Bayesian Matrix Factorization for Electricity Load Imputation

Xinqi Li, and Ke Li, *Senior Member, IEEE*

***Abstract*—The missing values in electricity load are a critical issue in grid applications. The electricity demand time series data are usually large-scaled with complicated patterns and cause difficulties for imputation. This paper presents a Bayesian matrix factorization (BMF)-based imputation method for large-scale electricity load missing value imputation. Through factorizing the original electricity load matrix into two latent matrices, the intrinsic information of the electricity load matrix can be discovered. Two Bayesian inference algorithms, Gibbs sampling and iterated conditional models are applied to solve the BMF model. The affect of the matrix rank to the electricity load imputation task is empirically studied. Experimental results on three real-world electricity load datasets are presented to show the superiority of the proposed method against five benchmark algorithms.**

***Index Terms*—Electricity Load Imputation, Bayesian Matrix Factorization, Multivariate Time Series**

I. Introduction

With the prosperous development of smart grid technologies, a large volume of electricity load data are collected from heterogeneous grid equipment, such as phasor measurement units, infrastructures, and baseload plants. These electricity load data are usually represented as multivariate time series for more than one input source. They have been widely used in a variety of applications including but not limited to grid simulations [1], fault-detection [2], load forecasting [3]–[7], and load management [8]. The performance of electricity load applications is determined by the data quality of input time series, which largely depend on various factors such as the accuracy of the data acquisition and the reliability of the data transmission. For example, noise and outliers encountered at the data acquisition process are usually recorded as missing values during the data pre-processing [9]. Moreover, unanticipated accidents such as power outages at the data transmission process can lead to the loss of data [10].

In statistics, imputation is the process of replacing missing data with substituted values [11]. As a classic imputation technique, interpolation aims to fill the missing values based on the range of observed data points, such as using the previous or the next observations in a given range of statistical indicators (i.e., mean and median) [12]. It is often used to impute time series with high-frequent patterns. As one of the most popular interpolation methods, K-nearest neighbor (KNN) has been successfully used to impute missing data in power time series [11] by mean predictions. Although KNN is simple and fast, it cannot provide reliable predictions when facing a large percentage of missing data and complex patterns in the time series [13]. Expectation maximization (EM) is another imputation approach that takes both the empirical mean and variance into account by maximizing the likelihood in an iterative procedure [14]. However, some empirical studies demonstrated that both EM and KNN can be biased when the percentage of the missing data is larger than 10% [15]. To mitigate this issue, copy-paste imputation (CPI) is a recently proposed method for univariate energy time series with a fixed interval of missing values [16]. Note that these aforementioned approaches are restricted to the univariate time series while the correlation between different time series has largely been ignored, in view of the strong dependency among the electricity load multivariate time series in a range of geographical and time-scale ranges. Multi-task Gaussian process (MTGP) [17] is an ideal choice for multivariate time series, and it has been applied to the financial signal processing in [17]. However, since the conventional MTGP is notorious for its cubic time complexity for the model training and inference, it is hardly useful for large-scale applications.

Nonnegative matrix factorization (NMF) seeks to reveal the hidden structure of nonnegative data by decomposing a nonnegative matrix $V \in \mathbb{R}^{m\times n}$ into a product of two low-rank nonnegative latent matrices $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{r\times n}$ [18], where $r \ll m$ and $r \ll n$. As a probabilistic variant of the NMF approach, Bayesian matrix factorization (BMF) has shown to be more effective to capture the correlation between data [19]. It has been applied to a wide range of applications, such movie recommendation system [19], [20], gene expression [21], and geographical modeling [22].

In this paper, we propose an data imputation framework based on BMF for electricity load data. By factorizing the electricity load data into a time latent and a location latent matrix, respectively, in a probabilistic manner, we can capture the corresponding spatio-temporal relationship. Furthermore, by tweaking the rank of the matrix factorization, the computational complexity and the accuracy of imputation can therefore be controlled in a principled manner. In our empirical study, the performance of our proposed BMF-based imputation approaches are compared against five selected peer methods in the literature on three benchmark datasets. Experimental results fully confirm the effectiveness and superiority of our proposed BMF-based imputation approaches.

The rest of this paper is organized as follows. Section II starts with a problem statement and followed by a pragmatic overview of selected works. Our proposed data imputation framework based on BMF are delineated in Section III. The

All authors are with the Department of Computer Science, University of Exeter, North Park Road, Exeter, EX4 4QF, UK (e-mail: {x.li5, k.li}@exeter.ac.uk)

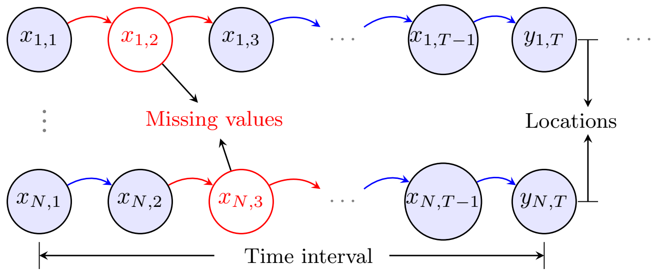


Fig. 1. Example of electricity load multivariate time series.

experimental settings are given in Section IV while the results are discussed in Section V. Finally, Section VI concludes this paper and sheds some lights on future directions.

II. Preliminary

This section starts with a formal problem statement from a matrix factorization perspective. Then, we give a pragmatic overview of the existing works on imputation for electricity load data.

A. Problem Statement

Let $X \in \mathbb{R}^{N\times T}$ be a multivariate time series of the electricity load data collected from N different locations for T time steps. Each column vector $\mathbf{x}^i = (x_1^i, \cdots, x_{t-1}^i, x_t^i, x_{t+1}^i, \cdots, x_T^i)^\top$ where $1 \le t \le T$ denotes the t-th time step and i is the i-th location. Fig. 1 shows a graphical representation of the electricity load multivariate time series where the missing values are marked in red.

Let X factorize into a product two latent matrices U and V. The factorization rank k is usually smaller than both row rank and column rank. The goal of matrix factorization is finding a product of two low-rank matrices that minimizes the difference between the original matrix and the factorization results. Different from the traditional matrix factorization applications, the rows (number of time stamps) of the electricity load prediction tasks are much higher than the column numbers (number of locations/infrastructures). The selection of the rank is crucial to the missing electricity load data imputation task as the rank affects both performance and computation cost of the matrix factorization.

B. Related works

In this subsection, four representative imputation methods with three variants of MTGP are selected to cover the techniques broadly applied in the literature and representative of various imputation strategies.

K-nearest neighbors (KNN) is a widely applied linear interpolation method because of its low complexity [11], [23]. It computes the interpolating weights w_i and provides a smoother approximation $G(x)$ based on the k nearest neighbors of x, such that:

$$G(\mathbf{x}) = \sum_{i=1}^{k} w_i(\mathbf{x}) f(\mathbf{x}^i), \tag{1}$$

where $f(x_i)$ is the observed value of x_i. From equation (1), it can be seen that KNN interpolation gives inaccurate approximation if there is a wide range of neighbour points missed.

EM is an iterative algorithm to find a maximum likelihood estimation [24], [25]. In each iteration, EM imputes the missing values with the most possible values given by the observed data's empirical mean and variance-covariance matrix [14]. The algorithm terminates when the imputed values do not change in new iterations.

Copy-paste imputation is a recently proposed imputation method for univariate energy time series with an interval of missing values [16]. CPI imputes the missing data with using a Prophet-based method to find weekly patterns, therefore it cannot be applied to the electricity load data having missing values within days. In addition, CPI is designed for univariate time series, the correlations between multivariate electricity load time series can hardly be captured.

Gaussian process is a non-parametric approach for nonlinear interpolation with placing a prior distribution over latent functions defined for any number of data points. MTGP is an extension of Gaussian process to multiple outputs [26].

Let $\mathcal{X} = \{\mathbf{x}^i\}_{i=1}^N$ denotes a set of N input data, $\mathcal{Y} = \{y_{1,}, \cdots, y_{1N}, \cdots, y_{T1}, \cdots, y_{TN}\}$ denotes a set of outputs of each input on T tasks. Let $\mathcal{Y}'$ be a subset of $\mathcal{Y}$ denotes the observed data points, the unseen values $\mathcal{Y}^c$ can be predicted by the latent function $f(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_T(\mathbf{x}))^\top$. For an arbitrary $\mathbf{x}$, a GP prior on the latent function is defined as $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$, where mean function $m(\mathbf{x})$ is assumed to be 0, kernel function $K(\mathbf{x}, \mathbf{x}')$ is a positive semi-definite matrix.

With more expressive kernels are developed, MTGP can describe spatial-temporal relations among infinitely many random variables while maintaining statistical dependence between tasks. The linear model of coregionalization (LMC) is the most general positive definite form of a multi-output covariance kernel. The covariance matrices obtained under the LMC assumption can be represented in the form of a sum of separable kernels [27]. A spectral mixture kernel is a positive-definite stationary kernel first introduced in [26] by modeling the spectral density as a weighted mixture of Q square-exponential functions with weights w_q, centers μ_q, and diagonal covariance matrix Σ_q:

$$k(\tau) = \sum_{q=1}^{Q} w_q \exp(-\frac{1}{2}\tau^\top \Sigma_q \tau) \cos(\mu_q^\top \tau), \tag{2}$$

where $\mu_q \in \mathbb{R}^n$, $\Sigma = \mathrm{diag}(\sigma_1^q, \cdots, \sigma_n^q)$, and $w_q, \sigma_q \in \mathbb{R}_+$.

The MTGP with spectral mixture-linear coregionalization (SM-LMC) model is the first introduces the spectral mixture kernel in the LMC framework [26] to represent complex spectral relationships between tasks. To further capture the cross-phase spectrum between tasks, MTGP with cross-spectral mixture (CSM) kernel is introduced in [28]. An extension work on CSM built on Cramér's Theorem named multi-output spectral mixture (MOSM) kernel is proposed in [29].

In this paper, the multi-task GP with these three spectral-based kernels are used in comparison with the proposed Bayesian matrix factorization approach. Due to the cubic

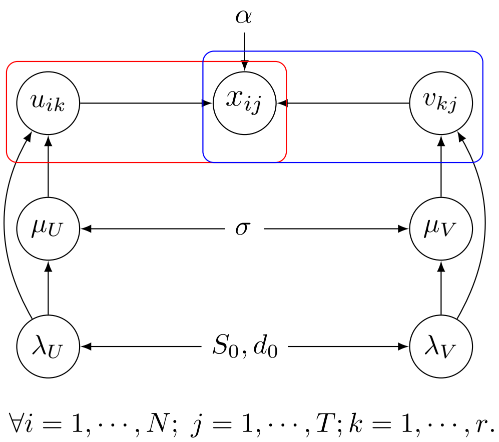


Fig. 2. The graphical model representation of the BMF approach for electricity load imputation.

time complexity of inference, the GP models cannot perform on large-scale dataset. Therefore, a month-long interval is extracted from the full dataset for MTGP comparison.

Remark 1. Most, if not all, existing energy load imputation methods in the smart grid literature lack of the ability to capture spatialtemporal relationship for large-scale multivariate time series. However, the real-world applications have the need of handling missing values for large dataset in daily uses.

III. Proposed Method

This section starts with the theoretical foundation of BMF, followed by the architecture of our proposed BMF-based framework for imputation of the electricity load data. At the end, we come up with two probabilistic inference methods of the BMF model.

A. The BMF Model

Let $U \in \mathbb{R}^{N\times r}$ be a latent time feature matrix and $V \in \mathbb{R}^{r\times T}$ be a latent location feature matrix, where r denotes the rank. The BMF approach decomposes the input matrix $X \in \mathbb{R}^{N\times T}$ into a product of U and V together with a residual matrix $E \in \mathbb{R}^{N\times T}$ as:

$$X = UV + E, \tag{3}$$

where x_{ij} denotes element on the i-th row and j-th column of matrix X.

In the probabilistic approach, the latent matrices U and V are treated as random variables.

The observed data x_{ij} in matrix X is expressed as a likelihood function comes from the product of U and V with additional Gaussian noise σ. The conditional distribution over the observed electricity load X is:

$$p(X|U, V, \sigma) = \prod_{i=1}^{N}\prod_{j=1}^{T}\mathcal{N}(x_{ij}|u_i \times v_j, \sigma^{-1}), \tag{4}$$

where u_i is the i-th row vector of U, v_j is the j-th column vector of V, $\mathcal{N}(x|\mu, \sigma)$ is the density of Gaussian distribution with mean μ and precision σ.

The prior distributions over the U and V can either be assumed to be Gaussian [19]:

$$\begin{aligned} p(U|\mu_U, \lambda_U) &= \prod_{i=1}^{N}\mathcal{N}(u_i|\mu_U, \lambda_U^{-1}) \\ p(V|\mu_V, \lambda_V) &= \prod_{j=1}^{T}\mathcal{N}(v_j|\mu_V, \lambda_V^{-1}), \end{aligned} \tag{5}$$

or can be assumed to be independently exponentially distributed with scales $\Gamma_U > 0$ and $\Gamma_V > 0$ [20]:

$$p(U|\Gamma_U) = \prod_{i=1}^{N}\mathcal{E}(u_i|\Gamma_U), \quad p(V|\Gamma_V) = \prod_{j=1}^{T}\mathcal{E}(v_j|\Gamma_V), \tag{6}$$

and an inverse Gamma density prior for the noise variance $p(\sigma) = \mathcal{G}^{-}1(\sigma|S_G, \Gamma_G)$ with shape S_G and scale Γ_G.

For the Gaussian prior distributions in (5), in order to reduce Bayesian updating to modifying the hyperparameter, multivariate Gaussian distribution's conjugate prior, Gaussian-Wishart prior, is placed on the location hyperparameters $\{\mu_U, \lambda_U\}$ and time hyperparameters $\{\mu_V, \lambda_V\}$ [19]:

$$\begin{aligned} &p(\{\mu_U, \lambda_U\}|\{\mu_0, \lambda_0\}) \\ &= \mathcal{N}(\mu_U|\mu_0, (\beta_0\lambda_U)^{-1})\mathcal{W}(\lambda_U|S_0, d_0) \\ &p(\{\mu_V, \lambda_V\}|\{\mu_0, \lambda_0\}) \\ &= \mathcal{N}(\mu_V|\mu_0, (\beta_0\lambda_V)^{-1})\mathcal{W}(\lambda_V|S_0, d_0), \end{aligned} \tag{7}$$

where $\mathcal{W}$ is the Wishart distribution with d_0 degrees of freedom and a scale matrix S_0 with dimension $r \times r$ and a normalizing constant c. Follow the non-informative setting in [19], the degree of freedom $d_0 = r$, S_0 is set as the identity matrix for both location and time hyperparameters, $\mu_0 = 0$. The graphical model representation of BMF is shown in Fig. 2.

The predictive distribution of the missing electricity load data x_{ij}^* for the location i at time j can be obtained by a marginal probability over its model parameters and hyperparameters.

Let $\mathcal{H}$ denote the set of the model parameters and hyperparameters, the predictive distribution is [19]:

$$\begin{aligned} p(x_{ij}^*|X, \mathcal{H}_0) = \iint & p(x_{ij}^*|u_i, v_j)p(U, V|X, \mathcal{H}_U, \mathcal{H}_V) \\ & p(\mathcal{H}_U, \mathcal{H}_V)|\mathcal{H}_0)d\{U, V\}d\{\mathcal{H}_U, \mathcal{H}_V\}. \end{aligned} \tag{8}$$

B. Probabilistic Inference

As equation (8) is analytically intractable, Bayesian inference is required to approximate the predictive distribution. Gibbs sampling is a Markov chain Monte Carlo-based algorithm that was first used in BMF and has been applied to collaborative filtering [30] and user/movie ratings [19]. However, since the electricity load data is dense and may have a high percentage of missing values up to 50% of total amounts of data, Gibbs sampling is computational expensive in this case. As an alternative, a more efficient block coordinate ascent type algorithm named iterated conditional models (ICM) [20] is used.

```
Algorithm 1: Gibbs Sampling
Input: $U_0 \in \mathbb{R}^{N\times r}, V_0 \in \mathbb{R}^{r\times T}$,
       maximum iteration $I_{\text{max}}$.
Output: $X$.
for iterations $i = 1$ to $I_{max}$ do
    Draw $U$'s hyperparameters by eqn. (9):
      $\mathcal{H}_U^i \sim p(\mu_U, \lambda_U | U^i, \mathcal{H}_0)$;
    for $j = 1$ to $N$ do
        Draw $u_j^{i+1} \sim p(u_j | X, V^i, \mathcal{H}_U^i)$
    Draw $V$'s hyperparameters:
      $\mathcal{H}_V^i \sim p(\mu_V, \lambda_V | V^i, \mathcal{H}_0)$;
    for $j = 1$ to $T$ do
        Draw $v_j^{i+1} \sim p(v_j | X, U^{i+1}, \mathcal{H}_V^i)$
return $X = U_{I_{\text{max}}+1} \times V_{I_{\text{max}}+1}$.
```

1) Gibbs Sampling: Gibbs sampling is a Markov chain Monte Carlo algorithm by alternatively drawing new samples from the conditional posterior densities of the model parameters to estimate the joint posterior [31].

In (5) and (7), since the Gaussian-Wishart prior is used as the prior, the conditional distribution can be expressed analytically. The conditional distributions of U and V have same form, so do their associated hyperparameters $\{\mu_U, \lambda_U\}$ and $\{\mu_V, \lambda_V\}$. In this case, only the conditional distributions of U and its hyperparameters are addressed herein.

For the hyperparameters $\{\mu_U, \lambda_U\}$, the conditional distribution over U and its hyperparameters is:

$$\begin{aligned} &p(\mu_U, \lambda_U | U, \mathcal{H}_0) \\ =&\mathcal{N}(\mu_U | \mu_U^*, (\beta_U^* \lambda_U)^{-1}) \mathcal{W}(\lambda_U | S_0 U^*, d_U^*), \end{aligned} \tag{9}$$

where

$$\begin{aligned} &\mu_U^* = \frac{\beta_0 \mu_0}{\beta_0 + N}, \quad \beta_U^* = \beta_0 + N, \quad d_U^* = d_0 + N, \\ &(S_U^*)^{-1} = S_0^{-1} + N S_U + \frac{\beta_0 N}{\beta_0 + N}(\bar{U} - \mu_0)(\bar{U} - \mu_0)^{\intercal} \\ &\bar{U} = \sum_{i=1}^{N} u_i, \quad \bar{S} = \frac{1}{N} \sum_{i=1}^{N} U_i U_i^{\intercal}. \end{aligned} \tag{10}$$

The conditional distribution of the location latent matrix U is a factorization of the products of the conditional distributions of its column vectors u_i, for $i \in \{1, \cdots, N\}$, i.e.:

$$p(U | X, V, \mathcal{H}_U) = \prod_{i=1}^{N} p(u_i, X, V, \mathcal{H}_U). \tag{11}$$

Therefore, the parallelization is enabled for each u_i. For each u_i, the conditional distribution $p(u_i | X, V, \mathcal{H}_U, \sigma)$ is a Gaussian distribution $\mathcal{N}(u_i | \mu_i^*, (\lambda_i^*)^{-1})$, where

$$\begin{aligned} \mu_i^* &= (\lambda_i^*)^{-1} (\mu_U \lambda_U + \sigma \sum_{j=1}^{M} (\mathbf{v}_j x_{ij})) \\ \lambda_i^* &= \lambda_U + \sigma \sum_{j=1}^{T} v_j v_j^{\top}. \end{aligned} \tag{12}$$

The full algorithm procedure is detailed in Algorithm 1.

2) Iterated Conditional Models (ICM): ICM is an iterative inference algorithm first introduced in [20]. It works similar to Gibbs sampling but it sets the latent matrices U and V and their hyperparameters as the modes of their corresponding distributions, instead of sampling from the conditional posterior distributions. The modes can be obtained through their closed-form expressions and reduces the computational cost.

Consider the exponential distributed case in (6), the conditional distribution of U is propotional to the product of a normal distribution and an exponential distribution, such that:

$$\begin{aligned} p(u_i | X, U_{\backslash(i)}, V, \sigma) \propto \mathcal{N}(u_i | \mu_{u_i}, \sigma) \times \mathcal{E}(u_i | \Gamma_U) \\ p(v_j | X, V_{\backslash(i)}, U, \sigma) \propto \mathcal{N}(v_j | \mu_{v_j}, \sigma) \times \mathcal{E}(v_j | \Gamma_V), \end{aligned} \tag{13}$$

where $U_{\backslash(i)}$ and $V_{\backslash(j)}$ denote all elements in U and V, except the i-th row in U and the j-th column in V. μ_{u_i} and μ_{v_j} are the mean vectors for U and V, respectively. σ is the precision.

When the algorithm converges, ICM gives a maximum a posterior estimation. Due to computer's precision issue, ICM may converge to solutions with multiple zero columns. To avoid poor approximation, all the zero entries are added by a small positive value during the iterations as in [32].

IV. Experimental Setup

This section introduces the setup of our empirical study including the dataset, peer algorithms and their parameter settings, evaluation metrics, and the statistical tests.

A. Datasets

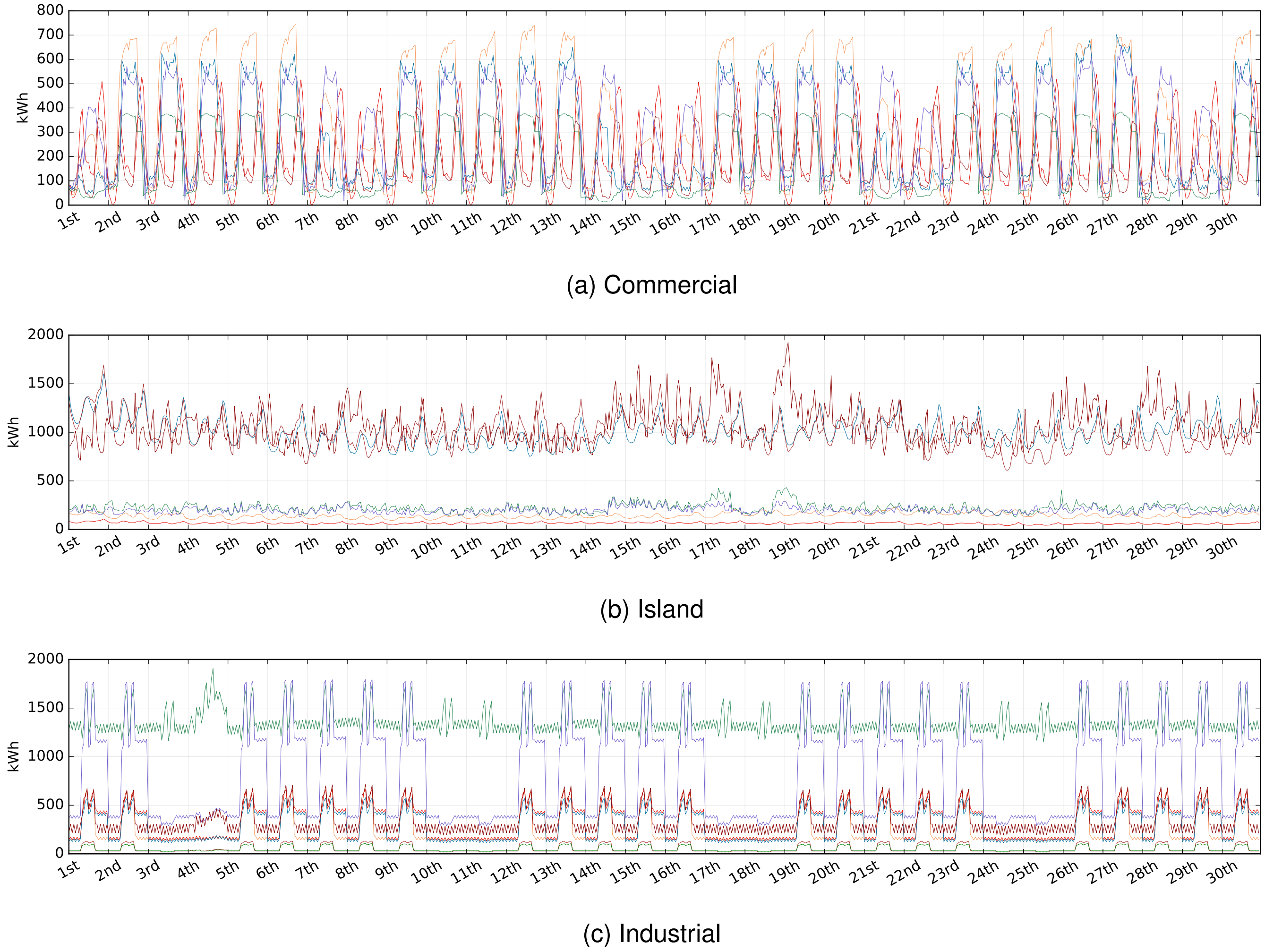
Three real world annual electricity load dataset are used in the experiments. Each dataset contains hourly average power values in kilowatt-hour, resulting in $8,760$ values (24 hours × 365 days) of six to eight locations a year. The commercial and industrial dataset are provided by [33]. It consists of a set of 6 and a set of 8 electricity load data collected from commercial and industrial end-use sectors respectively in 2020. Island dataset contains the electricity load over year collected from 7 islands.

The information of each dataset is detailed in Table I. The first month's data of there benchmark dataset are plotted in Fig. 3. Two separate experiments are conducted on a short-term (one month) and a long-term (one year) for comparison. The short-term dataset randomly selects an interval of 720 timestamps from the full dataset. For each dataset, ten to fifty percents of the data are randomly selected and set as missing points for testing.

B. Peer Algorithms and Parameter Settings

The performance of the two proposed BMF-based imputation methods, `Gibbs sampling` and `ICM`, are compared with five commonly used imputation methods: `KNN`, `EM`, and three `MTGPs` with different kernels introduced in Section II. The parameter settings of each algorithms are listed as follows.

- `Gibbs sampling` and `ICM`: The maximum iteration number is set as 200.
- `KNN`: The nearest neighbours number k is set as the number of locations in each experiment.



(a) Commercial

(b) Island

(c) Industrial

Fig. 3. The first month's plots of three benchmark datasets with 720 timestamps

TABLE I
DATASET DESCRIPTION

	No. of locations	No. of timestamps	Min	Max	Median	Mean	Variance
Commercial	6	8760	0	1000	233.18	291.7232	5.62×10^4
Island	7	8760	0	1975	188.50	381.5900	1.32×10^5
Industrial	8	8760	17.4	2387.7	300.00	490.9752	2.76×10^5

- EM: The maximum iteration number is set as 1000 and the error tolerance is set as 1e-06.
- MTGPs: The rank Q of three MTGP methods are set as the number of locations. The optimizers use a stochastic gradient descent named Adam. The learning rates are set as 0.1. The maximum iteration numbers are set as 200.

C. Performance metric and Statistical Tests

The performance of the algorithms are evaluated by the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{14}$$

where x_i is the observed value, $\hat{x}_i$ is the predicted value.

Besides, the CPU runtime are recorded to compare the computational cost of the algorithms. All the experiments are conducted on MacBook with M1 chip with 8-core CPU and 16 GB ram.

Wilcoxon signed-rank test [34] is applied for statistical interpretation of the significance of the comparison results. It is a non-parametric statistical test without making assumptions about the data's underlying distribution. The significance level is set to 0.05 in the experiments.

V. RESULTS AND DISCUSSIONS

The empirical study is driven by addressing the following four research questions (RQs).

- *RQ1*: What is the impact of matrix factorization rank to BMF-based imputation?
- *RQ2*: Does the BMF framework work for different patterns of electricity load dataset?
- *RQ3*: Does the percentage of missing values affect the BMF-based imputation, and how is the performance

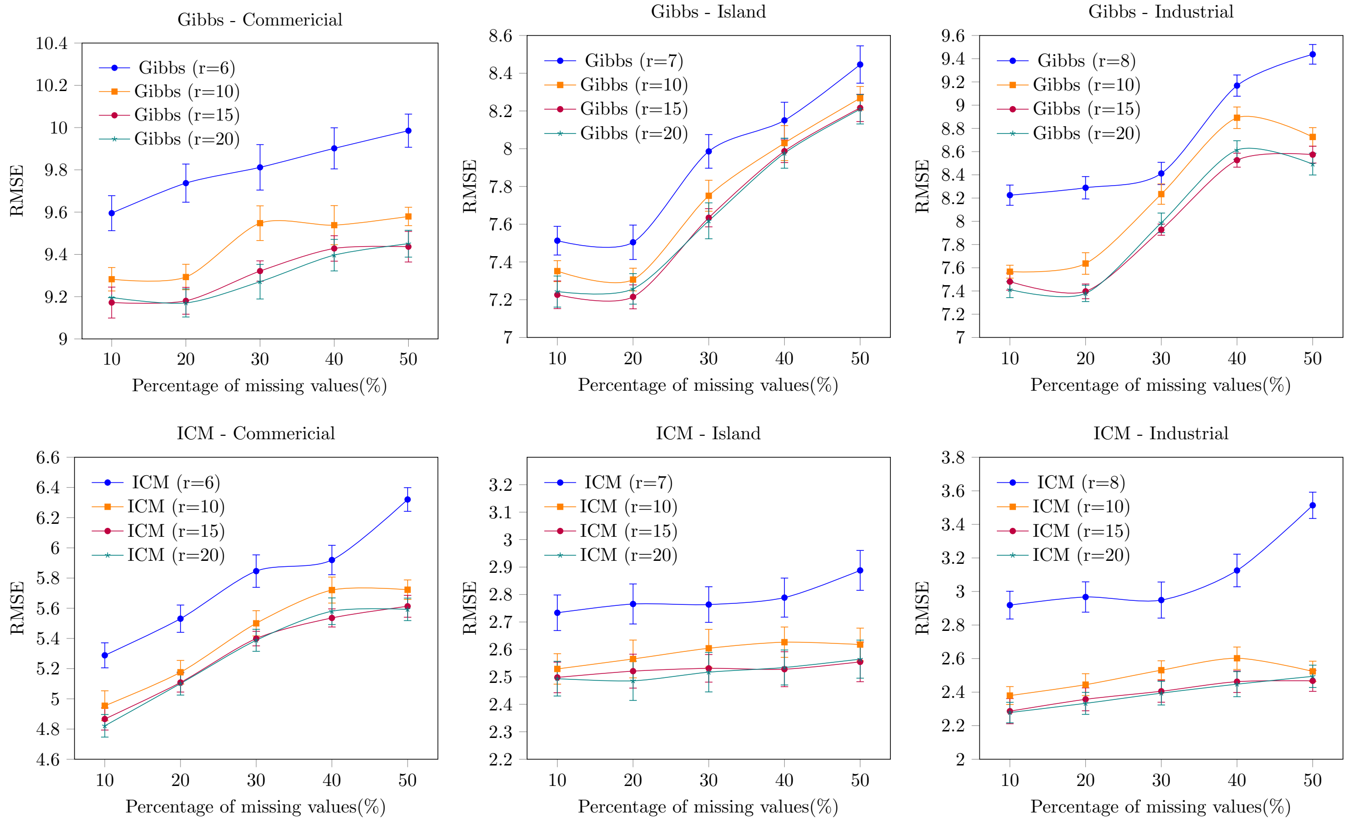


Fig. 4. Performance of Gibbs sampling and ICM with different ranks on three datasets.

TABLE II
RUNTIME (SECONDS) OF GIBBS SAMPLING (FIRST ROW IN EACH SUB-BLOCK) AND ICM (SECOND ROW IN EACH SUB-BLOCK) ON THREE BENCHMARK DATASETS WITH 50% MISSING VALUES

Dataset	Algorithm	rank=(6,7,8)	10	15	20
Commercial	Gibbs	72.28(6.26)	126.14(5.37)	195.73(7.37)	261.69(6.53)
	ICM	1.6(0.01)	2.49(0.01)	3.65(0.01)	4.91(0.01)
Island	Gibbs	69.23(5.76)	100.064(6.11)	155.07(6.44)	210.19(8.02)
	ICM	1.99(0.01)	2.73(0.02)	4.01(0.01)	5.34(0.01)
Industrial	Gibbs	76.98(6.20	99.92(5.85)	153.672(7.61)	226.20(7.92)
	ICM	2.33(0.01)	2.90(0.01)	4.14(0.01)	5.35(0.01)

comparing to the benchmark algorithms' performances?

- *RQ4*: Does the BMF-based method scale to long-term electricity load dataset?

A. Impact of factorization rank

Figure 4 shows the performance of Gibbs sampling and ICM with ranks ranging from the minimum number of locations 6 to 20 on three datasets. Subfigures (a) and (b) have similar RMSE performance when increase the percentage of missing values from 10% to 50%.

Response to RQ1: *From the experiment in this subsection, we find that the increment of the rank (from $r = 6$ to $r = 15$) leads to decrement RMSE result till $r = 15$ with 200 fixed iterations. In addition, a high-rank value leads to high computational time.*

It can be observed from the figures that the RMSE value decreases while the rank increases on all three datasets for both Gibbs sampling and ICM, since larger base matrices U and coefficient matrices V can store more information and give better approximation of X in theory. In addition, the gaps between blue and orange lines are larger than the gaps between orange and red lines. RMSE gap between rank 10 and 15 is improved less than the gap between rank 6 and rank 10 for the commercial dataset (rank 7 and 8 for island and industrial dataset, respectively). The gaps between red and green are the least among all. This indicates the improvement of RMSE is growing less, and it does not improve obviously when the rank is higher than 15.

Table II records the runtime of Gibbs sampling and ICM on three benchmark datasets with 50% missing values with different rank values. As the BMF computes the factorization of the original matrix, the number of missing values does not affect the runtime. The table shows that the runtime increases while the rank increases. Since more elements of U and V are required to be inferred during the algorithm process, more computational resources are required.

TABLE III
RMSE ON THREE SHORT-TERM ($T = 720$) DATASETS WITH MISSING VALUES RANGING FROM TEN TO FIFTY PERCENTS, WHERE THE BEST RESULTS ARE BOLDFACED, THE SECOND BEST RESULTS ARE UNDERLINED, † INDICATES THE BETTER RESULT (IN BOLD FACE) IS OF STATISTICAL SIGNIFICANCE ACCORDING TO THE WILCOXON'S RANK-SUM TEST AT THE 5% SIGNIFICANCE LEVEL.

Datasets	Algorithms	Missing Values = 10%	20%	30%	40%	50%
Commercial	KNN	67.18(1.19)†	76.21(1.25)†	90.78(2.96)†	89.65(3.42)†	92.78(3.08)†
	EM	79.23(1.50)†	87.81(2.88)†	98.67(4.01)†	103.84(3.52)†	110.84(3.25)†
	MTGP + SM-LMC	8.14(0.32)†	8.58(0.57)†	9.43(0.48)†	10.86(0.65)†	12.22(0.34)†
	MTGP + CSM	7.22(0.40)†	7.60(0.42)†	8.25(0.70)†	9.50(0.29)†	10.87(0.62)†
	MTGP + MOSM	7.31(0.45)†	7.74(0.47)†	8.37(0.52)†	9.63(0.51)†	11.28(0.52)†
	Gibbs	8.81(0.37)†	9.12(0.37)†	9.25(0.33)†	9.26(0.30)†	9.37(0.33)†
	ICM	**3.68(0.36)**	**3.99(0.36)**	**4.44(0.35)**	**4.36(0.38)**	**4.42(0.28)**
Island	KNN	86.92(2.89)†	94.51(2.46)†	105.92(4.51)†	107.41(3.77)†	108.90(3.11)†
	EM	50.49(3.35)†	52.45(2.67)†	66.28(3.50)†	63.37(5.02)†	69.45(4.59)†
	MTGP + SM-LMC	7.51(0.22)†	8.27(0.48)†	8.68(0.57)†	10.25(0.63)†	11.15(0.55)†
	MTGP + CSM	6.93(0.33)†	7.37(0.45)†	8.04(0.42)†	9.12(0.23)†	10.80(0.44)†
	MTGP + MOSM	7.09(0.30)†	7.66(0.29)†	7.82(0.47)†	8.80(0.39)†	10.43(0.37)†
	Gibbs	7.23(0.38)†	7.12(0.30)†	7.46(0.31)†	7.97(0.33)†	8.11(0.24)†
	ICM	**2.33(0.30)**	**2.33(0.29)**	**2.42(0.23)**	**2.49(0.30)**	**2.46(0.22)**
Industrial	KNN	33.00(0.82)†	47.53(1.07)†	77.49(2.82)†	101.68(4.06)†	106.34(4.96)†
	EM	38.63(1.01)†	45.28(1.52)†	60.85(2.09)†	75.23(2.93)†	100.34(3.56)†
	MTGP + SM-LMC	7.38(0.18)†	7.70(0.21)†	7.76(0.30)†	9.91(0.28)†	0.81(0.38)†
	MTGP + CSM	7.25(0.15)†	7.27(0.19)†	7.27(0.26)†	8.25(0.35)†	9.02(0.49)†
	MTGP + MOSM	7.17(0.22)†	7.04(0.25)†	7.64(0.17)†	8.22(0.28)†	8.30(0.41)†
	Gibbs	7.37(0.31)†	7.23(0.30)†	8.02(0.32)†	8.44(0.27)†	8.53(0.30)†
	ICM	**2.13(0.27)**	**2.21(0.29)**	**2.35(0.29)**	**2.38(0.29)**	**2.35(0.31)**

TABLE IV
RMSE ON THREE LONG-TERM ($T = 8760$) DATASETS WITH MISSING VALUES RANGING FROM TEN TO FIFTY PERCENTS.

Datasets	Algorithms	Missing Values = 10%	20%	30%	40%	50%
Commercial	KNN	69.24(1.24)†	119.97(3.05)†	137.57(3.30)†	144.39(3.67)†	147.06(3.93)†
	EM	68.73(1.35)†	92.56(2.79)†	111.19(3.51)†	127.94(3.11)†	136.04(3.68)†
	MTGP + SM-LMC	-	-	-	-	-
	MTGP + CSM	-	-	-	-	-
	MTGP + MOSM	-	-	-	-	-
	Gibbs	9.17(0.44)†	9.18(0.38)†	9.39(0.35)†	9.49(0.29)†	9.44(0.42)†
	ICM	**4.87(0.35)**	**5.11(0.24)**	**5.40(0.13)**	**5.54(0.34)**	**5.61(0.20)**
Island	KNN	81.60(1.77)†	106.85(2.10)†	110.61(3.52)†	120.30(2.84)†	127.61(2.55)†
	EM	82.49(1.41)†	92.44(3.54)†	102.0655(2.55)†	108.63(2.90)†	114.56(3.69)†
	MTGP + SM-LMC	-	-	-	-	-
	MTGP + CSM	-	-	-	-	-
	MTGP + MOSM	-	-	-	-	-
	Gibbs	7.23(0.25)†	7.22(0.28)†	7.63(0.53)†	7.99(0.36)†	8.2161(0.29)†
	ICM	**2.50(0.30)**	**2.52(0.22)**	**2.52(0.49)**	**2.53(0.43)**	**2.55(0.39)**
Industrial	KNN	37.83(0.98)†	71.1307(2.32)†	101.85(3.66)†	117.61(3.00)†	120.49(3.18)†
	EM	31.06(1.20)†	55.07(1.89)†	82.93(2.56)†	104.48(2.54)†	114.18(3.60)†
	MTGP + SM-LMC	-	-	-	-	-
	MTGP + CSM	-	-	-	-	-
	MTGP + MOSM	-	-	-	-	-
	Gibbs	7.48(0.19)†	7.40(0.41)†	7.93(0.32)†	8.53(0.23)†	8.57(0.25)†
	ICM	**2.29(0.29)**	**2.36(0.18)**	**2.41(0.20)**	**2.47(0.23)**	**2.47(0.35)**

Since the RMSE is not improved obviously from rank 15 to rank 20 and more runtime is required, the final ranks for `Gibbs sampling` and `ICM` are set as 15 on these three benchmark datasets in the later experiments.

B. Performance Evaluation of the Effectiveness of the BMF Imputation

Table III records the RMSE performance of seven imputation algorithms on three short-term (one month) electricity load datasets with missing values ranging from ten percent to fifty percent. It can be observed that `ICM` performs the best among all. It competes the second-best results by improving the RMSE up to 45.75% on 'commercial', 71.53% percent on 'island', and 68.74% percent on 'industrial'. The `MTGP` with CSM and MOSM achieves better performance than `Gibbs sampling` when the percentage of missing values is lower than 30% on 'commercial' and 'industrial' datasets, but are all competed by the two BMF inference methods on 'island'. When the number of missing values increases, `Gibbs Sampling` and `ICM` are less affected while the RMSEs of the other comparison algorithms increase dramatically, especially for `KNN` and `EM`.

Response to RQ2: *The experimental results show that BMF-based imputation methods work consistently well on three electricity load datasets with different patterns compared to the benchmark algorithms.*

TABLE V
RUNTIME (SECONDS) ON THREE LONG-TERM DATASETS WITH MISSING VALUES RANGING FROM TEN TO FIFTY PERCENTS.

Datasets	Algorithms	Missing Values = 10%	20%	30%	40%	50%
Commercial	KNN	**1.41**(**0.00**)	**2.45**(**0.00**)	**3.52**(**0.00**)	4.37(0.00)	4.95(0.00)
	EM	4.35(0.01)	12.52(0.08)	24.01(1.33)	45.29(3.53)	83.83(5.08)
	MTGPs	-	-	-	-	-
	Gibbs	190.97(10.24)	188.02(7.28)	185.55(10.19)	179.33(11.43)	195.73(13.54)
	ICM	3.64(0.00)	3.65(0.00)	3.65(0.00)	**3.64**(**0.00**)	**3.65**(**0.00**)
Island	KNN	**1.64**(**0.00**)	**2.57**(**0.00**)	)**4.00**(**0.00**)	4.85(0.00)	5.18(0.00)
	EM	4.83(0.01)	11.50(0.05)	23.24(1.72)	46.97(1.88)	91.19(3.35)
	MTGPs	-	-	-	-	-
	Gibbs	157.24(8.68)	152.13(9.15)	160.90(11.21)	152.87(10.54)	155.07(11.95)
	ICM	4.01(0.00)	4.00(0.00)	4.01(0.00)	**4.01**(**0.00**)	**4.01**(**0.00**)
industrial	KNN	**1.72**(**0.00**)	**2.84**(**0.00**)	**4.13**(**0.00**)	)5.25(0.00)	5.77(0.00)
	EM	4.45(0.01)	11.48(0.06)	23.66(0.32)	45.77(1.79)	89.52(4.70)
	MTGPs	-	-	-	-	-
	Gibbs	169.73(9.66)	155.02(10.03)	165.88(8.69)	158.52(11.71)	153.67(10.53)
	ICM	4.14(0.00)	4.13(0.00)	4.13(0.00)	**4.14**(**0.00**)	**4.14**(**0.00**)

Response to RQ3: *From the observations in this experiment, we see that the increment of missing values still leads to a decrement of the BMF's and Gibbs' RMSE results. The decrement ratio of the proposed methods is similar to MTGP methods but far lower than KNN and EM.*

C. Performance Evaluation of the BMF Imputation on Long-term Electricity load Dataset

Table IV records the RMSE performance on a long-term (one year) electricity load dataset with missing values ranging from ten percent to fifty percent. Three GP-based imputation methods are not recorded in this table since they cannot be computed on a large scale. Although the RMSE values are slightly increased for ICM and Gibbs sampling on these long-term datasets than on the short-term datasets, the results are similar to Table III. ICM defeats all the benchmark algorithms and Gibbs sampling. Gibbs sampling achieves the second best on all the datasets with the absence of MTGP models.

Table V records the runtime of each run in Table IV. ICM takes the least time to compute when the percentage of missing values is greater than 30%, and KNN takes the least time to compute when it is less or equal to 30%. Both Gibbs sampling and ICM have almost the same runtime for all percentage of missing values, whereas the runtime of KNN and EM increases when the percentage of missing values increase. As the MTGPs have a cubic time complexity, they averagely take 2 hours to compute on the short-term dataset and extremely long time to compute on the long-term dataset. The experiments were not conducted for the MTGPs on the long-term dataset, and thus the runtime are not recorded for them.

Response to RQ4: *From the experimental results recorded herein in comparison to the previous subsections, we conclude that ICM is able to scale to long-term electricity load dataset with a small increment of RMSE and runtime.*

VI. CONCLUSION

This paper presents a BMF model for imputing electricity load missing data for solving general smart grid data missing problems. The spatiotemporal relationship can be well captured by factorizing the electricity multivariate time series into two latent matrices. Two Bayesian inference methods, Gibbs sampling and ICM, are applied to solve the BMF model. Due to the large-scale of the electricity load dataset, Gibbs sampling cannot perform enough samplings to infer a good distribution during the algorithm procedure, and the performance is limited. Compared to Gibbs sampling, ICM is fast as it obtains the latent matrices and their hyperparameters by taking the local maximum of the joint probability with their closed form.

The experimental results show that BMF using the ICM inference method outperforms all the peer algorithms and BMF using the Gibbs sampling method. The percentage of missing values does not affect the runtime of Gibbs sampling of ICM as the algorithms find the factorization of the full data matrix in all runs no matter with the percentage of missing values. In addition, when the multivariate time series patterns are different (e.g. 'island' dataset), BMF is able to capture the intrinsic information by the latent matrices, unlike MTGP is better in filling time series with similar trends. Future investigations may aim at developing online Bayesian approaches based on other models or inference methods for online smart grid applications.

REFERENCES

[1] J. Rivera, P. Nasirifard, J. Leimhofer, and H. Jacobsen, "Automatic generation of real power transmission grid models from crowdsourced data," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5436–5448, 2019.

[2] M. Gilanifar, J. Cordova, H. Wang, M. Stifter, E. Ozguven, T. Strasser, and R. Arghandeh, "Multi-task logistic low-ranked dirty model for fault detection in power distribution system," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 786–796, 2020.

[3] J. Fiot and F. Dinuzzo, "Electricity demand forecasting by multi-task learning," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 544–551, 2016.

[4] W. Kong, D. Z, D. Hill, F. Luo, and Y. Xu, "Short-term residential load forecasting based on resident behaviour learning," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1087–1088, 2018.

[5] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu, and J. He, "Short-term load forecasting with deep residual networks," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3943–3952, 2019.

[6] J. Luo, T. Hong, and S.-C. Fang, "Robust regression models for load forecasting," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5397–5404, 2019.

[7] H. Aprillia, H. Yang, and C. Huang, "Statistical load forecasting using optimal quantile regression random forest and risk assessment index," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1467–1480, 2021.

[8] A. Baniasadi, D. Habibi, O. Bass, and M. A. Masoum., “Optimal real-time residential thermal energy management for peak-load shifting with experimental verification,” *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5587–5599, 2019.

[9] H. N. Akouemo and R. j. Povinelli, “Data improving in time series using arx and ann models,” *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3352–3359, 2017.

[10] C. King and J. Strapp, “Chapter 11—software infrastructure and the smart grid,” *Smart Grid: Integrating Renewable, Distributed and Efficient Energy*, p. 259–288, 2012.

[11] M. Kim, S. Park, J. Lee, Y. Joo, and J. Choi, “Learning-based adaptive imputation methodwith knn algorithm for missing power data,” *Energies*, vol. 10, no. 10, 2017.

[12] J. F. Steffensen, *Interpolation*. New York: Dover, Mineola, 2013.

[13] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: A critical evaluation,” *BMC Medical Informatics and Decision Making*, vol. 16, 07 2016.

[14] C. Do and S. Batzoglou, “What is the expectation maximization algorithm?” *Nature biotechnology*, vol. 26, no. 8, pp. 897–899, 2008.

[15] D. Bennett, “How can i deal with missing data in my study?” *Australian and New Zealand Journal of Public Health*, vol. 25, pp. 464 – 469, 10 2001.

[16] M. Weber, M. Turowski, H. K. Çakmak, R. Mikut, U. G. Kühnapfel, and V. Hagenmeyer, “Data-driven copy-paste imputation for energy time series,” *IEEE Transactions on Smart Grid*, vol. 12, pp. 5409–5419, 2021.

[17] T. Wolff, A. Cuevas, and F. Tobar, “Gaussian process imputation of multiple financial series,” May 2020, pp. 8444–8448.

[18] H. Lee and D. Seung, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[19] R. Salakhutdinov and A. Mnih, “Bayesian probabilistic matrix factorization using markov chain monte carlo,” in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML ’08. New York, NY, USA: Association for Computing Machinery, 2008, p. 880–887.

[20] M. Schmidt, O. Winther, and L. Hansen, “Bayesian non-negative matrix factorization,” *Independent Component Analysis and Signal Separation Lecture Notes in Computer Science*, p. 540–547, 2009.

[21] S. T.D., T. Gao, and E. Fertig, “Cogaps 3: Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structuresn,” *BMC Bioinformatics*, vol. 21, no. 453, Oct. 2020.

[22] B. Liu, H. Xiong, S. Papadimitriou, Y. Fu, and Z. Yao, “A general geographical probabilistic factor model for point of interest recommendation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, pp. 1167–1179, 05 2015.

[23] R. Sibson, “A brief description of natural neighbor interpolation (chapter 2),” *Interpreting Multivariate Data*, p. 21–36, 1981.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[25] M. Wang, C. Tsai, and W. Lin, “Towards missing electric power data imputation for energy management systems,” *Expert Systems with Applications*, vol. 174, p. 114743, 2021.

[26] A. Wilson and R. Adams, “Gaussian process kernels for pattern discovery and extrapolation,” in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 28:3. Atlanta, Georgia, USA: PMLR, 17-19 Jun 2013, pp. 1067–1075.

[27] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, “Kernels for vector-valued functions: A review,” *Found. Trends Mach. Learn.*, vol. 4, no. 3, p. 195–266, mar 2012.

[28] K. R. Ulrich, D. E. Carlson, K. Dzirasa, and K. Carin, “Gp kernels for cross-spectrum analysis,” in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[29] G. Parra and F. Tobar, “Spectral mixture kernels for multi-output gaussian processes,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[30] R. Salakhutdinov and A. Mnih, “Probabilistic matrix factorization,” in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*. Curran Associates, Inc., 2007, pp. 1257–1264.

[31] R. M. Neal, “Probabilistic inference using markov chain monte carlo methods,” *Technical Report CRG-TR-93-1*, 1993.

[32] B. Thomas, F. Jes, and L. Pietro, “Comparative study of inference methods for bayesian nonnegative matrix factorisation,” in *European Conference on Principles of Data Mining and Knowledge Discovery*, Jul 2017.

[33] F. Angizeh and A. G. M. Jafari, “Dataset on hourly load profiles for a set of 24 facilities from industrial, commercial, and residential end-use sectors,” *Mendeley Data, V1*, 2020.

[34] F. Wilcoxon, *Individual Comparisons by Ranking Methods*. Springer, 1992.