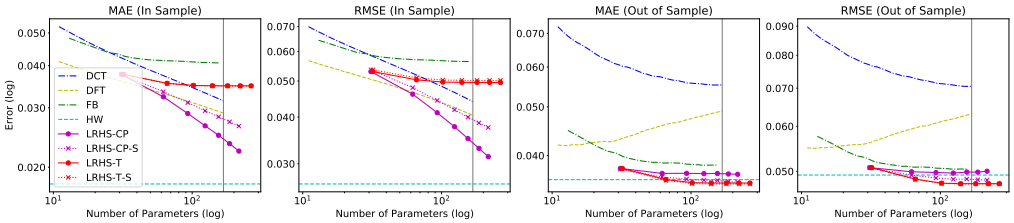
3.2 Additional Details on NYC YT Experiments

For NYC Yellow Taxi data, which has the dimensions $265 \times 265 \times 24 \times 7 \times 25$, we select among Tucker decomposition models with latent ranks $R = (R_{\text{loc}}, R_{\text{loc}}, R_{\text{time}}, R_{\text{time}}, R_{\text{week}})$, where $R_{\text{loc}} \in \{2, 3, 4\}$, $R_{\text{time}} \in \{2, 3, 4, 5\}$, and $R_{\text{week}} \in \{2, 3\}$. The model with the highest ELBO has $R = (4, 4, 5, 2, 2)$ and $a = 1000$. Our experiments show that further increasing the latent dimensions increase ELBO *only* for spatial dimensions, which are not of interest for this article. Therefore we upper bound R_{loc} to 4 for more expedient inference.

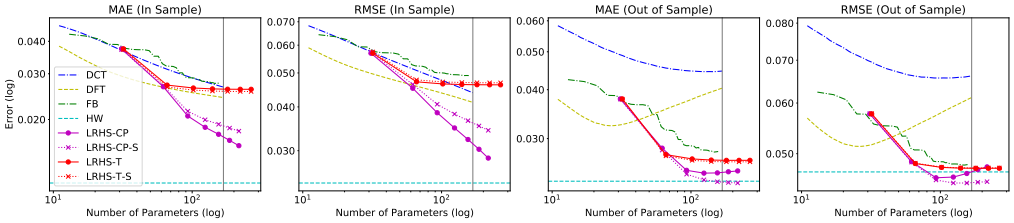
3.3 Additional Details on BART Experiments

For BART ridership data the observed tensor dimensions are $50 \times 50 \times 24 \times 7 \times 53 \times 12$. We select among Tucker decomposition models with latent ranks $R = (R_{\text{loc}}, R_{\text{loc}}, R_{\text{time}}, R_{\text{time}}, R_{\text{time}}, R_{\text{year}})$, where $R_{\text{loc}} \in \{2, 3, 4\}, R_{\text{time}} \in \{2, 3, 4, 5\}$, and $R_{\text{year}} \in \{2, 3\}$. The chosen model has $R = (4, 4, 4, 2, 4, 2)$ and $a = 1000$.

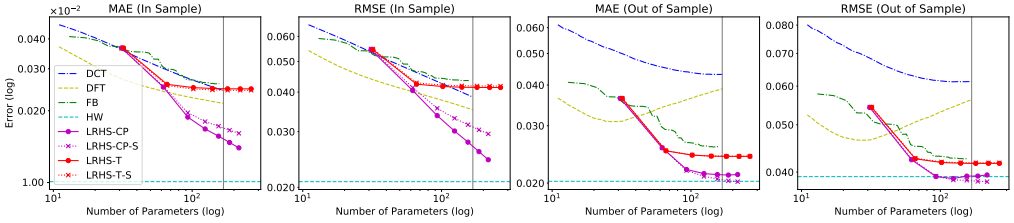
Note that we use a different cardinality for week of the year in multivariate experiments (53) compared to the univariate experiments (52). This is due to the irregularity of weeks according to years. ISO 8601's week date system prescribes some years to include an additional "leap week", making the total 53 for these select years. Although assuming 52 weeks and conducting cyclical folding is acceptable for the purposes of the univariate experiments, in the multivariate experiments we opt to observe consistency with calendar years, so that our Covid-related results are easier to interpret.



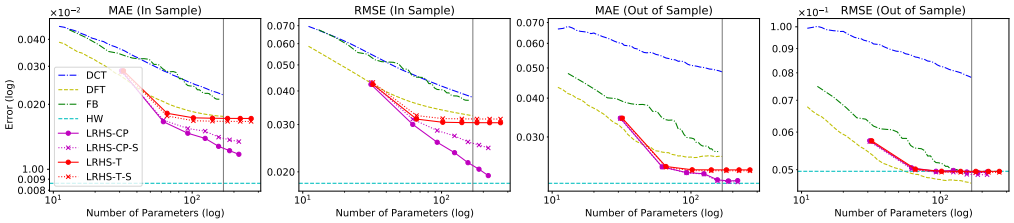
(a) Comparison of errors vs. the number of parameters required in the Electricity-85 dataset.



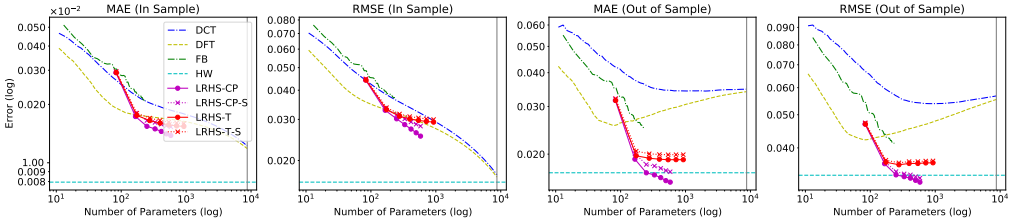
(b) Comparison of errors vs. the number of parameters required in the Traffic-75 dataset.



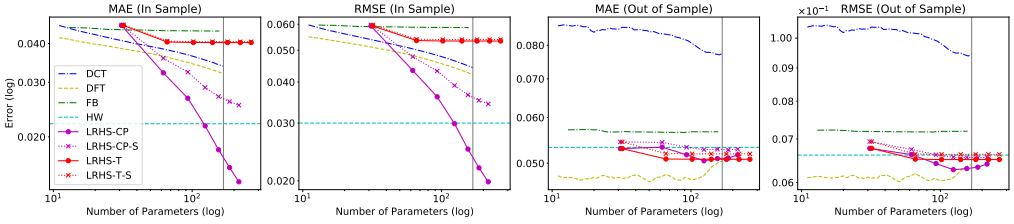
(c) Comparison of errors vs. the number of parameters required in the Traffic-85 dataset.



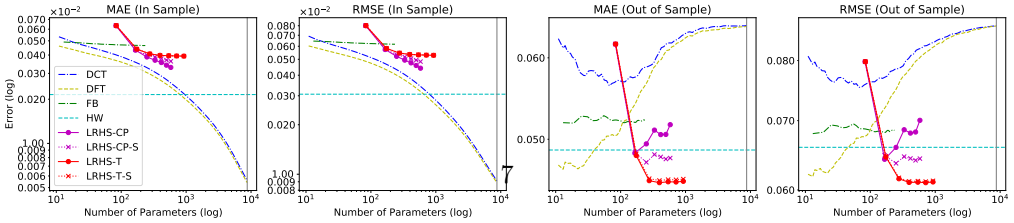
(d) Comparison of errors vs. the number of parameters required in the BART dataset.



(e) Comparison of errors vs. the number of parameters required in the BART-1Y dataset.



(f) Comparison of errors vs. the number of parameters required in the Energy dataset.



(g) Comparison of errors vs. the number of parameters required in the Energy-1Y dataset.

Figure 2: Comparison of errors vs. the number of parameters for different datasets in univariate experiments.

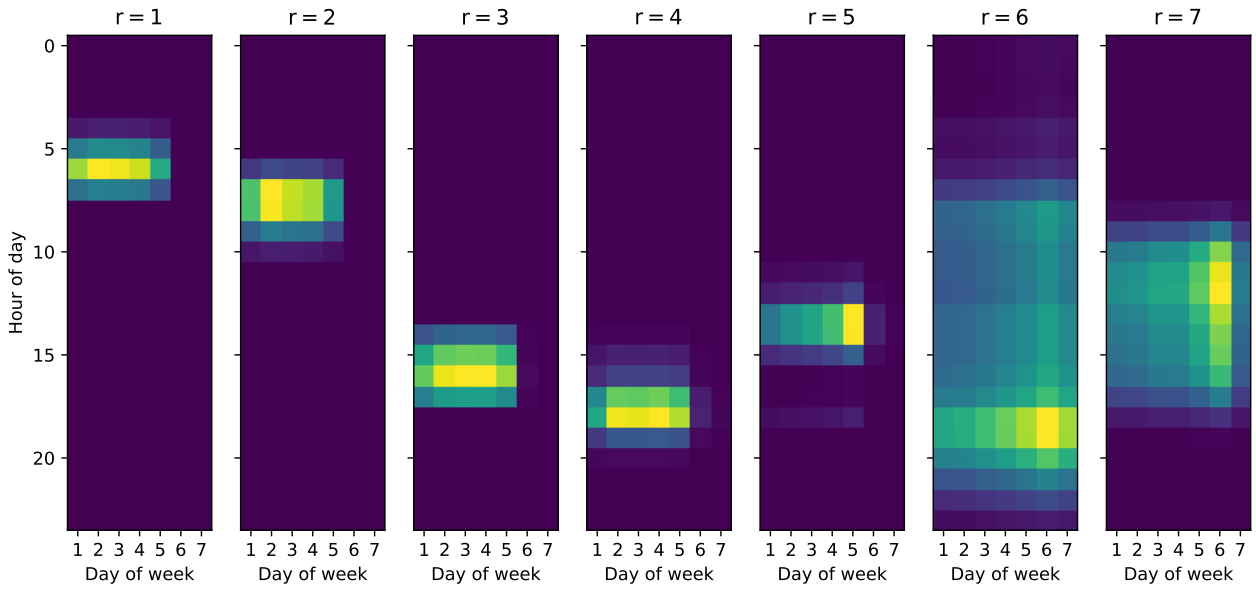


Figure 3: Estimates $\widehat{P}(\text{day}, \text{hour}|r)$ for all latent variable values r in Traffic-75 dataset. The model captures different cyclic patterns of human commute behavior.

Table 1: BART station mappings used in the main paper for spatiotemporal analysis.

Code	Station	Code	Station	Code	Station	Code	Station	Code	Station
12TH	0	16TH	1	19TH	2	24TH	3	ASHB	4
BAYF	5	CIVC	6	COLS	7	CONC	8	DALY	9
DBRK	10	DELN	11	EMBR	12	FRMT	13	FTVL	14
GLEN	15	LAFY	16	LAKE	17	MCAR	18	MLBR	19
MONT	20	PHIL	21	PITT	22	POWL	23	ROCK	24
SBRN	25	SSAN	26	UCTY	27	WOAK	28	BALB	29
DUBL	30	NBRK	31	ORIN	32	WCRK	33	NCON	34
RICH	35	SANL	36	COLM	37	HAYW	38	SHAY	39
CAST	40	PLZA	41	SFIA	42	WDUB	43	ANTC	44
WARM	45	PCTR	46	OAKL	47	BERY	48	MLPT	49

Lastly, the mapping used between station names and station indices while presenting BART readership analyses in the main paper's Figure 9 can be found in Table 1.

3.4 Additional Details on Missing Data Imputation Experiments

For the Guangzhou traffic data in the imputation experiments, which has the dimensions $214 \times 61 \times 144$, we use a CP decomposition and search among $R = \{5, 25, 50, 150, 300, 450\}$, corresponding to a range of 1000-fold to 10-fold parameter decrease. The model selection is conducted based on a held-out dataset that corresponds to 0.01 of the total cells, from among the uncensored data. The held-out dataset is used for early stopping the variational inference procedure, and the RMSE of the best epoch is used to compare models between latent ranks and hyperparameters. The selected hyperparameters can be seen in Table 2.

The fact that the dataset includes traffic speeds with real nonnegative values poses a problem given the Poisson likelihood assumed in the generative model. In this case, we avoid a discretization scheme described in Appendix 1.3 since it would prevent comparing the performance of BLRHS with previous related work directly. Therefore, in this case we proceed without any preprocessing, and apply the MFVI updates as described in Appendix 1.2. This renders marginal likelihood and ELBO undefined, preventing a likelihood-based model comparison. However, as described in the main paper, this is not necessarily prohibitive in this particular

problem, as model selection through performance on a held-out validation set is more appropriate here due to the performance metrics utilized.

Table 2: Selected hyperparameters for the missing data imputation experiments.

	Random Missing				
Missing Ratio	0.1	0.2	0.3	0.4	0.5
R	450	450	450	450	450
a	6×10^4	6×10^4	7.5×10^4	6×10^4	6×10^4
	Correlated Missing				
Missing Ratio	0.1	0.2	0.3	0.4	0.5
R	35	45	40	40	40
a	1.1×10^4	1×10^4	5×10^4	12.5×10^4	12.5×10^4

Similar to [10] we experiment with different data representations, and go up to 4 temporal indices creating a 5-order tensor with dimensions $214 \times 6 \times 24 \times 7 \times 9$. The cardinality 6 in the second mode is due to the observations having been made with 10 minute intervals. For this specific task, like [10] we find that a 3-order tensor with dimensions $214 \times 61 \times 144$ produces the best results, according to which we present our findings. This is not necessarily surprising given the nature of this data: there is no inherent reason for traffic speed to have an hourly seasonality. The opposite can be expected in other contexts. For example, in a dataset that records elevator use in a company, half and full hours might see increased travel between meeting rooms and offices due to meetings starting and ending, leading to a strong hourly seasonality.

References

[1] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2007.

[2] M. J. Beal and Z. Ghahramani, "Variational Bayesian learning of directed graphical models with hidden variables," *Bayesian Analysis*, vol. 1, Dec 2006.

[3] S. Yıldırım, M. B. Kurutmaz, M. Barsbey, U. Simsekli, and A. T. Cemgil, "Bayesian allocation model: Marginal likelihood-based model selection for count tensors," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, pp. 560–573, Apr 2021.

[4] R. G. Lyons, *Understanding Digital Signal Processing*. USA: Addison-Wesley Longman Publishing Co., Inc., 1st ed., 1996.

[5] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[6] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. USA: Academic Press Professional, Inc., 1990.

[7] A. C. Harvey and N. Shephard, "Structural time series models," *Handbook of Statistics*, vol. 11, pp. 261–302, 1993.

[8] S. J. Taylor and B. Letham, "Forecasting at scale," Tech. Rep. e3190v2, PeerJ Inc., Sep 2017.

[9] A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, *et al.*, "GluonTS: Probabilistic time series models in Python," *arXiv:1906.05264*, 2019.

[10] X. Chen, Z. He, and L. Sun, "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation," *Transportation Research Part C: Emerging Technologies*, vol. 98, pp. 73–84, Jan 2019.