Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model

Xinyu Chen, Zhaocheng He*, Yixian Chen, Yuhuan Lu

Guangdong Provincial Key Laboratory of Intelligent Transportation Systems, Research Center of Intelligent Transportation System, Sun Yat-Sen University, Guangzhou, Guangdong 510006, China

Abstract

Spatiotemporal traffic data, which represent multi-variate time series on considering different spatial locations, are ubiquitous in real-world transportation systems. However, the inevitable incompleteness makes data-driven intelligent transportation systems suffer from the incorrect response. Therefore, imputing missing data is of great importance but challenging as it is not easy to capture spatiotemporal traffic patterns, including explicit and latent features. In this study, we propose an augmented tensor factorization model by incorporating generic forms of domain knowledge. More specifically, we present a fully Bayesian framework for automatically learning parameters of this model using variational Bayes. Relying on the publicly available urban traffic speed data set collected in Guangzhou, China, experiments on two types of missing data scenarios (i.e., random and non-random) demonstrate that the Bayesian augmented tensor factorization (BATF) model achieves performance improvements, and outperforms the state-of-the-art baselines like Bayesian tensor factorization models as well. Meanwhile, we discover interpretable patterns from the experimentally learned global parameter, biases and latent factors that indeed conform to the basic trend of traffic states.

Keywords: Missing data imputation, Pattern discovery, Spatiotemporal traffic data, Bayesian tensor decomposition, Variational Bayes

1. Introduction

Incomplete data is a common and sometimes inevitable problem in the data-driven intelligent transportation systems, which also exists in several applications including traffic states monitoring. Although we have many advanced sensors to enable us to collect all of data as we want, unfortunately, it may be still impossible because some types of data are sparse by nature. Other types of urban traffic data may be restricted by the spatial coverage of sensors. The uncertainty like communication malfunctions and transmission distortions of sensors for data collection is another influential factor. Thus, in this context, making accurate data imputation and improving data quality support the success of any application which makes use of that type of data.

In general, the ideas in the urban traffic data completion task can be summarized as follows. If we have partially observed data with both spatial and temporal resolution, then a model is required to be capable of discovering spatiotemporal patterns. As we know, this is rather similar to the collaborative filtering in the field of recommender systems (Salakhutdinov and Mnih, 2008; Xiong et al., 2010). For example, given a road segment, concerning the case that we only have an incomplete time-series about traffic volume or speed for indicating traffic states, then its own partially observed time-series and the temporal trends of other road segments may be both considered into modeling (Laa et al., 2018).

To this end, there is a family of matrix factorization models has been tested for missing traffic data imputation in the past studies (Qu et al., 2008, 2009; Li et al., 2013). Qu et al. (2009) proposed a probabilistic principal component analysis (PPCA) based imputation method for traffic volume data completion, and in their experiments, this method was illustrated to make use of features including not only statistical information of traffic flow, but periodicity and local predictability. Within this work, BPCA evaluated by Qu et al. (2008) was proven to be inferior to PPCA. Following this work, Li et al. (2013) demonstrated that using spatial and temporal dependence could help reduce estimation errors significantly for PPCA based methods. Here, the assumption of strictly daily similarity is not required in these models.

Recently, Rodrigues et al. (2018) applied the multi-output Gaussian processes (GPs) to model the complex spatial and temporal patterns about incomplete traffic speed data. Since the model is capable of considering observation uncertainty and spatial dependencies between nearby road segments, their experiments showed that

*Corresponding author.

Email addresses: chenxy346@mail2.sysu.edu.cn (Xinyu Chen), hezhch@mail.sysu.edu.cn (Zhaocheng He), chenyx96@mail2.sysu.edu.cn (Yixian Chen), luyh6@mail2.sysu.edu.cn (Yuhuan Lu)

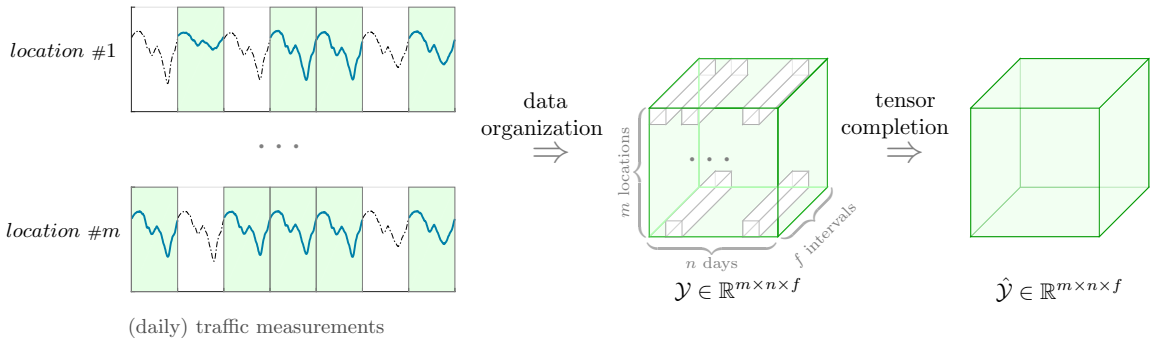


Figure 1: Graphical illustration of the tensor completion task and its framework including data organization and tensor completion, in which traffic measurements are partially observed.

the model achieves significantly better results than some popular state-of-the-art imputation methods including independent GP, PPCA and Bi-LSTM.

Another choice for modeling spatiotemporal traffic data is through organizing these data into tensors. In the existing studies, several tested tensor completion models for missing traffic data imputation can be summarized into two categories. The first is low-rank tensor completion models which include SiLRTC, FaLRTC and HaLRTC proposed by Liu et al. (2013), then, the experiments about traffic volume data imputation have indicated that HaLRTC supports the usage of spatial information from neighbouring locations (Ran et al., 2016). However, these models are sensitive to the observation noises and suffer from the sparsity issue. When dealing with an extremely sparse tensor, they are inferior to capture the global information of the tensor (Zhao et al., 2015a), hence, the imputation accuracy of these models is rather limited.

The second is tensor decomposition for an incomplete tensor, Tan et al. (2013a,b); Asif et al. (2016) employed multi-linear tensor decomposition as to estimate missing traffic data, and the extensive experiments demonstrated that the tensor decomposition models outperform the PCA based models. With a combination of the Bayesian inference methods, we have opportunities for tackling the non-convex optimization problem underlying tensor decomposition (Xiong et al., 2010; Rai et al., 2014; Zhao et al., 2015a,b). On the other hand, the sparsity issue can be alleviated to some extent under the Bayesian framework.

This paper has a new starting point for addressing the missing traffic data imputation problem in the existing Bayesian tensor factorization. Specifically, inspired by modeling the explicit patterns into matrix and tensor models (Koren et al., 2009; Chen et al., 2018), the aim in this work is to develop an augmented tensor factorization which combines explicit patterns and latent factors. In a variational Bayes framework, the model parameters formulated in the augmented tensor factorization can be learned by inferring their variational posteriors. In terms of Bayesian tensor factorization, Hu et al. (2015); Rai et al. (2015) also reported that deterministic inference methods such as variational Bayes and Expectation Maximization (EM) are more efficient than the close-formed Markov chain Monte Carlo (MCMC).

In this new approach to missing data imputation, we wish to further investigate the semantic interpretability of the augmented tensor factorization, in which we incorporate generic forms of domain knowledge from transportation systems. On considering the missing data scenario and by comparing to the Bayesian tensor factorization models, we finally intend to explore the advantages of newly formulated augmented tensor factorization with fully Bayesian treatment.

2. Preliminaries

A natural way of modeling multi-dimensional traffic data is in the form of a tensor. In this study, our task can be described as follows: given a partially observed tensor $\mathcal{Y}$, relying on its algebraic structure, learn from partial elements and further estimate the unobserved elements in this tensor (see Fig. 1). Formally, we use $\mathcal{Y}_\Omega$ to denote the partially observed elements in $\mathcal{Y}$, and where Ω is the set of observation index. For simplicity of notation, we only investigate the tensor factorization for third-order tensor $\mathcal{Y} \in \mathbb{R}^{m \times n \times f}$ in this study. And we further define $\mathcal{O} \in \mathbb{R}^{m \times n \times f}$ to be a binary tensor with such that $o_{ijt} = 1$ if y_{ijt} is observed (i.e., $(i, j, t) \in \Omega$), and $o_{ijt} = 0$ otherwise.

Regarding such formulated tensor completion task, the main developed tensor models can be summarized as follows:

(a) *Basic tensor factorization.* To identify an underlying low-dimensional representation of r latent factors, one well-established model is CANDECOMP/PARAFAC(CP) decomposition and we can factorize $\mathcal{Y}$ into factor matrices $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ and $X \in \mathbb{R}^{f \times r}$ (Kolda and Bader, 2009). For any (i, j, t)-th element of the tensor $\mathcal{Y}$, there exists an approximation which is a multi-linear combination of r latent factors from each factor matrix as

$$y_{ijt} \approx \sum_{k=1}^{r} u_{ik} v_{jk} x_{tk}, \forall i, j, t. \tag{1}$$

(b) *Low-rank tensor completion.* By introducing trace norm to approximate the rank of matrices, Liu et al. (2013) developed several low-rank tensor completion algorithms involving convex optimization. In their definition, the optimization can be briefly written as

$$\begin{aligned} \min_{\mathcal{X}} : & \sum_{i=1}^{3} \alpha_i \|\mathcal{X}_{(i)}\|_* \\ \text{s.t.} : & \mathcal{X}_\Omega = \mathcal{Y}_\Omega, \end{aligned} \tag{2}$$

where α_is are constants satisfying $\alpha_i \geq 0$ and $\sum_{i=1}^{3} \alpha_i = 1$. In the objective function, $\mathcal{X}_{(i)}$ denotes the matrix unfolded along i-th mode and $\|\mathcal{X}_{(i)}\|_*$ represents the trace norm of $\mathcal{X}_{(i)}$.

(c) *Bayesian tensor factorization.* The goal of tensor factorization is to find a low-rank approximation, thus, taking CP factorization as an example, we can in effect minimize the loss function to achieve a tensor factorization by

$$\mathcal{J} = \sum_{(i,j,t)\in\Omega} (y_{ijt} - \sum_{k=1}^{r} u_{ik} v_{jk} x_{tk})^2 + w_u \mathcal{R}_u + w_v \mathcal{R}_v + w_x \mathcal{R}_x, \tag{3}$$

where $\mathcal{R}_u, \mathcal{R}_v, \mathcal{R}_x$ are regularization terms related to the factor matrices U, V, X respectively, and their weights are $\{w_u, w_v, w_x\}$. Unfortunately, one common thing associated with this optimization is the non-convex problem, thus leading to the development of Bayesian Gaussian tensor factorization approaches (Xiong et al., 2010; Rai et al., 2014; Hu et al., 2015; Rai et al., 2015; Zhao et al., 2015a,b).

In terms of experimental evaluation, one real world data set collected from transportation system can be easily represented by a tensor. For example, relying on the urban traffic speed data set (publicly available at https://doi.org/10.5281/zenodo.1205229) and using Bayesian tensor decomposition approach, Chen et al. (2019) demonstrated that third-order tensor is the best algebraic structure for missing data imputation task, while comparing with matrix and fourth-order tensor.

3. Bayesian augmented tensor factorization model

In the following, we first introduce the mathematical formulation of the proposed augmented tensor factorization. Subsequently, we briefly discuss the Bayesian treatment for solving this model which we use to learn factorization parameters. Finally, we infer the variational posterior of parameters and hyper-parameters in the Bayesian graphical network and derive an implementation for the augmented tensor factorization using variational Bayes.

3.1. Augmented tensor factorization

Typically, CP decomposition maps multi-dimensional data to a joint latent factor space of dimensionality r, such that complicated interactions are modeled as inner products in that space (see Eq. (1)). In this work, we first assume that the data tensor collecting from transportation systems can be expressed by explicit patterns and latent factors. Given a third-order tensor $\mathcal{Y} \in \mathbb{R}^{m\times n\times f}$, we propose an augmented tensor factorization which is formally written as follows

$$y_{ijt} \approx \mu + \phi_i + \theta_j + \eta_t + \sum_{k=1}^{r} u_{ik} v_{jk} x_{tk}, \forall i, j, t, \tag{4}$$

where $\mu \in \mathbb{R}$ is a global parameter responsible for all tensor elements, $\boldsymbol{\phi} \in \mathbb{R}^m, \boldsymbol{\theta} \in \mathbb{R}^n, \boldsymbol{\eta} \in \mathbb{R}^f$ are bias vectors relative to each dimension, and $U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}, X \in \mathbb{R}^{f\times r}$ are factor matrices controlling the interactions among different dimensions. In this model, global parameter μ and bias vectors $\{\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta}\}$ indicate the explicit patterns, while factor matrices $\{U, V, X\}$ indicate the latent factors as indicated in Fig. 2.

In the proposed model, parameter μ serves as a global parameter for approaching the overall average of the tensor. Based on μ, bias along each dimension captures the explicit patterns or features (Koren et al., 2009). In the context of urban transportation, it is also valuable to model the bias of spatial and temporal attributes (Chen et al., 2018). Now, for example, say that the average speed of all road segments and periods is 39 km/h. Furthermore, suppose that the selected road segment tends to be 10 km/h above the average, and the period tends to be 5 km/h lower than the average. Then, the estimate for that speed value would be 44 km/h (i.e., $39 + 10 - 5 = 44$).

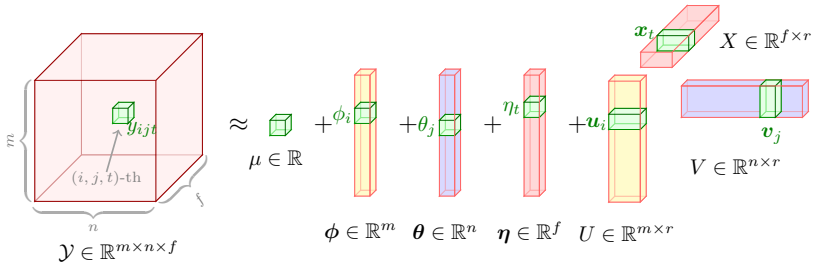


Figure 2: Proposed augmented tensor factorization to tensor completion.

3.2. Bayesian network

We propose to use Bayesian inference methods to learn the parameters $\{\mu, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta}, U, V, X\}$ from the data tensor $\mathcal{Y}$. Since Gaussian assumption over tensor factorization has an equivalent form to the common-used loss function (Xiong et al., 2010), it is convenient to assume that each element of $\mathcal{Y}$ follows independent Gaussian distribution, i.e.,

$$y_{ijt} \sim \mathcal{N}(\mu + \phi_i + \theta_j + \eta_t + \sum_{k=1}^{r} u_{ik} v_{jk} x_{tk}, \tau^{-1}), \forall i, j, t, \tag{5}$$

where the notation $\mathcal{N}(\cdot)$ denotes a Gaussian distribution and τ is the precision (inverse of variance), which is a universal parameter for all tensor elements. From a probability prospective, Eq. (5) also helps model the uncertainty and randomness of data tensor $\mathcal{Y}$.

The basic idea of Bayesian inference is to derive the posterior distribution as a consequence of prior distribution and likelihood function in a Bayesian setting. For training the model parameters in Eq. (5), we therefore need to place conjugate priors over model parameters, i.e.,

$$\begin{aligned} \mu, \phi_i, \theta_j, \eta_t &\sim \mathcal{N}(\mu_0, \tau_0^{-1}), \forall i, j, t, \\ \boldsymbol{u}_i &\sim \mathcal{N}(\boldsymbol{\mu}_u, \Lambda_u^{-1}), \forall i, \\ \boldsymbol{v}_j &\sim \mathcal{N}(\boldsymbol{\mu}_v, \Lambda_v^{-1}), \forall j, \\ \boldsymbol{x}_t &\sim \mathcal{N}(\boldsymbol{\mu}_x, \Lambda_x^{-1}), \forall t, \\ \tau &\sim \text{Gamma}(a_0, b_0), \end{aligned} \tag{6}$$

where the vector $\boldsymbol{u}_i \in \mathbb{R}^r$ is the i-th row of factor matrix $U \in \mathbb{R}^{m \times r}$ with dimensionality r, the vector $\boldsymbol{v}_j \in \mathbb{R}^r$ is the j-th row of factor matrix $V \in \mathbb{R}^{n \times r}$, and the vector $\boldsymbol{x}_t \in \mathbb{R}^r$ is the t-th row of factor matrix $X \in \mathbb{R}^{f \times r}$. The notation Gamma$(\cdot)$ denotes Gamma distribution, and its probability density function is given by

$$\text{Gamma}(\tau \mid a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau), \tag{7}$$

where a and b are shape and rate parameters respectively, and the notation $\Gamma(\cdot)$ denotes Gamma function.

Referring to the Bayesian probabilistic matrix factorization proposed by Salakhutdinov and Mnih (2008), we further place Gaussian-Wishart priors on hyperparameters $\{\boldsymbol{\mu}_u, \Lambda_u, \boldsymbol{\mu}_v, \Lambda_v, \boldsymbol{\mu}_x, \Lambda_x\}$ as follows

$$\begin{aligned} \boldsymbol{\mu}_u, \Lambda_u &\sim \mathcal{N}(\boldsymbol{\mu}_u \mid \boldsymbol{\mu}_0, (\beta_0 \Lambda_u)^{-1}) \times \mathcal{W}(\Lambda_u \mid W_0, \nu_0), \\ \boldsymbol{\mu}_v, \Lambda_v &\sim \mathcal{N}(\boldsymbol{\mu}_v \mid \boldsymbol{\mu}_0, (\beta_0 \Lambda_v)^{-1}) \times \mathcal{W}(\Lambda_v \mid W_0, \nu_0), \\ \boldsymbol{\mu}_x, \Lambda_x &\sim \mathcal{N}(\boldsymbol{\mu}_x \mid \boldsymbol{\mu}_0, (\beta_0 \Lambda_x)^{-1}) \times \mathcal{W}(\Lambda_x \mid W_0, \nu_0), \end{aligned} \tag{8}$$

where the marginal distribution over $\{\Lambda_u, \Lambda_v, \Lambda_x\}$ is a Wishart distribution (i.e., $\mathcal{W}(\cdot)$), and the conditional distribution over $\{\boldsymbol{\mu}_u, \boldsymbol{\mu}_v, \boldsymbol{\mu}_x\}$ given $\{\Lambda_u, \Lambda_v, \Lambda_x\}$ is a multivariate Guassian distribution. Specifically, the probability density function of Wishart distribution is given by

$$\mathcal{W}(\Lambda \mid W, \nu) = \frac{1}{C} |\Lambda|^{\frac{1}{2}(\nu - r - 1)} \exp(-\frac{1}{2} \text{tr}(W^{-1} \Lambda)), \tag{9}$$

where C is the normalization constant, and the notation tr$(\cdot)$ denotes the trace of a squared matrix. Λ follows Wishart distribution with ν degrees of freedom and a $r \times r$ scale matrix W.

In the following, we use Θ to represent $\{\mu, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta}, U, V, X, \tau, \boldsymbol{\mu}_u, \Lambda_u, \boldsymbol{\mu}_v, \Lambda_v, \boldsymbol{\mu}_x, \Lambda_x\}$ where it is clear just to reduce the verbosity and avoid confusion In terms of Eq. (4), the aim is to explicitly derive the model parameters $\{\mu, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta}, U, V, X\}$.

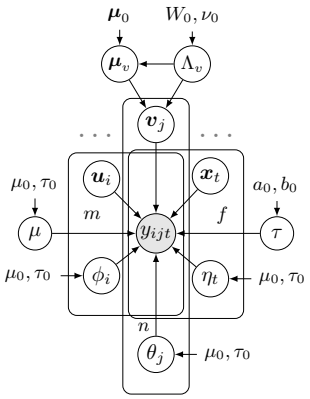


Figure 3: Graphical model of the Bayesian augmented tensor factorization. The observations $y_{jit}, (i, j, t) \in \Omega$ is shown by the shaded node, while the intersection of three plates illustrates that this third-order tensor is partially observed.

3.3. Posterior inference using variational Bayes

Bayesian tensor factorization models have attracted much interest in collaborative filtering (Xiong et al., 2010), image completion (Zhao et al., 2015a,b), and relational graph analysis (Schein et al., 2016) (e.g., social network and international relation). Previously, one good but relatively slow inference method for learning factorization parameters is to use MCMC to alternatively draw parameters and hyper-parameters from conditional posterior distributions in an iterative manner. Therefore, another inference method is variational Bayes, which tends to speed up the convergence.

3.3.1. Fundamentals of variational Bayes

Variational Bayes is a deterministic inference method for approximating posterior distributions. In this study, we wish to seek a distribution $q(\Theta)$ to approximate the true posterior distribution $p(\Theta \mid \mathcal{Y}_\Omega)$ by minimizing the Kullback-Leibler (KL) divergence as follows

$$\begin{aligned} \mathrm{KL}(q(\Theta) \parallel p(\Theta \mid \mathcal{Y}_\Omega)) &= \int q(\Theta) \ln \frac{q(\Theta)}{p(\Theta \mid \mathcal{Y}_\Omega)} d\Theta \\ &= \ln p(\mathcal{Y}_\Omega) - \int q(\Theta) \ln \frac{p(\mathcal{Y}_\Omega, \Theta)}{q(\Theta)} d\Theta, \end{aligned} \tag{10}$$

where $\ln p(\mathcal{Y}_\Omega)$ represents the model evidence which is a constant, and its lower bound is defined as

$$\mathcal{L}(q) = \int q(\Theta) \ln \frac{p(\mathcal{Y}_\Omega, \Theta)}{q(\Theta)} d\Theta.$$

According to the mean-field approximation, the variational posterior distribution $q(\Theta)$ is fully factorized by

$$\begin{aligned} q(\Theta) =& q(\mu) \times \prod_{i=1}^{m} q(\phi_i) q(\boldsymbol{u}_i) \times \prod_{j=1}^{n} q(\theta_j) q(\boldsymbol{v}_j) \times \prod_{t=1}^{f} q(\eta_t) q(\boldsymbol{x}_t) \\ & \times q(\tau) \times q(\boldsymbol{\mu}_u, \Lambda_u) \times q(\boldsymbol{\mu}_v, \Lambda_v) \times q(\boldsymbol{\mu}_x, \Lambda_x). \end{aligned} \tag{11}$$

For any s-th variable Θ_s, the equivalent form for maximizing the lower bound $\mathcal{L}(q)$ is given as follows

$$\ln q(\Theta_s) = \mathbb{E}_{q(\Theta \setminus \Theta_s)} \left[\ln p(\mathcal{Y}_\Omega, \Theta) \right] + \text{const}, \tag{12}$$

where the notation $\mathbb{E}_{q(\Theta \setminus \Theta_s)} [\cdot]$ denotes an expectation with respect to the distributions $q(\Theta \setminus \Theta_s)$ over all variables except Θ_s. Putting Eqs. (5), (6) and (8) together, the joint distribution $p(\mathcal{Y}_\Omega, \Theta)$ mentioned in Eq. (12) is

$$\begin{aligned} p(\mathcal{Y}_\Omega, \Theta) =& p(\mathcal{Y}_\Omega \mid \mu, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta}, U, V, X, \tau) \times p(\mu) \times \prod_{i=1}^{m} p(\phi_i) p(\boldsymbol{u}_i \mid \boldsymbol{\mu}_u, \Lambda_u) \\ & \times \prod_{j=1}^{n} p(\theta_j) p(\boldsymbol{v}_j \mid \boldsymbol{\mu}_v, \Lambda_v) \times \prod_{t=1}^{f} p(\eta_t) p(\boldsymbol{x}_t \mid \boldsymbol{\mu}_x, \Lambda_x) \times p(\tau) \\ & \times p(\boldsymbol{\mu}_u, \Lambda_u) \times p(\boldsymbol{\mu}_v, \Lambda_v) \times p(\boldsymbol{\mu}_x, \Lambda_x). \end{aligned} \tag{13}$$

3.3.2. The variational posterior distribution of μ

Starting with variational posterior distribution $q(\mu)$ with respect to the model parameter μ and applying Eqs. (12) and (13), we get the logarithm form of $q(\mu)$ as

$$\begin{aligned}\ln q(\mu) &= \mathbb{E}_{q(\Theta\backslash\mu)}[\ln p(\mathcal{Y}_\Omega, \Theta)] + \text{const} \\ &= -\sum_{(i,j,t)\in\Omega} \frac{1}{2}\mathbb{E}[\tau(z_{ijt}-\mu)^2] - \frac{1}{2}\tau_0\mathbb{E}[(\mu-\mu_0)^2] + \text{const} \\ &= -\frac{1}{2}(\mathbb{E}[\tau]\sum_{(i,j,t)\in\Omega} o_{ijt} + \tau_0)\mu^2 + (\mathbb{E}[\tau]\sum_{(i,j,t)\in\Omega} \mathbb{E}[z_{ijt}] + \tau_0\mu_0)\mu + \text{const},\end{aligned} \tag{14}$$

where the notation $\mathbb{E}_{q(\Theta\backslash\mu)}[\cdot]$ denotes an expectation with respect to the distributions $q(\Theta\backslash\mu)$ over all variables except μ. Equivalently, the variational posterior introduced in Eq. (14) is $q(\mu) = \mathcal{N}(\tilde{\mu}, \tilde{\tau}^{-1})$ with such that

$$\tilde{\mu} = \tilde{\tau}^{-1}(\mathbb{E}[\tau]\sum_{(i,j,t)\in\Omega} \mathbb{E}[z_{ijt}] + \tau_0\mu_0), \tilde{\tau} = \mathbb{E}[\tau]\sum_{(i,j,t)\in\Omega} o_{ijt} + \tau_0, \tag{15}$$

where $z_{ijt} = y_{ijt} - \phi_i - \theta_j - \eta_t - \sum_{k=1}^r u_{ik}v_{jk}x_{tk}$ and its variational expectation is given by

$$\mathbb{E}[z_{ijt}] = y_{ijt} - \mathbb{E}[\phi_i] - \mathbb{E}[\theta_j] - \mathbb{E}[\eta_t] - \sum_{k=1}^r \mathbb{E}[u_{ik}]\mathbb{E}[v_{jk}]\mathbb{E}[x_{tk}]. \tag{16}$$

3.3.3. The variational posterior distribution of $\{\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta}\}$

As can be seen from the Bayesian graphical model in Fig. 3 and the prior setting in Eq. (8), bias vectors $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta}$ are expressed by their independent Gaussian elements. Considering the i-th element ϕ_i of $\boldsymbol{\phi} \in \mathbb{R}^m$ as an example, we have

$$\begin{aligned}\ln q(\phi_i) &= \mathbb{E}_{q(\Theta\backslash\phi_i)}[\ln p(\mathcal{Y}_\Omega, \Theta)] + \text{const} \\ &= -\frac{1}{2}(\mathbb{E}[\tau]\sum_{j,t:(i,j,t)\in\Omega} o_{ijt} + \tau_0)\phi_i^2 + (\mathbb{E}[\tau]\sum_{j,t:(i,j,t)\in\Omega} \mathbb{E}[f_{ijt}] + \tau_0\mu_0)\phi_i + \text{const},\end{aligned} \tag{17}$$

where $\sum_{j,t:(i,j,t)\in\Omega}$ denotes the sum over $j \in \{1, 2, ..., n\}$ and $t \in \{1, 2, ..., f\}$ with specific i in the index set Ω.

We therefore derive the variational posterior $q(\phi_i) = \mathcal{N}(\tilde{\mu}_\phi, \tilde{\tau}_\phi^{-1})$ with such updates

$$\tilde{\mu}_\phi = \tilde{\tau}_\phi^{-1}(\mathbb{E}[\tau]\sum_{j,t:(i,j,t)\in\Omega} \mathbb{E}[f_{ijt}] + \tau_0\mu_0), \tilde{\tau}_\phi = \mathbb{E}[\tau]\sum_{j,t:(i,j,t)\in\Omega} o_{ijt} + \tau_0, \tag{18}$$

where

$$\mathbb{E}[f_{ijt}] = y_{ijt} - \mathbb{E}[\mu] - \mathbb{E}[\theta_j] - \mathbb{E}[\eta_t] - \sum_{k=1}^r \mathbb{E}[u_{ik}]\mathbb{E}[v_{jk}]\mathbb{E}[x_{tk}]. \tag{19}$$

Once we have the variational posterior distribution $q(\phi_i)$, we can also derive the variational posterior distributions $q(\theta_j)$ and $q(\eta_t)$ in a similar manner respectively.

3.3.4. The variational posterior distribution of $\{U, V, X\}$

Since factor matrices have multi-variate Gaussian prior over their row vectors, thus, for instance, we can write the variational posterior distribution $q(\boldsymbol{u}_i)$ for updating the factor matrix U as follows

$$\begin{aligned}\ln q(\boldsymbol{u}_i) =& \mathbb{E}_{q(\Theta\backslash\boldsymbol{u}_i)}[\ln p(\mathcal{Y}_\Omega, \Theta)] + \text{const} \\ =& -\frac{1}{2}\sum_{j,t:(i,j,t)\in\Omega} \mathbb{E}[\tau\left(w_{ijt} - \boldsymbol{u}_i^T(\boldsymbol{v}_j \circledast \boldsymbol{x}_t)\right)^2] - \frac{1}{2}\mathbb{E}[(\boldsymbol{u}_i - \boldsymbol{\mu}_u)^T\Lambda_u(\boldsymbol{u}_i - \boldsymbol{\mu}_u)] + \text{const} \\ =& -\frac{1}{2}\boldsymbol{u}_i^T(\mathbb{E}[\tau]\sum_{j,t:(i,j,t)\in\Omega} \mathbb{E}\left[(\boldsymbol{v}_j \circledast \boldsymbol{x}_t)(\boldsymbol{v}_j \circledast \boldsymbol{x}_t)^T\right] + \mathbb{E}[\Lambda_u])\boldsymbol{u}_i \\ &+ \frac{1}{2}\boldsymbol{u}_i^T(\mathbb{E}[\tau]\sum_{j,t:(i,j,t)\in\Omega} \mathbb{E}[\boldsymbol{v}_j \circledast \boldsymbol{x}_t]\mathbb{E}[w_{ijt}] + \mathbb{E}[\Lambda_u]\mathbb{E}[\boldsymbol{\mu}_u]) + \text{const},\end{aligned} \tag{20}$$

where the symbol $\circledast$ represents Hadamard product, and $\boldsymbol{u}_i^T(\boldsymbol{v}_j \circledast \boldsymbol{x}_t) = \sum_{k=1}^r u_{ik}v_{jk}x_{tk}$. For brevity, $\mathbb{E}[w_{ijt}] = y_{ijt} - \mathbb{E}[\phi_i] - \mathbb{E}[\theta_j] - \mathbb{E}[\eta_t]$. We have the variational posterior $q(\boldsymbol{u}_i) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_u, \tilde{\Lambda}_u^{-1})$ whose parameters are given by

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_u &= \tilde{\Lambda}_u^{-1}(\mathbb{E}\left[\tau\right] \sum_{j,t:(i,j,t)\in\Omega} \mathbb{E}\left[\boldsymbol{v}_j \circledast \boldsymbol{x}_t\right] \mathbb{E}\left[w_{ijt}\right] + \mathbb{E}\left[\Lambda_u\right] \mathbb{E}\left[\boldsymbol{\mu}_u\right]),\\ \tilde{\Lambda}_u &= \mathbb{E}\left[\tau\right] \sum_{j,t:(i,j,t)\in\Omega} \mathbb{E}\left[(\boldsymbol{v}_j \circledast \boldsymbol{x}_t)(\boldsymbol{v}_j \circledast \boldsymbol{x}_t)^T\right] + \mathbb{E}\left[\Lambda_u\right],\end{aligned} \tag{21}$$

where assuming that the vectors $\{\boldsymbol{v}_j, \boldsymbol{x}_t\}, \forall j, t$ are independent (Zhao et al., 2015a), then

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{v}_j \circledast \boldsymbol{x}_t)(\boldsymbol{v}_j \circledast \boldsymbol{x}_t)^T\right] &= \mathbb{E}\left[\boldsymbol{v}_j\boldsymbol{v}_j^T\right] \circledast \mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^T\right]\\ &= (\mathbb{E}\left[\boldsymbol{v}_j\right]\mathbb{E}\left[\boldsymbol{v}_j^T\right] + \mathrm{cov}(\boldsymbol{v}_j)) \circledast (\mathbb{E}\left[\boldsymbol{x}_t\right]\mathbb{E}\left[\boldsymbol{x}_t^T\right] + \mathrm{cov}(\boldsymbol{x}_t)),\end{aligned} \tag{22}$$

here, the notation $\mathrm{cov}(\cdot)$ denotes the covariance matrix of a vector.

In order to update the factor matrices V and X, we can do the same with vectors $\boldsymbol{v}_j, j \in \{1, 2, ..., n\}$ and $\boldsymbol{x}_t, t \in \{1, 2, ..., f\}$ while referring to $\boldsymbol{u}_i, i \in \{1, 2, ..., m\}$.

3.3.5. The variational posterior distribution of $\{(\boldsymbol{\mu}_u, \Lambda_u), (\boldsymbol{\mu}_v, \Lambda_v), (\boldsymbol{\mu}_x, \Lambda_x)\}$

According to Eq. (12), by taking derivative of Eq. (13) with respect to $(\boldsymbol{\mu}_u, \Lambda_u)$, the variational posterior $q(\boldsymbol{\mu}_u, \Lambda_u)$ can be analytically derived as

$$\begin{aligned}\ln q(\boldsymbol{\mu}_u, \Lambda_u) =& \mathbb{E}_{q(\Theta\setminus\boldsymbol{\mu}_u,\boldsymbol{\Lambda}_u)}[\ln p(\mathcal{Y}_\Omega, \Theta)] + \mathrm{const}\\ =& \frac{1}{2}\ln|\Lambda_u| - \frac{1}{2}(\boldsymbol{\mu}_u - \frac{m\bar{\boldsymbol{u}} + \beta_0\boldsymbol{\mu}_0}{m+\beta_0})^T\left[(m+\beta_0)\Lambda_u\right](\boldsymbol{\mu}_u - \frac{m\bar{\boldsymbol{u}} + \beta_0\boldsymbol{\mu}_0}{m+\beta_0})\\ &+ \frac{1}{2}(m + \nu_0 - r - 1)\ln|\Lambda_u|\\ &- \frac{1}{2}\mathrm{tr}((W_0^{-1} + \sum_{i=1}^{m}(\mathbb{E}\left[\boldsymbol{u}_i\right] - \bar{\boldsymbol{u}})(\mathbb{E}\left[\boldsymbol{u}_i\right] - \bar{\boldsymbol{u}})^T + \frac{m\beta_0}{m+\beta_0}(\bar{\boldsymbol{u}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{u}} - \boldsymbol{\mu}_0)^T)\Lambda_u) + \mathrm{const},\end{aligned} \tag{23}$$

recall that there is a Gaussian-Wishart prior placing on the hyper-parameters $(\boldsymbol{\mu}_u, \Lambda_u)$ as described in Eq. (8), we therefore have the variational posterior $q(\boldsymbol{\mu}_u, \Lambda_u) = \mathcal{N}(\boldsymbol{\mu}_u \mid \tilde{\boldsymbol{\mu}}_u^*, (\tilde{\beta}_u^*\Lambda_u)^{-1})\mathcal{W}(\Lambda_u \mid \tilde{W}_u^*, \tilde{\nu}_u^*)$ as follows

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_u^* &= \frac{m\bar{\boldsymbol{u}} + \beta_0\boldsymbol{\mu}_0}{m+\beta_0}, \tilde{\beta}_u^* = \beta_0 + m, \tilde{\nu}_u^* = \nu_0 + m,\\ (\tilde{W}_u^*)^{-1} &= W_0^{-1} + \sum_{i=1}^{m}(\mathbb{E}\left[\boldsymbol{u}_i\right] - \bar{\boldsymbol{u}})(\mathbb{E}\left[\boldsymbol{u}_i\right] - \bar{\boldsymbol{u}})^T + \frac{m\beta_0}{m+\beta_0}(\bar{\boldsymbol{u}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{u}} - \boldsymbol{\mu}_0)^T,\end{aligned} \tag{24}$$

where $\bar{\boldsymbol{u}} = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\boldsymbol{u}_i\right]$.

In such case, the Eqs. (23) and (24) can help us to derive the variational posterior of $(\boldsymbol{\mu}_v, \Lambda_v)$ and $(\boldsymbol{\mu}_x, \Lambda_x)$.

3.3.6. The variational posterior distribution of τ

Consider the precision term $\tau \in \mathbb{R}$ which controls all tensor elements, we write its variational posterior referring to the above derivations as

$$\begin{aligned}\ln q(\tau) =& \mathbb{E}_{q(\Theta\setminus\tau)}[\ln p(\mathcal{Y}_\Omega, \Theta)] + \mathrm{const}\\ =& (a_0 + \frac{1}{2}\sum_{(i,j,t)\in\Omega} o_{ijk} - 1)\ln\tau - (b_0 + \frac{1}{2}\sum_{(i,j,t)\in\Omega}\mathbb{E}\left[(y_{ijt} - g_{ijt})^2\right])\tau + \mathrm{const},\end{aligned} \tag{25}$$

and it is straightforward to have the variational posterior $q(\tau) = \mathrm{Gamma}(\tilde{a}_\tau, \tilde{b}_\tau)$ as follows

$$\begin{aligned}\tilde{a}_\tau &= a_0 + \frac{1}{2}\sum_{(i,j,t)\in\Omega} o_{ijk},\\ \tilde{b}_\tau &= b_0 + \frac{1}{2}\sum_{(i,j,t)\in\Omega}\mathbb{E}\left[(y_{ijt} - g_{ijt})^2\right],\end{aligned} \tag{26}$$

where we define $g_{ijt} = \mu + \phi_i + \theta_j + \eta_t + \sum_{k=1}^{r} u_{ik}v_{jk}x_{tk}$.

3.3.7. Lower bound of model evidence

The lower bound plays an essential role in the variational Bayes derivations. If in some cases we want to maximize the marginal probability, we can instead maximize its lower bound. As a result, when using variational Bayes to implement a tensor factorization model, we can check the value of lower bound to determine the convergence of the algorithm because $\mathcal{L}(q)$ at each iteration should increase sequentially. To be specific, the lower bound regarding Eq. (10) is given by

$$\begin{aligned}\mathcal{L}(q) =&\mathbb{E}_{q(\Theta)}\left[\ln p(\mathcal{Y}_\Omega,\Theta)\right]+H(q(\Theta))\\ =&\mathbb{E}_q\left[\ln p(\mathcal{Y}_\Omega\mid\Theta)\right]+\mathbb{E}_q\left[\ln p(\mu)\right]+\mathbb{E}_q\left[\ln p(\boldsymbol{\phi})\right]+\mathbb{E}_q\left[\ln p(\boldsymbol{\theta})\right]+\mathbb{E}_q\left[\ln p(\boldsymbol{\eta})\right]\\ &+\mathbb{E}_q\left[\ln p(U\mid\boldsymbol{\mu}_u,\Lambda_u)\right]+\mathbb{E}_q\left[\ln p(V\mid\boldsymbol{\mu}_v,\Lambda_v)\right]+\mathbb{E}_q\left[\ln p(X\mid\boldsymbol{\mu}_x,\Lambda_x)\right]\\ &+\mathbb{E}_q\left[\ln p(\boldsymbol{\mu}_u,\Lambda_u)\right]+\mathbb{E}_q\left[\ln p(\boldsymbol{\mu}_v,\Lambda_v)\right]+\mathbb{E}_q\left[\ln p(\boldsymbol{\mu}_x,\Lambda_x)\right]+\mathbb{E}_q\left[\ln p(\tau)\right]\\ &-\mathbb{E}_q\left[\ln q(\mu)\right]-\mathbb{E}_q\left[\ln q(\boldsymbol{\phi})\right]-\mathbb{E}_q\left[\ln q(\boldsymbol{\theta})\right]-\mathbb{E}_q\left[\ln q(\boldsymbol{\eta})\right]\\ &-\mathbb{E}_q\left[\ln q(U)\right]-\mathbb{E}_q\left[\ln q(V)\right]-\mathbb{E}_q\left[\ln q(X)\right]\\ &-\mathbb{E}_q\left[\ln q(\boldsymbol{\mu}_u,\Lambda_u)\right]-\mathbb{E}_q\left[\ln q(\boldsymbol{\mu}_v,\Lambda_v)\right]-\mathbb{E}_q\left[\ln q(\boldsymbol{\mu}_x,\Lambda_x)\right]-\mathbb{E}_q\left[\ln q(\tau)\right],\end{aligned}\tag{27}$$

where all expectations are with respect to the posterior distribution q. The first term is an expectation of the joint distribution. The second to the eighth terms are the expectations of log-priors over the global parameter, bias vectors, and factor matrices. The ninth to the eleventh terms denote the expectations of log-priors over hyper-parameters. The twelfth term is the expectation of log-prior over τ. In addition, the last 11 terms are entropy of the posterior distribution q over Θ.

3.4. Implementing BATF

In above, since our posterior inference based tensor factorization is inferred in a variational Bayesian framework, the question is how to learn our interested parameters $\{\mu,\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\eta},U,V,X\}$ (i.e., global parameter, bias vectors, and factor matrices) from the partially observed tensor $\mathcal{Y}_\Omega$. The feasible solution is by updating the model parameters and hyper-parameters (see Fig. 3) alternatively. We can trace back to the above derivations and see more details about it from Algorithm 1.

Algorithm 1 Bayesian augmented tensor factorization (BATF)

Input: incomplete data tensor $\mathcal{Y}_\Omega\in\mathbb{R}^{m\times n\times f}$, indicator tensor $\mathcal{O}\in\mathbb{R}^{m\times n\times f}$, global parameter μ, bias vectors $\{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\eta}\}$ and factor matrices $\{U,V,X\}$.
Output: estimated tensor $\hat{\mathcal{Y}}\in\mathbb{R}^{m\times n\times f}$, and updated μ, $\{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\eta}\}$ and $\{U,V,X\}$.
Initialize $\tau,a_0,b_0,\beta_0=1$, $\mu_0=0,\tau_0=1$, $\nu_0=r$, $\boldsymbol{\mu}_0=\mathbf{0}$, and $W_0=I$ (identity matrix).
repeat
  Update the posterior of global parameter $q(\mu)$ using Eq. (15).
  Update the posterior of hyper-parameters $q(\boldsymbol{\mu}_u,\Lambda_u)$, $q(\boldsymbol{\mu}_v,\Lambda_v)$ and $q(\boldsymbol{\mu}_x,\Lambda_x)$ using Eq. (24) and its similar inference results.
  for $i=1$ to m **do**
    Update the posterior of bias $q(\phi_i)$ using Eq. (18).
    Update the posterior of factor $q(\boldsymbol{u}_i)$ using Eq. (21).
  end for
  for $j=1$ to n **do**
    Update the posterior of bias $q(\theta_j)$ similar to Eq. (18).
    Update the posterior of factor $q(\boldsymbol{v}_j)$ similar to Eq. (21).
  end for
  for $t=1$ to f **do**
    Update the posterior of bias $q(\eta_t)$ similar to Eq. (18).
    Update the posterior of factor $q(\boldsymbol{x}_t)$ similar to Eq. (21).
  end for
  Update the posterior of precision $q(\tau)$ using Eq. (26).
  Evaluate the lower bound $\mathcal{L}(q)$ using Eq. (27).
until convergence.

4. Experiments

In this section, our goal is to learn an expressive representation of urban traffic state that is semantically meaningful, so that we can identify both explicit and latent patterns in the data. To this end, we carry on a wide range of empirical examinations to broadly investigate the usefulness of BATF. Relying on the urban traffic speed

data set, we first evaluate how well BATF works for tensor completion compared to the baseline models. We then survey the learned latent factors as well as the explicit patterns, and further show the semantic interpretations of each one and their combination. Finally, we demonstrate the robustness of BATF in the task of missing data imputation under different missing scenarios with varying missing rates. Our results suggest that BATF can capture significantly more expressive representations than baselines.

4.1. Details of experiment setting

Data set. We utilize a publicly available traffic speed data set (see https://doi.org/10.5281/zenodo.1205229) which is evaluated in our recent works (Chen et al., 2018, 2019). This data set is collected from 214 road segments in Guangzhou, China within two months (i.e., 61 days from August 1, 2016 to September 30, 2016) at 10-minute interval (144 time intervals per day). The speed data can be organized as a third-order tensor (road segment $\times$ day $\times$ time interval, with a size of $214 \times 61 \times 144$). There are about 1.29% missing values in the raw data set.

Experiment setup. The main task of this work is missing data imputation, therefore, we first follow two missing data scenarios including random missing and non-random (fiber) missing devised by Chen et al. (2018). Then, we set our tensor completion task with 10%, 30% and 50% missing rates under both two scenarios. When training BATF model, we use rank $r = 80$ in the case of random missing. In order to prevent overfitting with non-random missing, we consider rank $r = 20, 15, 10$ for BATF model at 10%, 30% and 50% missing rates, respectively. In terms of convergence, the maximum epoch for BATF model is set to 200. The Matlab code is available at https://github.com/sysuits/BATF.

Performance metrics. The mean absolute percentage error (MAPE) and root mean square error (RMSE) are used to evaluate model performance as follows

$$\begin{aligned} \text{MAPE} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}, \\ \text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \end{aligned} \tag{28}$$

where N is the total number of missing entries, and y_i and $\hat{y}_i$ are the actual value of a missing entry and its imputation, respectively.

Baselines. We consider two fully Bayesian tensor factorization models, the Bayesian CP factorization (BCPF, Zhao et al. (2015a)) and the Bayesian Gaussian CP decomposition (BGCP, Chen et al. (2019)), as evaluation baselines. BCPF and BGCP are implemented by variational Bayes and MCMC, respectively. Moreover, the Bayesian network of BGCP is more flexible than BCPF.

4.2. Performance of missing data imputation

With the above settings, we compare the proposed BATF model to three state-of-the-art models, including BCPF (Zhao et al., 2015b), BGCP (Chen et al., 2019) and STD (Chen et al., 2018). Table 1 shows the imputation performance of these models where BATF, BCPF and BGCP share the same rank r. Note that the comparison between BGCP and other models (e.g., daily average, kNN and HaLRTC) was demonstrated at the work of Chen et al. (2019). In this study, we only investigate the imputation performance of tensor based models.

Our first experiment examines the performance of different models in the random missing scenario. One can easily find that the Bayesian tensor factorization models have significant improvement over STD and are less sensitive to the increasing missing rate. Thus, it also supports that Bayesian inference methods for tensor factorization are effective for dealing with the sparsity issue Zhao et al. (2015a). Thanks to the flexible conjugate prior setting, BATF and BGCP get slightly better results than BCPF as they have more parameters to fit the data. However, when the tensor behaves with an increasing amount of missing data, these models accordingly exhibit growing errors.

In the second experiment, we present imputation performance in the non-random missing scenario, which is a more realistic temporally correlated scenario following Chen et al. (2018). Since Bayesian tensor factorization models are sensitive to the rank parameter, we choose the rank r as 20, 15 and 10 for the missing rate of 10%, 30% and 50%, respectively. From the comparison, we see that our BATF performs better than other two models, which shows the structural benefit of augmented tensor factorization. The results of Table 1 also suggest that the presentation learned by BATF is significantly more capable of imputing missing data than other competing models, and BATF's results are also less sensitive to the increasing missing rate.

Another thing should be further discussed is the completeness. Due to the temporally correlated corruption in the non-random missing scenario, it becomes difficult to utilize the algebraic structure and collaborative information. Comparing to the random missing scenario, we can find that the errors at the non-random missing scenario are relatively higher. Even with the same missing rate, it is easy to see that the non-random missing scenario is more difficult to tackle. In practice, we can observe an indication that BCPF fails to work in the non-random missing scenario (see Table 1).

Table 1: MAPE/RMSE scores of tensor completion models for the urban traffic speed data set.

	Random missing			Non-random missing		
	10% ($r = 80$)	30% ($r = 80$)	50% ($r = 80$)	10% ($r = 20$)	30% ($r = 15$)	50% ($r = 10$)
BATF	0.0825/3.5745	0.0834/3.5969	0.0841/3.6290	0.0976/4.1252	0.0995/4.2256	0.1029/4.3557
BCPF	0.0832/3.5988	0.0843/3.6340	0.0852/3.6784	-	-	-
BGCP	0.0823/3.5614	0.0827/3.5775	0.0833/3.6009	0.0980/4.1413	0.0999/4.2425	0.1048/4.4419
STD	0.0888/3.7708	0.0936/3.9286	0.0993/4.1253	0.1019/4.1881	0.1068/4.4029	0.1133/4.6291

4.3. Semantic interpretations of BATF

In this study, we are interested in BATF having not only imputation power but also discovering patterns. To provide more insights into the effectiveness of BATF, we explore the semantic interpretations of BATF in real experiments. First, we start by summarizing the explicit patterns of BATF (see Fig. 4). Fig. 4(a) presents the curves of global parameters of BATF by running 30 times. It is rather intuitive to find that the average of 30 global parameter curves are extremely close to the actual mean of observations (i.e., 39.01 km/h). Therefore, the global parameter controlling all tensor elements is used to approximate the mean standard of partially observed multi-dimensional data.

Fig. 4(b) and Fig. 4(c) give the value of biases corresponding to 214 road segments and 144 time intervals separately. We observe that there are about half of road segments obtaining biases above 0 and up to 20 km/h, meanwhile, other road segments obtaining biases between -15 km/h and 0 km/h (see Fig. 4(b)). The value of bias also has its real-world meanings. For example, if one road segment has a relatively large (positive) bias, we can generally say that the road segment is more fluent than the network's standard. In other words, the bias of one road segment is a relative value over the global average.

Fig. 4(c) illustrates that there are negative biases in the daytime as positive biases in the night. To be specific, the bias reaches its lowest during the evening peak hours, and the bias is relatively larger in the morning peak hours. Fig. 4(d) shows the heatmap of summed biases over day and time interval dimensions, and it can help us understand the time-evolving patterns from these two dimensions together. On the other hand, these findings are also consistent with the daily trend of traffic state provided by Chen et al. (2018).

To reinforce our interpretation that these explicit patterns are semantically meaningful, in Fig. 5, we present an example which covers the actual time series and its imputation of road segment #1. The simple combination of explicit patterns (i.e., global parameter and biases) provides rough trends for actual time series. Further taking explicit patterns and latent factors together, we can see that the estimated time series using BATF is closer to the actual time series. Thus, in terms of explicit patterns, our newly formulated Eq. (4) has more semantically meaningful representations than the conventional tensor factorization formulation (see Eq. (1)).

Regarding the failure of BCPF in the non-random missing scenario (see Table 1), we choose the experiment for BCPF at the 30% missing rate with its rank being $r = 5, 10$. Fig. 6 presents the RMSE and lower bound value of BCPF for investigating the train-test performance. In Table 1, it is worth noting that BCPF cannot work when setting the same rank r to BATF and BGCP models. However, observing Fig. 6(b), BCPF which places a smaller rank still suffers from the overfitting issue. There are some RMSE curves at a wrong direction as we see increasing lower bound curves. On the other hand, when setting $r = 5$, Fig. 6(a) shows the uncertainty of RMSEs (about 0.20 km/h), which illustrates that BCPF is not technically feasible in such case.

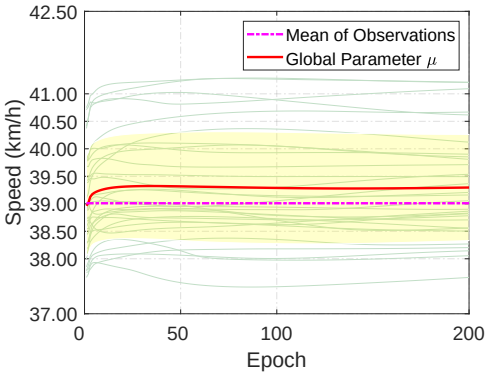
5. Conclusion

In this study, we propose an augmented tensor factorization with fully Bayesian treatment to impute the missing traffic data accurately. First, the factorization based on Bayesian inference is less sensitive to the data sparsity where the results reported by Bayesian tensor factorization models are in effect more tolerant to the increasing missing rate (see Table 1). Then, from the empirical studies, when setting the non-random missing rate ranging from 10% to 50%, we demonstrated that BATF performs best among its competing models. At the random missing scenario, BATF also achieves competitive imputation results.

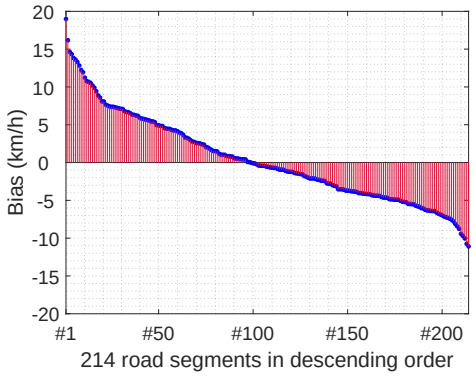
Finally, as our experiments demonstrated, competing tensor factorization models failed to capture explicit patterns and their application scenario is limited because of our complex data and the overfitting issue. Instead, the proposed BATF achieves generalization performance of Bayesian tensor factorization and combines explicit patterns and latent factors together. Our formulation (see Eq. (4)) incorporating generic forms of domain knowledge also provide more insights into the effectiveness of tensor factorization.

Acknowledgement

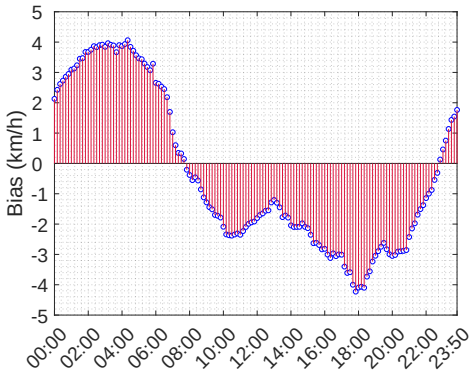
The authors would like to thank anonymous referees for their valuable comments. This research is supported by the project of National Natural Science Foundation of China (No. U1811463), the Science and Technology



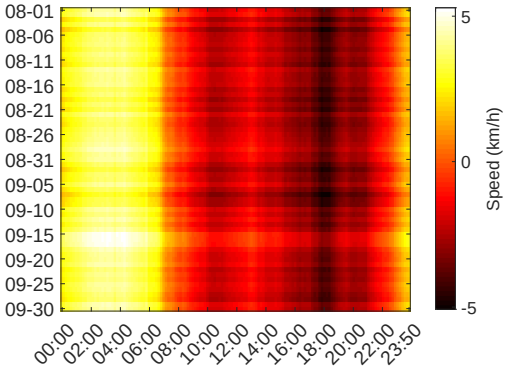
(a) Global parameter at each epoch.



(b) Biases of 214 road segments.

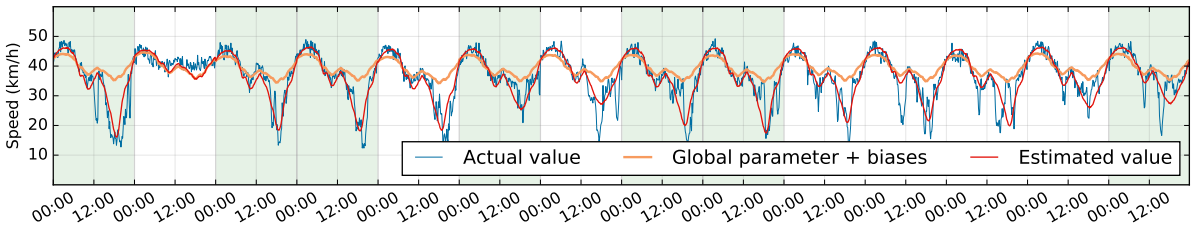


(c) Biases of 144 time intervals.

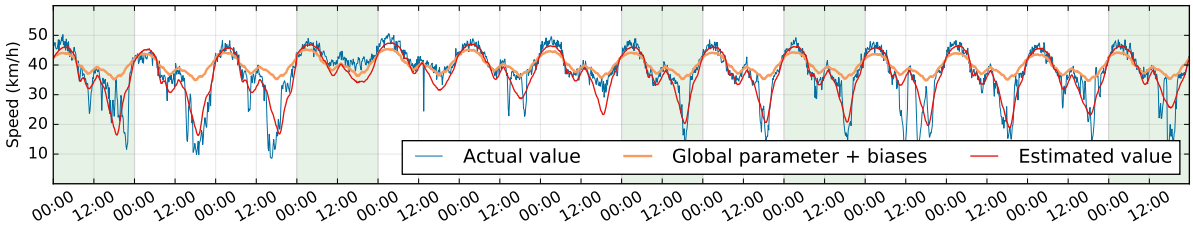


(d) Summed biases over day and time interval dimensions.

Figure 4: The explicit patterns (i.e., global parameter and biases) of BATF at the 50% non-random missing rate with $r = 10$.

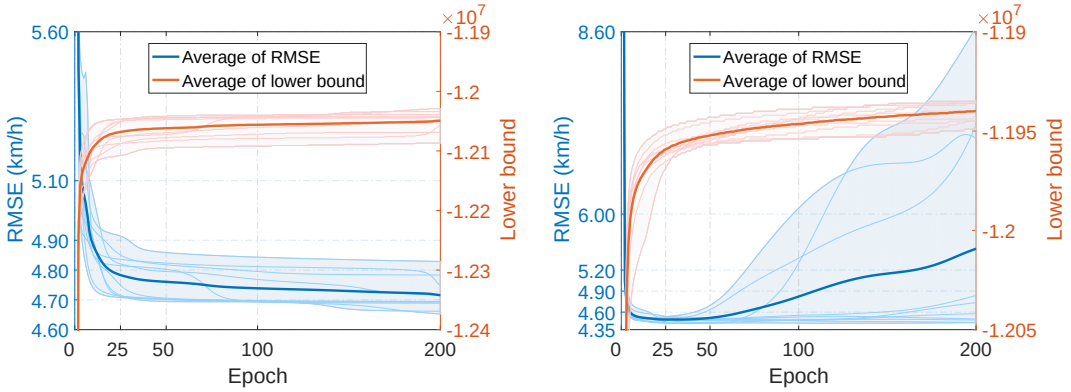


(a) Time series of actual and estimated speed within two weeks from August 1 to 14.



(b) Time series of actual and estimated speed within two weeks from September 12 to 25.

Figure 5: The imputation performance of BATF at the 50% non-random missing rate with $r = 10$, where the estimated result of road segment #1 is selected as an example. In the both two panels, white rectangles represent fiber missing (i.e., speed observations are lost in a whole day), and green rectangles indicate partially observed data.



(a) Rank $r = 5$.

(b) Rank $r = 10$.

Figure 6: RMSEs and lower bound values of BCPF model ran 10 times at the 30% missing rate.

Planning Project of Guangzhou, China (No. 201804020012), and the Natural Science Foundation of Guangdong Province, China (No. 20187616042030004).

References

Asif, M. T., Mitrovic, N., Dauwels, J., Jaillet, P., 2016. Matrix and tensor based methods for missing data estimation in large traffic networks. IEEE Transactions on Intelligent Transportation Systems 17 (7), 1816–1825.

Chen, X., He, Z., Sun, L., 2019. A bayesian tensor decomposition approach for spatiotemporal traffic data imputation. Transportation Research Part C: Emerging Technologies 98, 73 – 84.
URL http://www.sciencedirect.com/science/article/pii/S0968090X1830799X

Chen, X., He, Z., Wang, J., 2018. Spatial-temporal traffic speed patterns discovery and incomplete data recovery via svd-combined tensor decomposition. Transportation Research Part C: Emerging Technologies 86, 59–77.

Hu, C., Rai, P., Chen, C., Harding, M., Carin, L., 2015. Scalable bayesian non-negative tensor factorization for massive count data. In: Proceedings, Part II, of the European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9285. ECML PKDD 2015. Springer-Verlag New York, Inc., New York, NY, USA, pp. 53–70.
URL http://dx.doi.org/10.1007/978-3-319-23525-7_4

Kolda, T. G., Bader, B. W., 2009. Tensor decompositions and applications. SIAM Reviw 51 (3), 455–500.

Koren, Y., Bell, R., Volinsky, C., Aug 2009. Matrix factorization techniques for recommender systems. Computer 42 (8), 30–37.

Laa, I., Olabarrieta, I. I., Vlez, M., Ser, J. D., 2018. On the imputation of missing data for road traffic forecasting: New insights and novel techniques. Transportation Research Part C: Emerging Technologies 90, 18 – 33.
URL http://www.sciencedirect.com/science/article/pii/S0968090X18302535

Li, L., Li, Y., Li, Z., 2013. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transportation research part C: emerging technologies 34, 108–120.
URL http://dx.doi.org/10.1016/j.trc.2013.05.008

Liu, J., Musialski, P., Wonka, P., Ye, J., 2013. Tensor completion for estimating missing values in visual data. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1), 208–220.

Qu, L., Li, L., Zhang, Y., Hu, J., 2009. Ppca-based missing data imputation for traffic flow volume: A systematical approach. IEEE Transactions on Intelligent Transportation Systems 10 (3), 512–522.

Qu, L., Zhang, Y., Hu, J., Jia, L., Li, L., June 2008. A bpca based missing value imputing method for traffic flow volume data. In: 2008 IEEE Intelligent Vehicles Symposium. pp. 985–990.

Rai, P., Hu, C., Harding, M., Carin, L., 2015. Scalable probabilistic tensor factorization for binary and count data. In: Proceedings of the 24th International Conference on Artificial Intelligence. IJCAI'15. AAAI Press, pp. 3770–3776.
URL http://dl.acm.org/citation.cfm?id=2832747.2832775

Rai, P., Wang, Y., Guo, S., Chen, G., Dunson, D., Carin, L., 2014. Scalable bayesian low-rank decomposition of incomplete multiway tensors. In: Xing, E. P., Jebara, T. (Eds.), Proceedings of the 31st International Conference on Machine Learning. Vol. 32 of Proceedings of Machine Learning Research. PMLR, Bejing, China, pp. 1800–1808.
URL http://proceedings.mlr.press/v32/rai14.html

Ran, B., Tan, H., Wu, Y., Jin, P. J., 2016. Tensor based missing traffic data completion with spatial-temporal correlation. Physica A: Statistical Mechanics and its Applications 446, 54–63.

Rodrigues, F., Henrickson, K., Pereira, F. C., 2018. Multi-output gaussian processes for crowdsourced traffic data imputation. IEEE Transactions on Intelligent Transportation Systems, 1–10.

Salakhutdinov, R., Mnih, A., 2008. Bayesian probabilistic matrix factorization using markov chain monte carlo. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08. ACM, New York, NY, USA, pp. 880–887.
URL http://doi.acm.org/10.1145/1390156.1390267

Schein, A., Zhou, M., Blei, D., Wallach, H., 20–22 Jun 2016. Bayesian poisson tucker decomposition for learning the structure of international relations. In: Balcan, M. F., Weinberger, K. Q. (Eds.), Proceedings of The 33rd International Conference on Machine Learning. Vol. 48 of Proceedings of Machine Learning Research. PMLR, New York, New York, USA, pp. 2810–2819.
URL http://proceedings.mlr.press/v48/schein16.html

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y.-J., Li, F., 2013a. A tensor-based method for missing traffic data completion. Transportation Research Part C: Emerging Technologies 28, 15 – 27, euro Transportation: selected paper from the EWGT Meeting, Padova, September 2009.
URL http://www.sciencedirect.com/science/article/pii/S0968090X12001532

Tan, H., Yang, Z., Feng, G., Wang, W., Ran, B., 2013b. Correlation analysis for tensor-based traffic data imputation method. Procedia-Social and Behavioral Sciences 96, 2611–2620.

Xiong, L., Chen, X., Huang, T.-K., Schneider, J., Carbonell, J. G., 2010. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In: Proceedings of the 2010 SIAM International Conference on Data Mining. SIAM, pp. 211–222.
URL http://epubs.siam.org/doi/abs/10.1137/1.9781611972801.19

Zhao, Q., Zhang, L., Cichocki, A., 2015a. Bayesian cp factorization of incomplete tensors with automatic rank determination. IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (9), 1751–1763.

Zhao, Q., Zhang, L., Cichocki, A., 2015b. Bayesian sparse tucker models for dimension reduction and tensor completion. CoRR abs/1505.02343.
URL http://arxiv.org/abs/1505.02343