Transportation Research Part C 124 (2021) 102916

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Integrating probabilistic tensor factorization with Bayesian supervised learning for dynamic ridesharing pattern analysis

Zheng Zhu [a], Lijun Sun [b], Xiqun Chen [c,d,*], Hai Yang [a]

[a] Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology, Kowloon, Hong Kong
[b] Department of Civil Engineering and Applied Mechanics, McGill University, Montreal, Quebec, Canada
[c] College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, China
[d] Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Hangzhou, China

ARTICLE INFO

Keywords:
Ridesharing
Classification
Supervised learning
Tensor factorization
Latent class analysis

ABSTRACT

In the era of transportation big data, the analysis of mobility patterns generally involves large quantities of datasets with high-dimensional variables recording individual travelers' activities and socio-economic attributes, bringing new challenges to researchers. Conventional regression-based models commonly require complex structures in depicting random or fixed effects with a considerable number of parameters to estimate, and state-of-the-art machine learning models are regarded as black-boxes that are not clear in interpreting the mechanism in human mobility. To overcome the challenges of capturing complex high-order relationships among variables of interest, this paper proposes a Bayesian supervised learning tensor factorization (BSTF) model for the classification of travel choices in the mobility pattern analysis. The BSTF model induces a hierarchical probabilistic structure between predictor variables and the dependent variable, which offers a nature supervised learning foundation via Bayesian inference. Latent class (LC) variables are considered in the BSTF model to discover hidden preferences/states among travelers associated with their mobility patterns. We apply the BSTF model to analyze passenger-side choice patterns between diverse service options on a ride-sourcing platform, drawing increasing attention during recent years. A case study with a real-world dynamic ridesharing dataset in Hangzhou, China, is conducted. Different cases of training sizes are utilized to fit the proposed BSTF model as well as some other state-of-the-art machine learning models. By identifying significant variables and derive their probabilistic relationship between service types (i.e., ridesharing, non-sharing, and taxi), the proposed BSTF model offers good performance in both classification accuracy and the interpretability of shared mobility.

1. Introduction

Modeling mobility patterns is an essential component in transportation planning and travel demand analysis. Accurate prediction of mobility attributes, such as aggregate travel mode choice and destination choice, is critical to a wide range of operations and planning applications, such as traffic management, congestion mitigation, and transportation system design. With the rapid advances in transportation systems, understanding the patterns of travel choices becomes increasingly important due to the introduction of new services/modes (e.g., scooters, electric bikes, and shared mobility) and the changes in people's lifestyle and preferences (e.g.,

* Corresponding author at: College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, China.
E-mail address: chenxiqun@zju.edu.cn (X. Chen).

https://doi.org/10.1016/j.trc.2020.102916
Received 2 December 2019; Received in revised form 29 November 2020; Accepted 30 November 2020
Available online 25 December 2020

encouraging sustainable transport).

One of the main tasks in analyzing mobility patterns is to build classification models which can (a) capture the complex dependencies and interactions among a set of predictor variables such as trip attributes (e.g., travel time, and distance), and socio-economic attributes (e.g., land-use attributes of the origin/destination); and (b) characterize the underlying human mobility mechanism based on behavior or probability theorems. A good model is expected to identify critical attributes from the broad set of variables and characterize the interactions/dependencies among them. Researchers in transportation and economics have begun to study mobility patterns with respect to travelers' decision-making behaviors since the introduction of discrete choice models in the 1970s (McFadden, 1973). A large number of choice models have been developed based on different techniques, including psychological, economic, and supervised machine learning approaches (e.g., Ben-Akiva and Lerman, 1985; Hensher and Ton, 2000; Train, 2009; Dias et al., 2017; Zhu et al., 2019a, 2021). These models are fundamental to the establishment of human mobility science.

Thanks to the fast development of information and communication technologies, large quantities of temporal and spatial datasets that record individual travelers' activities and socio-economic attributes become available. In the analysis of travel choices and mobility patterns, since a categorical setting is usually required for variables such as income, education, point of interest, mode type, enlarged dimensionality, and sample size in mobility datasets, bring new challenges for model development. Conventional regression-based models become less appealing due to their limited capability in dealing with the nonlinearity and higher-order dependencies among a large number of attributes. One needs to include model parameters for each categorical level of the categorical variables; many more parameters are created when modeling the interactions among these variables. Supervised learning models—such as Neural Networks (NN) and Random Forest (RF)—provide a data-driven paradigm to characterize higher-order interactions in mobility analysis problems, and these models have shown superior performance compared to traditional regression-based models (e.g., Hensher and Ton, 2000; Zhang and Xie, 2008; Zhu et al., 2018). However, due to the data-driven nature, the theoretical basis of these machine learning models for interpreting and understanding the human mobility mechanism becomes an emerging concern that prevents practitioners from adopting them in real-world applications (Brathwaite et al., 2017). Some researchers have tried to conduct dimensionality reduction, such as principal component analysis (PCA), to the datasets, and then used the first few principal components to build a simple regression model (Miller and Mohammadian, 2003). However, the components are not precise in interpretation, and the accuracy is not good due to the missing information. Brathwaite et al. (2017) provided the microeconomic basis for decision tree (DT) models in travel decision-making. More efforts are awaited, and a significant and urgent research issue is to understand people's modern travel patterns and make accurate choice classification simultaneously.

In this paper, we propose to integrate probabilistic tensor factorization with Bayesian supervised learning for travel choice and mobility pattern analysis. By capturing the nonlinear and higher-order dependencies among the predictor variables, the proposed Bayesian supervised learning tensor factorization (BSTF) model offers both high classification accuracy interpretability on mobility patterns. The BSTF model incorporates latent class (LC) structures to provide hidden probabilistic dependencies between travelers' choices and predictor (manifest) variables. A mobility mechanism is built upon the Bayesian inference for the LC variables and the dependent variable given the predictor variables. This hierarchical structure allows us to build efficient Markov chain Monte Carlo (MCMC) sampling algorithms for model estimation. A categorical variable setting is used in this BSTF framework to better interpret the high-order dependencies via probability tables.

We apply the BSTF model to analyze the mobility pattern of ridesharing, which has a high-dimensional dataset with categorical variables. At present, the percentage of ridesharing orders in ride-sourcing markets is still low (Li et al., 2019). It was pointed out that a good understanding of passengers' preferences is essential to the success of a dynamic ridesharing program (Agatz et al., 2012). As demonstrated in previous studies, the passengers' service type choice will, in return, affect the operation and performance of each ride-sourcing or other service options (Ke et al., 2017; Zhu et al., 2020). Moreover, the modeling results also help researchers and TNCs better understand the aggregate performance (i.e., patterns of passenger waiting time and travel time) of the ridesharing program. Pre-trip information provision can be implemented by the platform to attract more ridesharing passengers. However, far efforts have primarily been directed toward the design of algorithms to efficiently match drivers and passengers on short notice in a dynamic ridesharing environment (e.g., Agatz et al., 2012; Wang et al., 2017). More research efforts are needed to identify critical factors affecting passengers' willingness to "share a ride" and examine their quantitative effects on ridesharing decisions (Dias et al., 2017). As a real-world application, Didi Chuxing's order data and land-use GIS data in Hangzhou, China, are fused for the ridesharing mobility pattern analysis. We evaluate the performance of BSTF with respect to the classification accuracy and compare it with other state-of-the-art supervised machine learning models as benchmarks. Also, we look into travelers' preferences via the latent patterns and interpret the choice-making mechanism of dynamic ridesharing in the ride-sourcing market.

The major contributions of this paper include: (a) proposing a supervised learning approach, i.e., the BSTF model, which provides a closed-form hierarchical probabilistic structure and leads to high classification accuracy in mobility pattern analysis studies; (b) interpreting and understanding passengers' dynamic ridesharing patterns via the hierarchically probabilistic structure of the BSTF model. The BSTF model is reliable in solving classification problems with high-dimensional and massive datasets, which is not limited to mobility analysis. Additionally, the LC structure and the underlying conditional probabilistic relationship offers an intuitive way to understand the complex human mobility mechanism behind critical predictor variables and travelers' choices.

The remainder of the paper is organized as follows. Section 2 presents a literature review on travel choice models and travel decision classification models, as well as ridesharing studies based on these models. We also review some signature applications of tensor factorization (TF) in transportation research. Section 3 presents the proposed BSTF model, including its key assumption, formulation, and MCMC-based model estimation. Section 4 undertakes a real-world case study on dynamic ridesharing pattern analysis with DiDi ride-sourcing order data. The introduction of the dataset, the classification accuracy of the BSTF model and the benchmarks, and the characterized mechanism of passengers' ridesharing choices via probabilistic inference are covered. Finally, Section 5 concludes this

paper and discusses future research directions.

2. Literature review

In this section, we first explore the state-of-the-art methodologies on travel choice modeling and classification. Identifying research gaps leads us to propose a classification approach that is superior in both model interpretability and classification accuracy. We also provide existing studies on ridesharing.

Researchers in transportation and economics have begun to study travelers' decision-making mechanism since the introduction of the multinomial Logit (MNL) model in the 1970s (McFadden, 1973). Originated from econometric formulation, the behavior foundation of the MNL model is restricted to the random utility (RU) theorem, which assumes that a person chooses the alternative with the maximal utility. The contribution of each predictor variable towards the dependent variable is clear through the linear formulation of utility. The estimation of utility function parameters can help understand the relative importance of different travel attributes and socio-economic variables in decision-making. The model also considers the hidden causal factors not included in the utility function as one alternative-specific constant. The MNL family has been extended by considering the correlation term between different alternatives, including nested Logit models (de Dios Ortuzar, 1983), generalized extreme value model (Small, 1987), Bayesian nested Logit (Poirier, 1996), continuous cross-nested Logit models (Lemp et al., 2010), LC Logit models (Kamargianni et al., 2015), and continuous-discrete choice models (Bhat, 2018).

Another major trend of travel choice modeling is to use machine learning techniques to classify a traveler's choice given predictor variables. Without the requirement of a predetermined model structure, machine learning focuses on data and tries to find connections among variables. The flexibility in the model structure makes it possible to offer insights into the relations that RU models cannot recognize. Machine learning methods can handle large-size datasets, saving much time in the model estimation, finding more complex relations compared with RU-based discrete choice models, and providing high classification accuracy. However, in travel choice and mobility pattern research, it is not clear to interpret humans' decision-making mechanisms by using most machine learning methods. There are several types of machine learning models frequently adopted in the classification of travelers' choices, e.g., DT models (Wets et al., 2000), NN models (Hensher and Ton, 2000), mixed Bayesian network (BN) models (Zhu et al., 2018), support vector machine (SVM) (Zhang and Xie, 2008), and ensemble learning (Chen et al., 2017).

To interpret the patterns of travelers' choices and maintain high classification accuracy, we propose a Bayesian supervised learning approach based on TF in this paper. TF is a method to summarize a high-dimensional dataset into a tensor (i.e., a high-order array). It has gained popularity in various fields such as data mining, signal processing, statistics, etc. (Shashua and Hazan, 2005; Kolda and Bader, 2009). In the field of transportation research, TF also attracts increasing attention. Tan et al. (2013) integrated the expectation–maximization algorithm with Tucker decomposition to impute the missing data in a temporal traffic dataset. CANDECOMP/PARAFAC (CP) decomposition has been used in analyzing temporal (Dunlavy et al., 2011) and spatial–temporal patterns of traffic flow (Han and Moutarde, 2016). Sun and Axhausen (2016) proposed an unsupervised probabilistic TF model to understand urban mobility patterns. The research team also integrated the probabilistic TF model with an LC model and a rejection sampling model for population synthesis (Sun et al., 2018) and utilized probabilistic TF for spatial imputation (Zhang et al., 2019; Chen et al., 2020). In contrast to unsupervised TF models, Tan et al. (2016) developed a dynamic TF model for short-term traffic prediction. Chen et al. (2019a) and Chen et al. (2019b) proposed an augmented TF model for missing traffic data imputation. In these studies, TF based models show superior performance compared with existing prediction methods. By adopting the probabilistic TF in a supervised learning approach, the proposed BSTF model provides a convenient approach for mobility patterns and other data-driven transportation research.

Based on but not limited to those above conventional discrete choice and machine learning approaches, there have been studies focusing on ridesharing mobility patterns. Miller et al. (2005) developed a tour-based mode choice mixed Logit model, in which traditional ridesharing was incorporated by adding a constraint on household vehicle allocation. Concerning travelers' intra-household interactions, activity-based logit models were developed to identify significant predictors (e.g., work schedules, auto availability, and presence of children) related to shared rides (Gliebe and Koppelman, 2002). The analyses of traditional ridesharing were primarily based on stated preference survey datasets. The coordination of traditional ridesharing was recognized as a within-household activity (Morency, 2007). In terms of non-household ridesharing, factors such as time conflicts (Giuliano, 1992; Ferguson, 1997), monetary saving (Correia and Viegas, 2011), parking (Su and Zhou, 2012), environmental awareness (Wang et al., 2020), and walking distance (Hunt and McMillan, 1997) were found of great importance. A detailed review can be found in Neoh et al. (2017), in which the authors applied meta-analysis to explore key predictors of ridesharing. Due to the availability of timely door-to-door service, participants' (i.e., both passengers' and drivers') mobility patterns under the dynamic ridesharing market can be substantially different compared with traditional ridesharing. Chen et al. (2017) applied ensemble learning to classify dynamic ridesharing decisions, and they claimed trip attributes (e.g., travel time, surge pricing ratio, trip fee, and trip distance) were critical for ridesharing classification. Dong et al. (2018) utilized an unsupervised learning approach to classify ridesharing drivers based on commuting styles and detour patterns. Exploring the modeling results of these studies, we note that: (a) classification accuracy is one of the most critical issues, for which reason machine learning approaches are preferable in recent ridesharing studies; (b) more research efforts are awaited to improve the interpretability (i.e., the basis for modeling humans' choice-making in machine learning) and to understand the quantitative influence of different predictors on travelers' dynamic ridesharing patterns. To address these issues, we apply the proposed BSTF model to real-world dynamic ridesharing analysis. The specific application also demonstrates the superiority of the BSTF model.

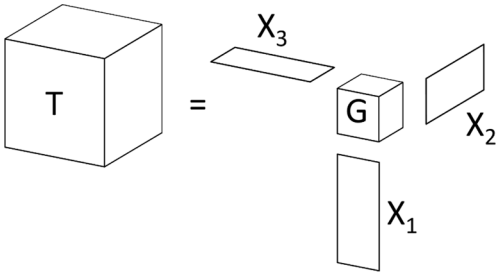

Fig. 1. An illustration of HOSVD.

3. Bayesian supervised learning tensor factorization

In this paper, identifying a traveler's choice (on ride-sourcing service type) is recognized as a classification problem, which is also a supervised learning problem. This section introduces the concept of Bayesian supervised learning and formulates the BSTF model. The analytical foundation and model estimation algorithm are discussed at last.

3.1. Problem description of Bayesian supervised learning

Statistical classification is needed across numerous scientist disciplines, in which some kinds of supervised learning algorithms are often required. The focus of supervised learning with high-dimensional predictor variables will not be directly on selecting variables but on the description of the classification and inference rules.

Let $X_1, \ldots, X_p$ denote predictor variables, and Y denote the categorical classification variable. The goal of a general classification model is to induce rule $f : x_1, \ldots, x_p \rightarrow y$, where x_j denotes the value of predictor X_j, y denotes the categorical value of Y. Unlike other classification approaches, which assume f to be some deterministic mapping from $X_1, \ldots, X_p$ to Y (e.g., NN, and SVM), Bayesian supervised learning treats $X_1, \ldots, X_p$ and Y as random variables that are sampled from some joint probability distribution. The probabilistic formulation is as follows:

$$\boldsymbol{W}|\alpha p_{\alpha}^{\sim} \tag{1}$$

where $\boldsymbol{W}$ denotes the random vector of $\{X_1, \ldots, X_p, Y\}$. The probability density function (PDF) of $\boldsymbol{W}$ is denoted as p_α, which is determined based on parameter α. Thereby, the classification rule can be expressed as a probabilistic inference below:

$$f : \underset{y \in \Upsilon}{argmax} P\left(y|x_1, \ldots, x_p\right) \tag{2}$$

where Υ denotes the set of Y, $P(\cdot)$ denotes probability operator. The conditional probability of y given $x_1, \ldots, x_p$ is calculated based on joint PDF p_α. Suppose we have a sample of n data records, i.e., $\boldsymbol{w}_i = \{x_{i,1}, \ldots, x_{i,p}, y_i\}$, $i = 1, \ldots, n$. A general approach to estimating α is to use likelihood maximization:

$$\underset{\alpha}{argmax} \sum_i L(\alpha|\boldsymbol{w}_i) \tag{3}$$

where $L(\alpha|\boldsymbol{w}_i) = p_\alpha(\boldsymbol{w}_i)$ denotes the likelihood of observing sample i.

In a travelers' choice classification problem, $X_1, \ldots, X_p$ can be a list of socio-economic variables and travel attributes variables, and Y is the categorical decision (i.e., mode, departure time or service type). The assumption on the joint probability distribution in Eq. (1) plays an essential role in classification accuracy and model interpretability.

Conventional approaches, such as MNL discrete choice models, have been preferable among researchers to solve the classification problem. Based on the formulation of utility functions with respect to different choices (e.g., modes or departure times), researchers attempt to estimate the coefficients of different features associated with the decision-making mechanism. However, once the dataset is with high-dimensional variables of interest, it can be difficult for conventional models to comprehensively consider all possible outcomes. This is because the utility functions should include not only the interested variables themselves but also the variables created based on their interactions. Thereby, many model structures and their corresponding large number of parameters need to be estimated and evaluated. To overcome the high-dimensional issue, we utilize a BSTF model for this research task. The BSTF model can characterize the high-order data structure to a lower-dimensional data of only vital predictors in a supervised learning problem.

Moreover, insights on humans' mobility preferences and patterns are usually expected in travel choice classification problems. For instance, a discrete choice model from the MNL family generally needs a hypothesis on whether travelers' choice is sensitive to some specific variables. Machine learning models, which attract increasing attention due to their high accuracy in solving classification problems, are usually treated as black-boxes. Although some of them (e.g., DT model) characterize the choice-making mechanism as a rule-based process, most state-of-the-art machine learning models do not provide deep human mobility insights. Unlikely, the BSTF

model is based on a full hypothesis in which all the variables are correlated with the dependent variable. The model estimation process of the BSTF model is to identify critical variables and provide interpretability on mobility patterns through a hierarchically probabilistic structure. The hierarchical structure with LC variables provides an alternative to interpret nonlinear relationships between the vital variables and the dependent variable. Detailed formulations will be introduced in Sections 3.2 and 3.3.

3.2. Bayesian supervised learning based on tensor factorization

The BSTF model was developed based on the integration between supervised learning and higher-order singular value decomposition (HOSVD). HOSVD was first proposed by Tucker (1966) as a three-way data decomposition that fits the relation between a three-way target tensor $T = \{t_{x_1,x_2,x_3}\}$ and variables X_1, X_2, and X_3 through a core tensor G (i.e., Fig. 1).

The model was extended to high-order by de Lathauwer et al. (2000), such that, with p predictor variables $X_1, \ldots, X_p$, the target tensor $T = \{t_{x_1,\ldots,x_p}\}$ is generated by all the predictor variables as follows:

$$t_{x_1,\ldots,x_p} = \sum_{z_1=0}^{k_1-1} \cdots \sum_{z_p=0}^{k_p-1} g_{z_1,\ldots,z_p} \prod_{j=1}^{p} u^{j}_{z_j x_j} \tag{4}$$

where $k_j, j = 1, \ldots, p$ denotes the rank of predictor variable X_j, $G = \{g_{z_1,\ldots,z_p}\}$ denotes the core tensor, $U^j = \{u^j_{z_j x_j}\}$ denotes the factor matrix of variable X_j. Since $k_j \leqslant |X_j|$, the HOSVD model incorporates a natural LC structure. We consider a categorical setting, such that variable X_j can take d_j states (i.e., $|X_j| = d_j$ and the states are indexed by $0, \ldots, d_j - 1$). Thereby, k_j can represent the number of LCs of variable X_j. We use an LC variable Z_j, whose value is denoted as z_j, to refer to the LC of X_j. The mapping from category x_j to LC z_j is characterized by the factor matrix. Hereafter, predictor variables $X_1, \ldots, X_p$ are also referred as manifest variables to distinguish from LC variables $Z_1, \ldots, Z_p$. The introduction of LC provides a lower rank approximation of T, and it can lead to high accuracy (Vannieuwenhoven et al., 2012). Moreover, the lower rank approximation can reduce the number of parameters for model estimation. This feature brings convenience for high-dimensional probabilistic factorization (Sun and Axhausen, 2016; Yang and Dunson, 2016).

A nonnegative version of HOSVD was proposed, which offered insights to formulate an unsupervised learning probabilistic TF model (Kim and Choi, 2007). Following the same structure as Eq. (4), a probabilistic TF is formulated as follows:

$$P(Z_1 = z_1, \ldots, Z_p = z_p) = \beta_{z_1,\ldots,z_p} \tag{5}$$

$$P(X_j = x_j | Z_j = z_j) = \kappa^j_{z_j}(x_j) \tag{6}$$

$$P(X_1 = x_1, \ldots, X_p = x_p) = \sum_{z_1=0}^{k_1-1} \cdots \sum_{z_p=0}^{k_p-1} \beta_{z_1,\ldots,z_p} \prod_{j=1}^{p} \kappa^j_{z_j}(x_j) \tag{7}$$

$$\kappa^j_{z_j}(x_j) \geqslant 0, \quad \beta_{z_1,\ldots,z_p} \geqslant 0, \quad \sum\nolimits_{z_j=0}^{k_j-1} \kappa^j_{z_j}(x_j) = 1 \tag{8}$$

where $B = \{\beta_{z_1,\ldots,z_p}\}$ denotes the core probability tensor; $K^j = \{\kappa^j_{z_j}(x_j)\}$ denotes the probability factor matrix of manifest variable X_j. The core probability tensor captures the interaction between LC variables, and a factor matrix is regarded as a probability matrix for mapping a manifest variable to its LC. The target tensor (see Eq. (7)) acts as the probability that an observed $X_i = x_1, \ldots, X_p = x_p$ is characterized by the combination of the corresponding LCs $Z_i = z_1, \ldots, Z_p = z_p$. As an unsupervised learning model, the number of LCs for each manifest variable is predetermined, and the underlying data stream from X_i to Z_i then to X_i is not capable of solving classification problems (Sun and Axhausen, 2016; Sun et al., 2018).

Inspired by Yang and Dunson (2016), the proposed BSTF model aims to integrate Bayesian supervised learning into probabilistic factorization to deal with high-dimensional classification problems. Suppose that dependent variable Y has d labels indexed by $\{0, \ldots, d-1\}$, $|\Upsilon| = d$, the BSTF model is formulated in a conditional probability context as follows:

$$P(Y = y | Z_1 = z_1, \ldots, Z_p = z_p) = \lambda_{z_1,\ldots,z_p}(y) \tag{9}$$

$$P(Z_j = z_j | X_j = x_j) = \pi^j_{z_j}(x_j) \tag{10}$$

$$P(Y = y | X_1 = x_1, \ldots, X_p = x_p) = \sum_{z_1=0}^{k_1-1} \cdots \sum_{z_p=0}^{k_p-1} \lambda_{z_1,\ldots,z_p}(y) \prod_{j=1}^{p} \pi^j_{z_j}(x_j) \tag{11}$$

$$\pi^j_{z_j}(x_j) \geqslant 0, \quad \lambda_{z_1,\ldots,z_p}(y) \geqslant 0, \quad \sum\nolimits_{z_j=1}^{k_j} \pi^j_{z_j}(x_j) = 1, \quad \sum\nolimits_{y=1}^{d} \lambda_{z_1,\ldots,z_p}(y) = 1 \tag{12}$$

where $\Lambda(y) = \{\lambda_{z_1,\ldots,z_p}(y)\}$ denotes the core conditional probability tensor for the dependent category to be y; $\Pi^j = \{\pi^j_{z_j}(x_j)\}$ denotes

the conditional probability factor matrix for manifest variable X_j. For simplification, we use $\mathbf{\Pi}$ to denote matrix $\{\Pi^j\}$, $\mathbf{\Lambda}$ to denote tensor$\{\Lambda(y)\}$ and K to denote vector $\{k_j\}$. Variables $X_1, ..., X_p$ are mutually independent from each other[1]; and the category of variable Y only depends on LC variables $Z_1,...,Z_p$. The BSTF model is suitable for the high-dimensional classification problem defined in Section 3.1. Substituting Eq. (11) into Eq. (2), we obtain the classification rule for the BSTF model as below.

$$f: \underset{y\in Y}{argmax} \sum_{z_1=0}^{k_1-1} ... \sum_{z_p=0}^{k_p-1} \lambda_{z_1,...,z_p}(y) \prod_{j=1}^{p} \pi_{z_j}^{j}(x_j) \tag{13}$$

For categorical variables, the BSTF model shown in Eqs. (9)-(12) is equivalent to the following conditional probabilistic setting:

$$Y|Z_1 = z_1, ..., Z_p = z_p \sim Cat\left(\lambda_{z_1,...,z_p}(0), ..., \lambda_{z_1,...,z_p}(d-1)\right) \tag{14}$$

$$Z_j|X_j = x_j \sim \mathrm{Cat}\left(\pi^j{}_0(x_j), ..., \pi^j_{k_j-1}(x_j)\right) \tag{15}$$

where $\mathrm{Cat}(u_1, ..., u_S)$ denotes a Categorical distribution with S states and the probability of observing state s is u_s. The categorical distribution is a Multinomial distribution with only one trial. The conditional probabilistic formulation in Eqs. (14)-(15) provide a convenient way for Bayesian model estimation. With some specific assumptions on the prior distributions of $\mathbf{\Pi}$ and $\mathbf{\Lambda}$, the observed data samples and Bayesian conditional probability theorem are adopted to estimate their posterior distribution. We utilize the Dirichlet distribution, which is the conjugate prior for the categorical distribution, as the prior distribution of $\mathbf{\Pi}$ and $\mathbf{\Lambda}$, such that

$$\lambda_{z_1,...,z_p}(0), ..., \lambda_{z_1,...,z_p}(d-1) \sim \mathrm{Dir}\left(\frac{1}{d}, ..., \frac{1}{d}\right) \tag{16}$$

$$\pi^j{}_0(x_j), ..., \pi^j_{k_j-1}(x_j) \sim \mathrm{Dir}\left(\frac{1}{k_j}, ..., \frac{1}{k_j}\right) \tag{17}$$

where $\mathrm{Dir}(v_1, ..., v_S)$ denotes a Dirichlet distribution with parameters $v_1, ..., v_S$. The adoption of Dirichlet distribution is not merely motivated by the convenience in the model estimation, but also by its capability in modeling travelers' choices among alternatives that are independent except for a constraint (Zhu et al., 2019b).

There are two types of unknown parameters that make the model estimation a complex problem: (a) model structural vector K; and (b) conditional probabilistic matrices $\mathbf{\Pi}$ and $\mathbf{\Lambda}$. We regard the estimation of the former parameters as structure learning and the latter as parameter learning. The structure vector K determines the model complexity. That is, a higher k_j leads to more probabilistic parameters to estimate. If some k_j equal 1, the corresponding manifest variable can be excluded from the model to simplify the parameter learning process. Once K is determined, $\mathbf{\Pi}$ and $\mathbf{\Lambda}$ can be estimated via existing statistical approaches, such as MCMC sampling, expectation–maximization, etc.

3.3. Structure learning of the BSTF model

With the aforementioned probabilistic formulation and assumption on the prior probability distribution, it is difficult to derive the likelihood function of a structure K directly. Thereby, the goodness of a structure K is evaluated based on marginal likelihood (ML), which is generally used for a Bayesian statistical model with the following setting:

$$W|\theta \sim p_\theta, \quad \theta|\beta \sim p_\beta \tag{18}$$

In Eq. (18), the PDF of random variable (or vector) W is denoted as p_θ with parameter θ, and θ has a prior PDF p_β and parameter β. The ML (for data sample i) is defined to marginalize out the uninterested parameter θ as follows:

$$ML(\beta|w_i) = \int_\theta p_\theta(w_i) p_\beta(\theta) \mathrm{d}\theta \tag{19}$$

where w_i is an observed data sample. In the BSTF model, the ML given K can be formulated as:

$$ML(K|\boldsymbol{x}_i, y_i) = \int_{\mathbf{\Lambda}} \int_{\mathbf{\Pi}} \sum_{\boldsymbol{z}_i} P(y_i|\mathbf{\Lambda}, \boldsymbol{z}_i) P(\boldsymbol{z}_i|\boldsymbol{x}_i, \mathbf{\Pi}) p_{\psi(K)}(\mathbf{\Lambda}, \mathbf{\Pi}) \mathrm{d}\mathbf{\Pi} \mathrm{d}\mathbf{\Lambda} \tag{20}$$

where $\boldsymbol{x}_i$ denotes vector $\{x_{i,1}, ..., x_{i,p}\}$, $\boldsymbol{z}_i$ denotes vector $\{z_{i,1}, ..., z_{i,p}\}$, and $\psi(K)$ denotes the parameters for the prior distributions in Eqs. (16)-(17).

The calculation of Eq. (20) is complicated due to multi-layer integrals on Dirichlet distributions. A simplified approximation was proposed by Yang and Dunson (2016), which transfers the soft clustering from X_j to Z_j to a hard clustering. In other words, $\pi^j_{z_j}(x_j)$ can only take 0 or 1 in the approximation, eliminating the Dirichlet priors for $\mathbf{\Pi}$. To obtain $\mathbf{\Pi}$, we need to specify the hard cluster rule R_K.

[1] This assumption will be released via a variables grouping process in Section 3.3

Initialize K, R_K, $AML(K, R_K)$
for *itr* from *1* to M_1
  for j from *1* to p
    # obtain new cluster
    if k_j equals to *1*
      randomly cluster X_j with *2* LCs
    else if k_j equals to d_j
      randomly combine two LCs
    else
      randomly choice from {*1*, *2*, *3*}
      case *1*
        randomly cluster X_j with k_j+*1* LCs
      case *2*
        randomly combine two LCs
      case *3*
        randomly re-cluster X_j with k_j LCs
    obtain K^1, $R_K{}^1$ from the **if-else** process
    # compare the new ML with the current best
    $AML(K^1, R_K{}^1)$ vs $AML(K, R_K)$, **if** accept
      update K, R_K with K^1, $R_K{}^1$
Return K, R_K, $AML(K, R_K)$

Fig. 2. Structure learning stage-one: determining K.

Initialize **Gro**, KG, R_{KG}
for *itrg* from *1* to M_G
  # group significant manifest variables
  if |**Gro**| equals to $|\mathbf{J}_K|$
    randomly combine two groups in **Gro**
  else if |**Gro**| equals to 1
    randomly split the manifest variables into two groups
  else
    randomly combine or split groups in **Gro**
  obtain new **Gro**^{1}, KG^1, $R_{KG}{}^1$ from the **if-else** process
  # obtain new structure and cluster rule for the new group
  for *itrc* from *1* to M_C
    for g in **Gro**^{1} and $|g|$ larger than *1*
      randomly increase or decrease k_g
    obtain KG^2, $R_{KG}{}^2$ from the **for** process
    # compare the new AML with the local best under **Gro**^{1}
    $AML(KG^2, R_{KG}{}^2)$ vs $AML(KG^1, R_{KG}{}^1)$, **if** accept
      update KG^1, $R_{KG}{}^1$ with KG^2, $R_{KG}{}^2$
  # compare the new AML with the current global best
  $AML(KG^1, R_{KG}{}^1)$ vs $AML(KG, R_{KG})$, **if** accept
    update **Gro**, KG, R_{KG} *with* **Gro**^{1}, KG^1, $R_{KG}{}^1$
 # compare the grouping with the original structure of stage-one
$AML(KG, R_{KG})$ vs $AML(K, R_K)$, **if** accept
  update K based on KG
  group X based on **Gro**
Return K

Fig. 3. Structure learning stage-two: manifest variable grouping.

For instance, a manifest variable $X_j \in \{0, 1, 2, 3\}$, the LC variable $Z_j \in \{0, 1\}$, a cluster rule can be $f_{X_j \to Z_j} : \{\{0, 2, 3\} \to 0, \{1\} \to 1\}$; namely, categories 0, 2, 3 for X_j are clustered as LC 0 for Z_j, and state 1 is clustered as LC 1. Consequently, the approximated marginal likelihood (AML) becomes:

$$AML(K, R_K|\boldsymbol{x}_i, y_i) = \int_{\Lambda} \sum_{z_i} P(y_i|\boldsymbol{\Lambda}, \boldsymbol{z}_i) P(\boldsymbol{z}_i|\boldsymbol{x}_i, R_K) p_{\psi(K)}(\boldsymbol{\Lambda}) \mathrm{d}\boldsymbol{\Lambda} \tag{21}$$

Since $P(\boldsymbol{z}_i|\boldsymbol{x}_i, R_K)$ takes either 0 or 1, AML reduces to a product of ML of Categorical distribution with Dirichlet priors:

$$AML(K, R_K|\{\boldsymbol{w}_1, ..., \boldsymbol{w}_n\}) = \prod_i AML(K, R_K|\boldsymbol{x}_i, y_i) = \prod_{\boldsymbol{z}} \frac{\prod_{y=0}^{d-1}\Gamma\left(\frac{1}{d} + N_{ZY}(\boldsymbol{z}, y)\right)}{\Gamma(1 + N_Z(\boldsymbol{z}))\Gamma\left(\frac{1}{d}\right)^d} \tag{22}$$

$$N_{ZY}(\boldsymbol{z}, y) = \sum_{i=1}^{n} I\left(Z_{i,1} = z_1, ..., Z_{i,p} = z_p, Y_i = y\right) \tag{23}$$

$$N_Z(\boldsymbol{z}) = \sum_{i=1}^{n} I\left(Z_{i,1} = z_1, ..., Z_{i,p} = z_p\right) \tag{24}$$

where $I(\cdot)$ is the indicator function; $N_{ZY}(\cdot)$ and $N_Z(\cdot)$ are counting functions for the specified combination of LC variables $Z_1, ..., Z_p$ and dependent variable Y. Many algorithms (e.g., genetic algorithm, simulation-based optimization (Chen et al., 2015), and MCMC) can be adopted to seek the optimal structure with a high AML.

In this paper, we propose a two-stage structure learning process as follows: (a) stage-one determines an initial model structure and excludes insignificant manifest variables; and (b) stage-two attempts to group the significant manifest variables obtained from stage-one, which releases the assumption in Section 3.2 such that variables $X_1, ..., X_p$ are mutually independent of each other. Stage-two is necessary because the mutually independent assumption may ignore the correlation among some manifest variables. For instance, the conditional dependency structure $Z_{0,1}|\{X_0, X_1\}$ may be more reasonable than $Z_0|X_0$ and $Z_1|X_1$ if manifest variables X_0 and X_1 are highly correlated. Compared with the simplified one-stage algorithm provided by Yang and Dunson (2016), our two-stage algorithm extends the BSTF model to generalized classification problems with variables of different ranks (number of categories) and correlations.

The proceeding of the stage-one algorithm is depicted in Fig. 2, in which M_1 is the total number of iterations for determining the optimal K. We initialize the process by setting $k_j = 1$, and calculate the corresponding AML. Within each iteration, the algorithm visits each of the manifest variables to randomly increase or decrease k_j. After changing k_j, temporary K^1 and R_K^1 are obtained to evaluate a corresponding AML. The acceptance of the new K^1 and R_K^1 will be based on the comparison between the current AML and the temporary AML via the stochastic search MCMC (SS-MCMC) algorithm (George and McCulloch 1997).

Fig. 3 illustrates the manifest variable grouping algorithm for stage-two. In this figure, M_G denotes the number of iterations, ***Gro*** denotes the grouping rule of the manifest variables, KG denotes the model structure after grouping, R_{KG} denotes the clustering rule with KG, and M_C denotes the number of sub-iterations to determine the LC structure KG of a grouping ***Gro***. We use a sample example to present the proceeding of the algorithm. Suppose manifest variables X_0, X_3 and X_4 are found significant in the first stage (i.e., $k_0 > 1$, $k_3 > 1$, and $k_4 > 1$). To initialize, ***Gro*** will have three groups, i.e., ***Gro*** $= \{\{0\}, \{3\}, \{4\}\}$; KG and R_{KG} are obtained from stage-one such that $KG = \{k_0, k_3, k_4\}$ and $R_{KG} = R_K$. For each iteration, the algorithm randomly combines or splits the groups in the existing ***Gro*** to obtain a temporal ***Gro***^{1}. If a group g only has one manifest, i.e., $g = \{j\}$, the corresponding k_g is set to be k_j obtained from structure learning stage-one; otherwise, its k_g is set to 2. Continue with the example, after a random combination, the model structure can be ***Gro***$^1 = \{\{0, 4\}, \{3\}\}$, $KG^1 = \{2, k_3\}$. A random mapping from ***Gro***^{1} to KG^1 is adopted to generate an initial classification rule R_{KG}^1. After obtaining ***Gro***^{1}, KG^1, and R_{KG}^1, the algorithm will determine the local best KG^1 under grouping rule ***Gro***^{1}. Similar to structure learning stage-one, for each sub-iteration, the algorithm randomly increases or decreases k_g for the groups g which have more than one manifest variable; then temporal KG^2 and R_{KG}^2 are obtained. The local best KG^1 and R_{KG}^1 will be updated based on the comparison of the AMLs. After the sub-iteration loop, the temporal grouping rule ***Gro***^{1} has a local best KG^1 and R_{KG}^1. The decision to update the grouping rule is made after comparing the AML of ***Gro***^{1} and the AML of the current grouping rule ***Gro***, KG and R_{KG}. After the entire iteration loop (i.e., M_G iterations), the AML obtained from stage-one will be compared with the AML with ***Gro***, KG and R_{KG} to make a final decision on the acceptance of the grouping rule. All the acceptance decisions are based on the SS-MCMC algorithm (George and McCulloch, 1997).

3.4. Parameter learning of the BSTF model

After learning the model's structure, we adopt Gibbs sampling as a general approach in Bayesian statistics for parameter learning. Based on the prior conjugate assumption in Eqs. (16)-(17), we utilize a four-step Gibbs sampling, which completes a stationary Markov chain, to estimate the posterior probabilistic parameters of $\boldsymbol{\Pi}$ and $\boldsymbol{\Lambda}$. For initialization, a random clustering of $\boldsymbol{x}_i$ to z_i is conducted to label each manifest variable of training sample i with a starting LC. Then, we will repeat Steps 1 through 4 below for M_2 iterations such that the Markov chain can converge:

Step 1: Sample and update $\boldsymbol{\Lambda}$ based on Eq. (25).

$$\lambda_{z_1,...,z_p}(0), ..., \lambda_{z_1,...,z_p}(d-1)|- \sim \mathrm{Dir}\left(\frac{1}{d} + N_{ZY}(z_1, ..., z_p, 1), ..., \frac{1}{d} + N_{ZY}(z_1, ..., z_p, d)\right) \tag{25}$$

Step 2: Switch labels for the LCs to speed up the convergence of the algorithm.

Step 3: Sample and update $\boldsymbol{\Pi}$ based on Eqs. (26)-(27).

$$\pi^j{}_0(x_j), ..., \pi^j_{k_j-1}(x_j)|- \sim \mathrm{Dir}\left(\frac{1}{k_j} + N_{ZX_j}(0, x_j), ..., \frac{1}{k_j} + N_{ZX_j}(k_j - 1, x_j)\right) \tag{26}$$

$$N_{ZX_j}(z_j, x_j) = \sum_{i=1}^{n} I(Z_{i,j} = z_j, X_{i,j} = x_j) \tag{27}$$

Step 4: Sample and update z_i based on Eq. (28).

$$P(Z_{i,j} = z_j|-) \propto \lambda_{z_{i,1},\ldots,z_{i,j-1},z_j,z_{i,j+1},\ldots,z_{i,p}}(y_i)\pi^j_{z_j}(x_{i,j}) \tag{28}$$

In the equations above, symbol $|-$ means the probabilistic relationship is dependent on the entire dataset, i.e., $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ in the observed manifest dataset and the clustered LC states $z_1, \ldots, z_n$.

4. A case study based on Hangzhou DiDi and land use data

Urban mobility has undergone drastic changes in recent years with the introduction of on-demand ride services (or ride-sourcing services) provided by transportation network companies (TNCs), such as Uber, Lyft, and Didi Chuxing. By efficient connecting passengers and dedicated drivers through an online platform (or smartphone app), ride-sourcing services are now playing an increasingly important role in meeting mobility needs[2]. Recently, TNCs are launching and promoting dynamic ridesharing services (also termed as ridesplitting services)[3]. Unlike traditional ridesharing programs, which require users to schedule their trips in advance, dynamic ridesharing programs can accommodate on-demand requests and bring advantages in many aspects, including improving the vehicle utilization rate, reducing passengers' cost, increasing drivers' income, alleviating traffic congestion, saving energy, and mitigating air pollution (Ferguson, 1997; Chan and Shaheen, 2012; Chen et al., 2018)[4].

To examine the performance of the proposed BSTF model and understand the conditional probabilistic relationship between ridesharing choice and the key predictor variables (i.e., manifest variables) through a hierarchical LC structure, we undertake a real-world case study in this section. In Section 4.1, we introduce the dataset used in this case study. The performance of the BSTF model in terms of classification accuracy is shown in Section 4.2, in which we also compare different state-of-the-art machine learning approaches (e.g., NN and RF). The identified key manifest variables, the LC variables, and the conditional probabilistic relationship are discussed in Section 4.3.

4.1. Ridesharing and land-use data

The dataset used in this case study is fused based on DiDi order data and land use GIS data of Hangzhou, China. Didi Chuxing has been operating dynamic ridesharing programs in many cities in China, such as Beijing, Chengdu, and Hangzhou, for a few years. DiDi Hitch and DiDi Express Carpool were launched in July 2015 and November 2015, respectively; the former is dynamic ridesharing between a driver and a passenger with similar routes, while the latter is ridesharing between two passengers provided by a dedicated driver. Detailed information of individual ride orders is recorded during operations, including order ID, driver ID, passenger ID, passenger order time, service type (i.e., ridesharing, non-sharing or taxi), location (latitude and longitude) of the origin and destination, pickup and drop-off time, trip distance, car level, etc. The order data used in this case study, which includes 251,344 records in total, were randomly sampled at a rate of 20% from all the completed order data between September 7 and 13, 2015, in Hangzhou, China. Since DiDi Express Carpool was launched after collecting this dataset, in this case study, ridesharing refers to DiDi Hitch, and non-sharing refers to DiDi Express. The land-use data were provided by Hangzhou Transportation Research Center, Hangzhou, China, which records traffic analysis zone (TAZ) based socio-economic data, including population, number of houses, number of business of public units, number of bus and subway stations, areas of various land-use types (e.g., residential, business, education, industry, and hospital), multiple types of points of interest (POIs), etc. We fuse the order data and land use data by adding TAZ based socio-economic variables into individual orders. Namely, based on the location information in each order, we identify the TAZ IDs of the origin and destination and fuse the corresponding socio-economic variables. After cleaning order data with missing and/or extreme values, we finally obtain 212,310 samples for the pattern analysis of dynamic ridesharing.

The variables of the fused dataset are shown in Table 1, in which continuous variables are discretized into category variables. We present pie charts for all the variables in Fig. 4. It is noted that the majority of the orders are non-sharing (DiDi Express), while ridesharing and taxi only take up 8% and 14%, respectively. Most socio-economic (i.e., land-use) variables show closed distribution patterns at the origin and the destination except for variables "O_bus" and "D_bus". Note that some of the trip features (e.g., order waiting time and in-vehicle time) are known after the trip. Incorporating these variables in the model is beneficial for TNCs to better understand the performance of ridesharing programs in both the demand and supply sides (Chen et al., 2017). Based on the probabilistic relationship between the dependent variable and the post-trip features, the TNC can design operational strategies to address the demand–supply imbalance and improve the service quality of different service options. For instance, the TNC may display an estimated

[2] It is reported that Uber has developed its business to 24 countries and served over five billion trips since its birth (Uber's official website. https://www.uber.com/newsroom/5billion-2/); DiDi is serving over 25 million trips every day which covers over 400 cities in China (China Daily. http://global.chinadaily.com.cn/a/201801/09/WS5a541c98a31008cf16da5e76.html).

[3] It is reported that Lyft aimed to have 50 percent of rides being shared by 2022 (Schaller, 2018).

[4] Ridesharing has a long history tracing back to the Second World War when the U.S. government established the Car-Sharing Club for fuel conservation. Traditional ride-sharing generally involves commonly commuters who plan to share a ride with others; the riders are willing to share travel costs such as gasoline consumption.

Table 1
Variables considered in the case study.

variable name	Variable meaning	Category code
type	Order service type	0: taxi; 1: non-sharing; 2: ridesharing
create_DoW	Order creating day of week	0: Mon; 1: Tue, Wed, Thu; 2: Fri; 3: Sat, Sun
create_time	Order creating time	0: 0 to 5o'clock, and 19 to 23o'clock 1: 6 to 8o'clock, and 16 to 18o'clock 2: 9 to 15o'clock
waiting_time	Order waiting time (matching time plus pickup time)	0: smaller or equal to 3 min; 1: 3 to 5 min; 2: 5 to 7 min; 3: 7 to 10 min; 4: 10 to 15 min; 5: over 15 min
vehicle_time	In-vehicle travel time	0: smaller or equal to 5 min; 1: 5 to 10 min; 2: 10 to 15 min; 3: 15 to 20 min; 4: 20 to 30 min; 5: over 30 min
distance	Actual recorded travel distance	0: smaller or equal to 3 km; 1: 3 to 6 km; 2: 6 to 10 km; 3: 10 to 20 km; 4 over 20 km
O_house (D_house)	Number of houses in the origin (destination) zone	0: smaller or equal to 200; 1: 200 to 600; 2: 600 to 1100; 3: 1100 to 1600; 4: 1600 to 2400; 5: over 2400
O_bus (D_bus)	Number of bus stops in the origin (destination) zone	0: smaller or equal to 20; 1: 20 to 50; 2: 50 to 70; 3: 70 to 120; 4: over 120
O_metro (D_metro)	Number of metro stations in the origin (destination) zone	0: none; 1: 1 or 2; 2: 3 or more
O_beauty (D_beauty)	Number of beauty shops in the origin (destination) zone	0: smaller or equal to 10; 1: 10 to 50; 2: 50 to 100; 3: 100 to 200; 4: over 200
O_restaurant (D_restaurant)	Number of restaurants in the origin (destination) zone	0: smaller or equal to 50; 1: 50 to 200; 2: 200 to 400; 3: 400 to 700; 4: over 700
O_education (D_education)	Number of schools in the origin (destination) zone	0: smaller or equal to 3; 1: 3 to 10; 2: 10 to 20; 3: 20 to 40; 4: over 40
O_enterprise (D_enterprise)	Number of enterprises in the origin (destination) zone	0: smaller or equal to 200; 1: 200 to 400; 2: 400 to 900; 3: 900 to 1500; 4: over 1500
O_hospital (D_hospital)	Number of hospitals in the origin (destination) zone	0: smaller or equal to 10; 1: 10 to 30; 2: 30 to 70; 3: 70 to 120; 4: over 120
O_shop (D_shop)	Number of shops in the origin (destination) zone	0: smaller or equal to 200; 1: 200 to 400; 2: 400 to 600; 3: 600 to 1000; 4: over 1000

waiting time or in-vehicle time to passengers through the smartphone app to encourage them to opt for ridesharing or other social preferable options (Li et al., 2018). The data contains multiple trip records from one driver or passenger such that the analysis may be biased from the individual perspective. Due to privacy-preserving issues, we anonymize the personally identifiable information of drivers and passengers[5]. The individual-level behavioral analysis is not suitable due to the aforementioned biased sample issue and the lack of personal demographic attributes. With land-use attributes of each trip, this paper focuses on the aggregate-level classification and mobility pattern analysis of the three ride-sourcing service options provided on a ride-sourcing platform. There are some other limitations of the fused dataset. For instance, there is no estimated pre-trip travel information about service types that a traveler did not choose; the order type can be affected by DiDi's dispatching algorithm and priority, the analysis results may be different given data from other TNCs[6]; and the missing of pricing information makes it hard to analyze the sensitivity of the trip fare to service type choice.

4.2. Results on classification accuracy

We train the proposed BSTF model based on the fused dataset with variable "type" as the dependent variable (service type choice) and the other variables as manifest variables. In addition to the proposed BSTF model, we also train other machine learning models in the literature for comparison. All the models are briefly introduced below:

- BSTF: the Bayesian supervised learning tensor factorization model presented in Section 3.
- DT: conditional inference DT model, an implementation of conditional inference trees that embed tree-structured regression models into a well-defined theory of conditional inference procedures (Zeileis et al., 2008). It dominates structured datasets on classification and regression predictive modeling problems.
- NB: naïve Bayesian model, a simple probabilistic classifier that applies Bayes conditional probability theory with strong independence assumptions among variables.

[5] Among the final 212,310 cleaned samples, there are 40,292 drivers and 54,776 passengers. The average number of orders made by a driver/passenger is 5.3/3.8, respectively. 29.7%/33.1% of the drivers/passengers took only 1 order, 39.4%/46.1% of them took 2 to 5 orders, 27.6%/19.7% took 6 to 20 orders, and 3.3%/1.1% took over 20 orders.

[6] In the on-demand ride service market of 2015, DiDi was the only ride-sourcing platform providing the taxi dispatching service. DiDi still dominates the e-hailing service for taxis now.

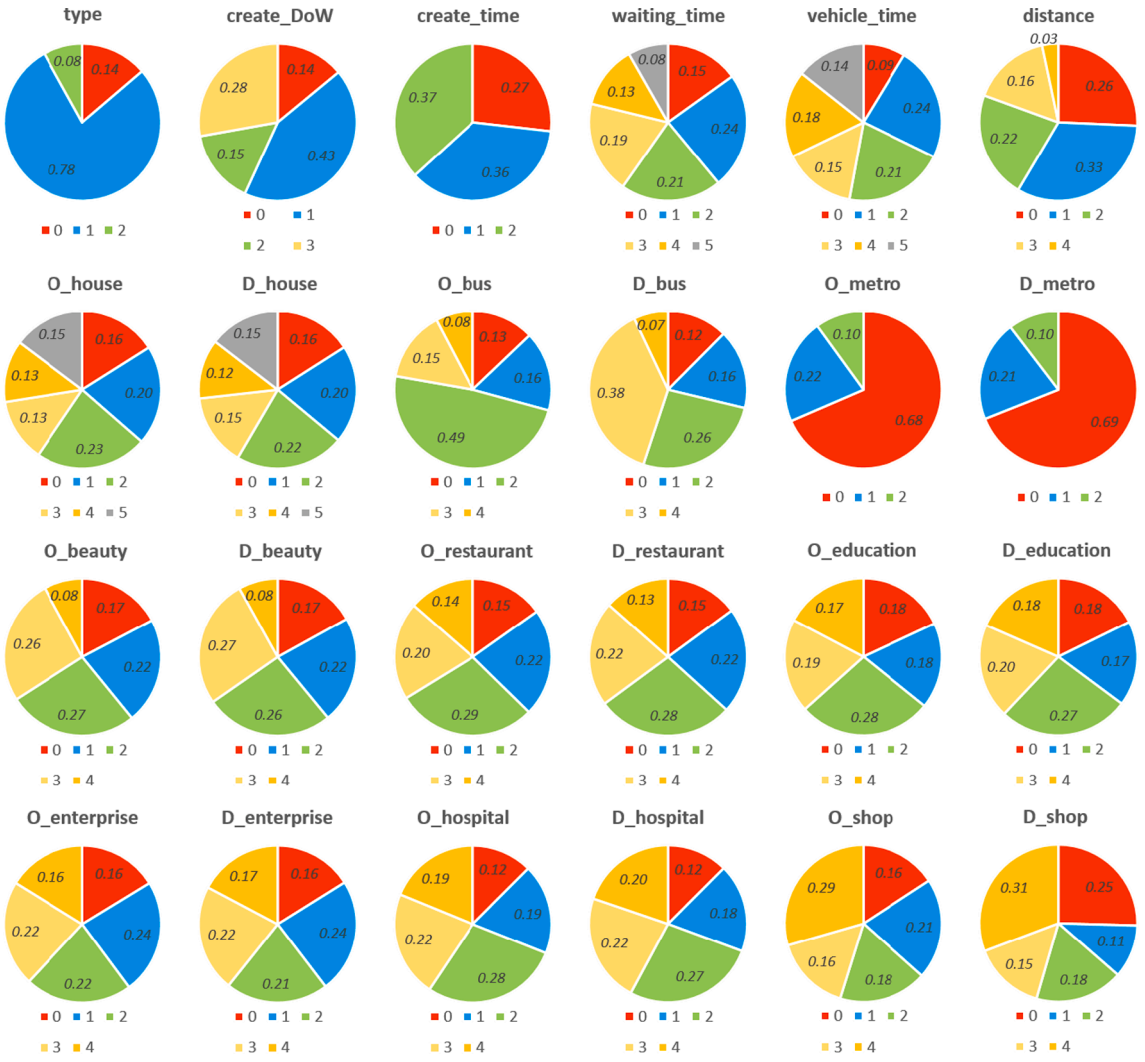

Fig. 4. Pie charts for all the variables.

- NN: the basic neural network contains at least three layers, i.e., input layer, hidden layer, and output layer. There are three hidden layers in this study, and the numbers of neurons are 3, 5, and 3 in each hidden layer, respectively.
- RF: random forest model, is the model trained by bootstrapped samples of each decision tree. The number of trees is 100 in this paper.
- SVM: support vector machine model, which constructs a set of hyperplanes in a domain with infinite-dimensional variables to classify them into categories. It can be used as a classification machine, as a regression machine, or for novelty detection. We use a linear kernel in this paper.

Table 2
Models for comparison in the case study.

Model	Algorithm	Training Time (s)			
		10%	20%	50%	80%
BSTF	MCMC and Gibbs sampling	18.2	33.9	99.2	151.5
DT	Zeileis et al. (2008)	1.1	2.7	9.8	21.6
NB	Dimitriadou et al. (2009)	0.3	0.8	1.9	3.2
NN	Ripley et al. (2016)	7.1	15.7	39.4	70.9
RF	Breiman (2001)	1.4	3.1	11.7	25.5
SVM	Chang and Lin (2011)	15.2	28.4	76.3	126.8

Table 3
Classification error of the case study.

Model	Percentage of training set			
	10%	20%	50%	80%
BSTF	**0.186**	**0.183**	**0.181**	**0.180**
DT	0.192	0.190	0.184	0.183
NB	0.278	0.260	0.251	0.247
NN	0.222	0.208	0.197	0.189
RF	0.189	0.187	0.185	0.183
SVM	0.196	0.196	0.192	0.192

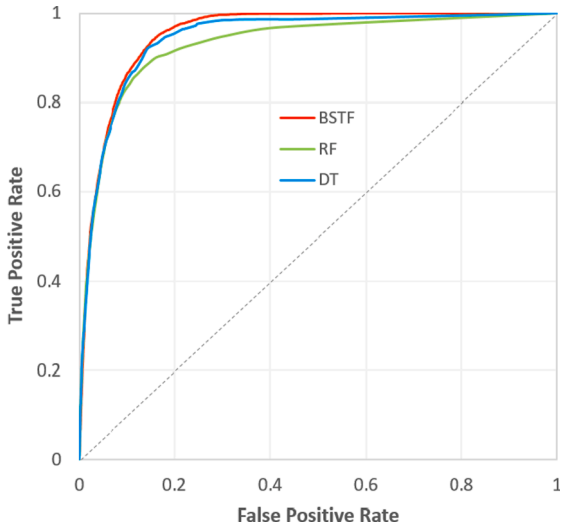

Fig. 5. ROC curves under 80% training size.

These models are trained and tested via existing algorithms, and the proposed BSTF model is estimated by the proposed MCMC-based two-stage structure learning and Gibbs sampling-based parameter learning (shown in Table 2). We train these machine learning models in four cases with different training sizes (i.e., 10%, 20%, 50%, and 80%). Taking the case with 10% training size as an instance, the dataset is randomly split as a training set with 21,231 samples (10% of 212,310) and a test set with the remaining 191,079 samples (90% of 212,310); the training set is used for model estimation, and the test set is for the classification accuracy test. Table 2 also presents the computational time of these models with a 6th-Gen Intel Core i5 Cup and a 6 GB RAM. Since the Bayesian estimation algorithm requires MCMC and Gibbs sampling, it generally takes a long time for model training. Given the complex hierarchical structure, it is necessary to apply the proposed algorithm to identify all the key variables.

The classification errors for all the cases and models are shown in Table 3. For all four cases, the proposed BSTF models perform the lowest classification error. This result agrees with the claim that the tensor formulation of high-order interaction can maintain predictive information in the supervised learning approach (Vannieuwenhoven et al., 2012). Moreover, the superiority of the BSTF models becomes more significant with smaller training datasets. It is general to note that the classification accuracy for all the models will improve with a larger training size. The classification error of the RF model is closed to the BSTF model in all four cases, which means RF is also a preferable approach for high-dimensional classification. The NB, NN, and DT models are significantly sensitive to the training size; under the 80% training size scenario, the DT model can also provide high classification accuracy.

To better illustrate the classification performance for ridesharing, we present the receiver operating characteristic (ROC) curve in Fig. 5. The ROC curve is the general measurement for the diagnostic ability of a classification model; and the higher the curve above the diagonal line (dashed line in Fig. 5), the more trustable the classifier is. We plot the ROC curves with "ridesharing" as a positive classification for the BSTF model, the DT model, and the RF model under the case of 80% training size. We select these three models because they are superior in overall classification accuracy compared to the other models. We note that the BSTF model provides the most reliable classification for ridesharing since its ROC curve is the highest. The DT model and the RF model also show good performance. The areas under the curve (AUC) for the BSTF model, the DT model, and the RF model are 0.952, 0.944, and 0.923, respectively. We also provide detailed classification errors for different service types in Appendix A.

4.3. Latent class analysis of ridesharing choice patterns

The tensor formulation of high-order interaction not only provides a high classification accuracy, as shown in Section 4.2, but also

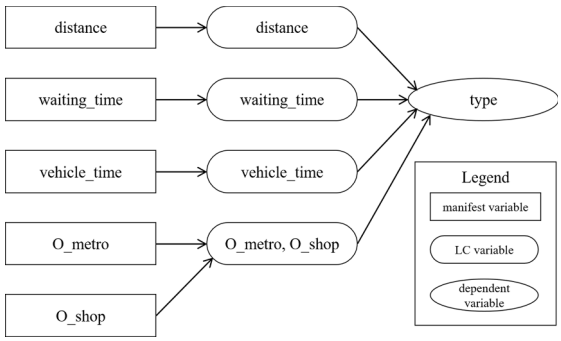

Fig. 6. A Bayesian Network presentation of the BSTF model.

Table 4
Probability table **Π**.

(a) variable "distance"

	LC 0	LC 1	LC 2
0	0.946	0.042	0.012
1	0.110	0.863	0.026
2	0.002	0.439	0.559
3	0.015	0.014	0.971
4	0.009	0.018	0.973

(b) variable "waiting_time"

	LC 0	LC 1	LC 2	LC 3
0	0.998	0.001	0.001	0.001
1	0.740	0.217	0.000	0.042
2	0.463	0.442	0.000	0.095
3	0.267	0.355	0.270	0.108
4	0.157	0.077	0.554	0.211
5	0.014	0.176	0.125	0.685

(c) variable "vehicle _time"

	LC 0	LC 1	LC 2	LC 3
0	0.985	0.013	0.001	0.001
1	0.007	0.982	0.002	0.009
2	0.001	0.167	0.673	0.159
3	0.005	0.019	0.672	0.304
4	0.003	0.008	0.477	0.512
5	0.004	0.002	0.147	0.848

(d) variable "O_metro, O_shop"

	LC 0	LC 1
0,0	0.995	0.005
0,1	0.920	0.080
0,2	0.867	0.133
0,3	0.991	0.009
0,4	0.996	0.004
1,0	0.485	0.515
1,1	0.646	0.354
1,2	0.977	0.023
1,3	0.152	0.848
1,4	0.862	0.138
2,0	0.237	0.763
2,1	0.858	0.142
2,2	0.924	0.076
2,3	0.054	0.946
2,4	0.968	0.032

offers probabilistic interpretability among the manifest variables, LC variables, and dependent variable (service type choice). In this section, the BSTF model in the 80% training size case is used to analyze passengers' choice patterns. The probabilistic relationship among the key variables is depicted in Fig. 6. Similar to the Bayesian network (Zhu et al., 2016, 2018), an arrow from variable *A* to

Table 5
Marginal probability $P(y|Z_j = z_j)$.

Manifest Variable	LC	$y = 0$	$y = 1$	$y = 2$
distance	0	0.344	0.537	0.119
	1	0.351	0.488	0.161
	2	0.374	0.419	0.207
waiting_time	0	0.607	0.389	0.003
	1	0.334	0.587	0.079
	2	0.262	0.523	0.215
	3	0.222	0.427	0.351
vehicle_time	0	0.618	0.241	0.140
	1	0.298	0.560	0.142
	2	0.189	0.671	0.139
	3	0.320	0.454	0.226
O_metro, O_shop	0	0.244	0.584	0.172
	1	0.469	0.379	0.152

variable B means that the probability distribution of B depends on A. We note that stage one of structure learning identifies five key manifest variables (i.e., "distance", "waiting_time", "vehicle_time", "O_metro" and "O_shop") that have a significant probabilistic relationship with the service type. Based on stage two of the structure learning algorithm, variables "O_metro" and "O_shop" are grouped and have one LC variable "O_metro, O_shop". Variables "distance", "O_metro", and "O_shop" are known by passengers (also by the TNC) before the trips. The probabilistic relationship between these pre-trip variables and the order type can reflect passengers' preferences on ride-sourcing service types given their travel needs. Variables "waiting_time" and "vehicle_time" are post-trip variables correlated with the service type. Since "waiting_time" consists of both matching time and pickup time, it may reflect DiDi's order searching time window and searching radius; and it is not a pure post-trip variable because some passengers may make an early schedule that leads to an idle time window and a long pickup time. The variable "vehicle_time" is positively correlated to "distance", but it also reflects congestion during the trip. As the key factors of the classification problem, the two post-trip variables are strongly correlated with the order type, such that each service type can have a distinctive pattern of service quality. Note that the conditional dependency (arrow) from an LC variable to the order type does not indicate a causal relationship. For instance, "waiting_time" is caused by the order type, not vice versa.

As mentioned in Section 3, the probabilistic relationship is presented by the probability tables between the LC variables and their manifest variables ($\mathbf{\Pi}$) and the probability tables between the dependent variable "type" and the LC variables ($\mathbf{\Lambda}$). We summarize probability table $\mathbf{\Pi}$ in Table 4, which is estimated based on Eq. (26) after parameter learning. The decimal numbers in Table 4 denote $\pi_z^j(x_j)$ with x_j to be the manifest category (row index) and z_j to be the LC (column index). Variable "distance" is found to have three LCs. Since the index of the category for variable "distance" increases with respect to travel distance (Table 1), we may explain LC 0 as a short distance, LC 1 as a medium distance, and LC 2 as a long distance. Similar to manifest variable "distance", LC 0, 1, 2, and 3 for manifest variable "waiting_time" can be interpreted as short wait, medium wait, long wait, and extremely long wait, respectively; and LC 0, 1, 2, and 3 for manifest variable "vehicle_time" can be interpreted as short ride, medium ride, long ride, and extremely long ride, respectively. Two LCs are found for the grouped manifest variable "O_metro, O_shop". It is found that for origins without a metro station, the probability of LC 1 is very low. However, with only the manifest variables, the meaning of LCs 0 and 1 is not clear; and further examination is provided later in this section.

To interpret the probabilistic relationship between each LC variable and the dependent variable, we illustrate the marginal probability $P(y|Z_j = z)$, which is calculated as flows:

$$P(y|Z_j = z_j) = \sum_{Z_j = z_j} \lambda_{z_1,\ldots,z_p}(y) \tag{29}$$

Here the values of $\lambda_{z_1,\ldots,z_p}(y)$ are obtained via Eq. (25). Based on the marginal probabilities shown in Table 5, the classification

Table 6
LC Patterns for high ridesharing classification probability.

$Z_1 = z_1, \ldots, Z_p = z_p$				$\lambda_{z_1,\ldots,z_p}(y)$		
distance	waiting_time	vehicle_time	O_metro, O_shop	$y = 0$	$y = 1$	$y = 2$
2	3	1	0	0.067	0.168	0.765
2	2	1	0	0.292	0.019	0.689
1	3	0	0	0.411	0.015	0.575
2	3	3	0	0.023	0.409	0.568
0	3	3	0	0.128	0.318	0.554
2	3	2	0	0.055	0.409	0.536
1	3	2	1	0.071	0.406	0.523
2	3	3	1	0.067	0.432	0.501

probability of ridesharing is found to monotonically increase with respect to the LCs for variables "distance" and "waiting_time". A longer trip distance makes passengers inclined to ridesharing, probably due to the larger monetary saving. The positive relationship between the LC of "waiting_time" and the probability of ridesharing may be because there is a shortage of ridesharing (i.e., DiDi Hitch) drivers, or many ridesharing passengers used DiDi Hitch to schedule their trips in advance (Chen et al., 2017). As the LC of "distance" varies from 0 to 2, the classification probability of non-sharing (i.e., DiDi Express) decreases significantly, but it may not affect the probability of taking a taxi. This means taxi passengers are not sensitive to trip distance. We also note that the probability of taking a taxi monotonically decreases with the LC of "waiting_time", such that taxis are faster than DiDi's ride-sourcing services in picking up passengers. Unlikely, the LC for "vehicle_time" does not have a notable monotonic relationship with the probabilities of all the three service types. When the LC for "vehicle_time" is 0, the classification chance of taxi is 61.8%. For service types taxi and non-sharing, the monotonic patterns of their marginal probabilities with respect to LC variables "distance" and "vehicle_time" are not consistent. The results indicate that traffic congestion is significantly correlated with service types, affecting a passenger's choice among taxi and non-sharing. For the grouped variable "O_metro, O_shop", we note that LC 1 results in a 46.9% chance of a taxi service, while LC 0 reduces the taxi probability to 24.4%. Based on the marginal probabilities, we can regard LC 1 of variable "O_metro, O_shop" as a taxi-convenient class, while LC 0 as a taxi-inconvenient class. It generally makes sense because the number of cruise taxis in a TAZ is related to its socio-economic variables (i.e., "O_metro" and "O_shop" in this case study).

We provide the combinations of LCs $Z_1 = z_1, \ldots, Z_p = z_p$ that lead to high classification probability for ridesharing (i.e., high $\lambda_{z_1,\ldots,z_p}(2)$) in Table 6 to better understand passengers' choice patterns of dynamic ridesharing. There are notable common features among the illustrated combinations of LC variables that lead to a high probability of ridesharing: the LC of "distance" is long (i.e., 2), the LC of "waiting_time" is long or extremely long (i.e., 2 or 3), and the LC of "O_metro, O_shop" is "taxi-inconvenient" (i.e., 0). The general pattern is representative in real-world cases when ridesharing is usually used. That is, regardless of the in-vehicle travel time, passengers tended to use ridesharing in taxi-inconvenient areas for long-distance trips. Moreover, as discussed in the finding mentioned above in Table 5, many passengers scheduled ridesharing services in advance, leading to an extremely long waiting time (i.e., LC 3). In addition to the general LC pattern, there are also other representative LC combinations. For instance, LC combination (1, 3, 0, 0) has a high ridesharing classification probability because passengers may schedule a medium-distance ridesharing trip in a taxi-inconvenient area; LC combination (0, 3, 3, 0) indicates that in a highly congested (LC 3 for "vehicle_time") and taxi-inconvenient area, a trip is more likely a ridesharing order once the waiting time is extremely long.

More interesting findings are awaiting once additional manifest variables (e.g., gender, and pricing) are available. Moreover, as mentioned in Section 4.1, we only know post-trip travel information in the current dataset, which is common in many big data analyses, resulting in the limited predicting capability of the BSTF model and other state-of-art machine learning models. To make full use of the Hangzhou dataset, we present additional results in Appendix B, such that models with estimated "pre-trip" information for different service types and models without any post-trip/pre-trip information are tested.

5. Conclusions

This paper provides a high-dimensional machine learning classification model, i.e., the BSTF model, for solving classification problems in mobility pattern analysis. The BSTF model integrates supervised learning and probabilistic TF, which utilizes a hierarchical LC structure to interpret the probabilistic relationship between the predictor variables and the dependent variable (i.e., travel choice or other mobility measurements). The consideration of LC variables enables the BSTF model to discover hidden patterns in transportation classification problems. With high-dimensional datasets, the BSTF model is capable of identifying critical variables that are related to mobility patterns. The estimation of the BSTF model consists of a two-stage structure learning process and a parameter learning process. The former process identifies the significant predictor (manifest) variables and the corresponding LC structure via MCMC, and the latter process estimates the parameters of the conditional probability relation based on Bayesian posterior inference.

A real-world case study on dynamic ridesharing analysis with Hangzhou DiDi ride-sourcing data and land use data is conducted in this paper. Cases with different training sizes are utilized to examine the classification accuracy of the BSTF model and some other state-of-the-art machine learning approaches. The BSTF model is found superior in classification accuracy. Even with small training datasets, the BSTF model can capture important features for highly accurate classification. Also, the interpretability ability of the BSTF model is witnessed by the identification of crucial predictor variables (i.e., travel distance, waiting time, in-vehicle time, and the accessibility of metro and shops of the origin) and the corresponding LC structure in passengers' choice patterns on different ride-sourcing service types.

One of the major limitations is that this paper only considers categorical variables. Therefore, one future research direction can be the adaption for mixed variables or continuous variables. The BSTF approach can also be implemented in unsupervised learning research such as population synthesis, activity synthesis, etc.; and other supervised learning research like traffic prediction. The scalability of this model for more complex problems can also be an interesting research direction. For instance, one may extend the current static BSTF model to a dynamic model. Moreover, the dataset in this study does not include pricing information and individual-level demographic variables. With more comprehensive datasets, the model is expected to provide more behavioral insights into transportation research.

CRediT authorship contribution statement

Zheng Zhu: Conceptualization, Software, Writing - original draft, Visualization. **Lijun Sun:** Methodology, Writing - original draft.

Xiqun Chen: Data curation, Writing - review & editing, Project administration. **Hai Yang:** Supervision, Project administration.

Acknowledgments

The work described in this paper is partially supported by Hong Kong Research Grants Council under project HKUST16208920, and is partially supported by the Hong Kong University of Science and Technology - DiDi Chuxing (HKUST-DiDi) Joint Laboratory. The opinions in this paper do not necessarily reflect the official views of HKUST-DiDi Joint Laboratory. The authors are responsible for all statements. Dr. Xiqun Chen is financially supported by the National Key Research and Development Program of China (2018YFB1600900), Zhejiang Provincial Natural Science Foundation of China (LR17E080002), National Natural Science Foundation of China (71922019, 71771198, 71961137005), and Young Elite Scientists Sponsorship Program by CAST (2018QNRC001).

Appendix A. . Classification results by the service type

In this appendix, we illustrate the detailed classification percentages of different models under the 80% training size scenario in Section 4.2. In Table A.1, the numbers in the cells represent the average classification rate (i.e., percentage) over the 42,462 samples (20% of 212,310). For instance, the number 2.16 in Table A.1(a) means that 2.16% of the classifications correctly predict service type 0; while the number 11.08 means that 11.08% of the classifications predict the type to be 1, but the actual type is 0.

Based on Table A.1, we compute the precision and recall of different service types in Table A.2. The precision (also referred to as the

Table A1
Detailed classification percentages for different service types.

(a) The BSTF model

Percentage		Predicted Type		
		0	1	2
Actual Type	0	2.16	11.08	0.38
	1	0.67	75.61	2.11
	2	0.10	3.66	4.27

(b) The DT model

Percentage		Predicted Type		
		0	1	2
Actual Type	0	2.22	11.01	0.38
	1	1.07	75.06	2.25
	2	0.09	3.53	4.41

(c) The BN model

Percentage		Predicted Type		
		0	1	2
Actual Type	0	3.33	9.64	0.65
	1	5.55	66.75	6.09
	2	0.09	2.72	5.21

(d) The NN model

Percentage		Predicted Type		
		0	1	2
Actual Type	0	0.91	12.42	0.29
	1	0.41	75.87	2.10
	2	0.00	3.70	4.32

(e) The RF model

Percentage		Predicted Type		
		0	1	2
Actual Type	0	2.53	10.74	0.34
	1	1.31	74.98	2.10
	2	0.09	3.77	4.17

(f) The SVR model

Percentage		Predicted Type		
		0	1	2
Actual Type	0	1.02	12.15	0.44
	1	0.50	75.23	2.65
	2	0.01	3.47	4.55

Table A.2
Precision and recalls of different service types.

Model	Type 0		Type 1		Type 2	
	Precision	Recall	Precision	Recall	Precision	Recall
BSTF	73.72	15.86	83.69	96.45	63.17	53.18
DT	65.68	16.31	83.77	95.76	62.64	54.92
BN	37.12	24.45	84.38	85.15	43.60	64.96
NN	68.94	6.68	82.48	96.80	64.38	53.87
RF	64.38	18.59	83.79	95.65	63.09	51.93
SVR	66.67	7.49	82.81	95.98	59.55	56.66

positive predictive value) is the fraction of true positive data records among the total positive data records; and recall (also referred to as the sensitivity) is the fraction of true positive records among the total true records[7].

Appendix B. . Extra numerical results

One limitation of this paper lies in the lack of important variables such as pre-trip waiting times and in-vehicle times for different service types, price information (the TNC may display such information to passengers in the app), and personal social-demographical information. Although we have demonstrated the superiority of the BSTF model in classification accuracy and probabilistic interpretability, the models based on the existing dataset may not be suitable for individual behavior analysis or travel choice prediction. In this appendix, we develop the BSTF models and other machine learning models with two additional scenarios to fulfill the utilization of the Hangzhou dataset.

In the first scenario, the service type classification model is developed based on the land-use data, order creation time, trip distance, estimated pre-trip waiting time, and in-vehicle time for different service types. Similar to conventional discrete choice models, we assume a passenger knows the pre-trip waiting time and in-vehicle time before making a decision. We assume the pre-trip in-vehicle times of taxi, DiDi Express and DiDi Hitch for one trip request are the same. This is because DiDi would estimate the pre-trip in-vehicle time of one service type based on the shortest path; without time-dependent roadway congestion information, we simply assume the shortest paths for the three types are identical (note that there is no detour in DiDi Hitch). Moreover, since DiDi is capable of making accurate travel time prediction based on advanced algorithms and real-time data (Li et al., 2018; Wang et al., 2018; Fu et al., 2020), the pre-trip in-vehicle times are assumed to be equal to post-trip in-vehicle times in the Hangzhou dataset. For one trip record, the pre-trip waiting time of the chosen service type is the same as the post-trip waiting time in the dataset; the waiting times of the other two unchosen types are estimated based on linear regression models. We conduct regressions based on the profile (percentile) of post-trip waiting times in the dataset. Fig. B.1 presents the two regression lines: the red line is obtained with $x$and y to be the 0.1, 0.2, …, 0.9 percentile points of taxi and DiDi Hitch post-trip waiting times, respectively; and the blue line is estimated based on the percentile points of taxi and DiDi Express post-trip waiting times. The intercept of the blue line is small (i.e., 18 s), and the slope is 1.14. This means the waiting time of DiDi Express is around 14% longer than taxi, so there can be more taxis than DiDi Express vehicles in the supply pool. For the red line, the intercept is 196 s, and it indicates DiDi Hitch services are generally booked in advance, or drivers need some time to confirm a shared ride. All the estimated pre-trip waiting times and in-vehicle travel times are discretized based on the rules in Table 1.

In the second scenario, we eliminate all the post-trip and estimated pre-trip information and only use land-use data, order creation time and trip distance to fit the BSTF model and other state-of-art machine learning models. We use this scenario to illustrate the capability of passenger choice prediction when the TNC only knows the land-use characteristics of a passenger's origin and destination and the distance between them.

The classification accuracy of different models in the two extra scenarios is shown in Table B.1. For each scenario, we note that the BSTF model still provides the lowest classification error among all the machine learning methods used in this paper. However, comparing Table B.1(a), Table B.1(b), and Table 3, we find that the BSTF model offers the highest, medium, and lowest accuracy with actual post-trip information, estimated pre-trip information, and no pre-trip/post-trip information, respectively. This means the estimation of pre-trip information is not perfectly accurate but still makes some sense. In the future research, a more comprehensive dataset is needed to fully explore the value of the proposed BSTF model in other applications such as travel behavior analysis, choice prediction, etc.

Furthermore, as stated in footnote 5, there are 54,776 passengers and some of them could make multiple orders with similar travel patterns. For the individual-level travel behavior analysis, repeated trips made by the same passengers from the same origins and destinations around the same departure time could lead to biased results.

To examine the impact of repeated/multiple orders, we train the BSTF models with different subsets of the DiDi Hangzhou data. The subsets are created by randomly excluding some repeated/multiple data records with the same combination of passenger ID, origin zone ID, destination zone ID, variables "create_DoW" and "create_time" (defined in Table 1); such a subset is referred to as a combination-based subset. We define the repeated exclusion rate (RER) as the percentage of repeated orders to be excluded. For

[7] Please refer to webpage https://en.wikipedia.org/wiki/Precision_and_recall for a more detailed introduction.

instance, with an RER of 40%, if a passenger made 5 orders that share the same origin zone, destination zone, "create_DoW" and "create_time" in the original dataset, we randomly pick 1 order to add into the combination-based subset, but for the rest 4 orders, each one has a 40% chance to be dropped and excluded for the subset; however, if a passenger only made 1 order with a specific combination, we directly add the data record into the combination-based subset. We randomly divide a combination-based subset into a training set with 80% data and a test with 20% data and estimate the BSTF model. The results are given in Table B.2, in which we try different RERs of repeated records from 0 (i.e., the original dataset) to 100% (i.e., one order per combination). First, we note that even after excluding 100% of such repeated orders, there are still 171,226 data records with a unique combination. This indicates that there are not many repeated orders with respect to the combination of passenger, origin, destination, and departure time in the original dataset. Second, we find that as more repeated/multiple data records are excluded, the classification error nearly stays unchanged. This is probably because the size of the subset does not change significantly. Therefore, the aforementioned biased issue can be negligible in this paper, given the tiny change in the classification error before and after removing the repeated orders in terms of the predefined combination.

Based on the same exclusion method, we drop repeated orders made by the same passengers and use the passenger-based subsets to train the BSTF models. The measurements of different subsets/models are illustrated in Table B.3. Without individual-level demographic variables, the results indicate that repeated/multiple data records from the same passengers are still helpful to the

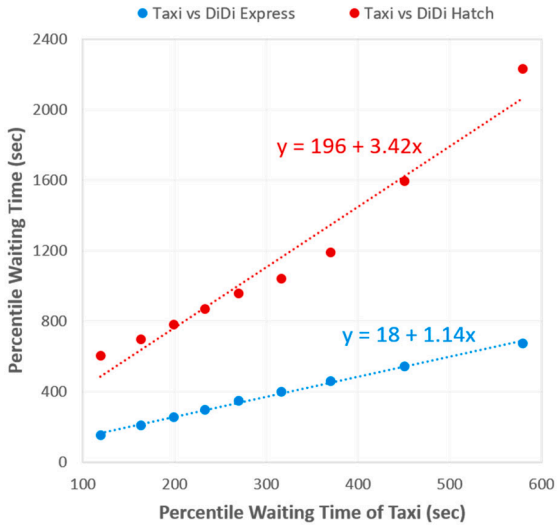

Fig. B1. Linear regression of waiting times.

Table B.1
Classification error of extra case studies.

(a) With estimated pre-trip information				
Model	Percentage of Training Set			
	10%	20%	50%	80%
BSTF	**0.211**	**0.208**	**0.205**	**0.203**
DT	0.215	0.211	0.207	0.206
NB	0.275	0.267	0.262	0.262
NN	0.218	0.211	0.208	0.205
RF	0.219	0.214	0.211	0.209
SVM	0.214	0.212	0.212	0.212
(b) Without post-trip or pre-trip information				
Model	Percentage of Training Set			
	10%	20%	50%	80%
BSTF	**0.213**	**0.212**	**0.212**	**0.211**
DT	0.217	0.215	0.213	0.213
NB	0.257	0.241	0.232	0.231
NN	0.216	0.214	0.214	0.212
RF	0.225	0.218	0.215	0.214
SVM	0.215	0.213	0.213	0.213

Table B.2
Summary of models trained by subsets of DiDi Hangzhou data (combination-based).

Size of Subset	RER (%)	Avg. Num. of Orders per Combination	Classification Error
212,310	0	1.24	0.180
205,470	20	1.20	0.180
198,613	40	1.16	0.180
190,062	60	1.11	0.181
181,419	80	1.06	0.181
171,226	100	1.00	0.182

Table B.3
Summary of models trained by subsets of DiDi Hangzhou data (passenger-based).

Size of Subset	RER (%)	Avg. Num. of Orders per Passenger	Classification Error
212,310	0	3.85	0.180
180,871	20	3.28	0.181
127,304	40	2.71	0.182
100,520	60	2.14	0.185
73,736	80	1.57	0.188
54,776	100	1.00	0.191

ridesharing pattern analysis in this paper. Since this study is mainly about aggregate ridesharing pattern analysis, including these repeated/multiple trips could help identify principal patterns and key latent classes.

Appendix C. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trc.2020.102916.

References

Agatz, N., Erera, A., Savelsbergh, M., Wang, X., 2012. Optimization for dynamic ride-sharing: a review. Eur. J. Oper. Res. 223 (2), 295–303.
Ben-Akiva, M., Lerman, S., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press.
Bhat, C., 2018. A new flexible multiple discrete-continuous extreme value (MDCEV) choice model. Transport. Res. Part B: Methodol. 110, 261–279.
Brathwaite, T., Vij, A., Walker, J.L., 2017. Machine learning meets microeconomics: The case of decision trees and discrete choice. arXiv preprint arXiv:1711.04826.
Breiman, L., 2001. Random forests. Machine Learn. 45 (1), 5–32.
Chan, N., Shaheen, S., 2012. Ridesharing in North America: past, present, and future. Transp. Rev. 32 (1), 93–112.
Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 1–27.
Chen, X., He, Z., Chen, Y., Lu, Y., Wang, J., 2019a. Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. Transport. Res. Part C: Emerg. Technol. 104, 66–77.
Chen, X., He, Z., Sun, L., 2019b. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. Transport. Res. C: Emerg. Technol. 98, 73–84.
Chen, X., Yang, J., Sun, L., 2020. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. Transport. Res. C: Emerg. Technol. 117, 102673.
Chen, X., Zahiri, M., Zhang, S., 2017. Understanding ridesplitting behavior of on-demand ride services: an ensemble learning approach. Transport. Res. C: Emerg. Technol. 76, 51–70.
Chen, X., Zheng, H., Wang, Z., Chen, X., 2018. Exploring impacts of on-demand ridesplitting on mobility via real-world ridesourcing data and questionnaires. Transportation in press.
Chen, X., Zhu, Z., He, X., Zhang, L., 2015. Surrogate-based optimization for solving a mixed integer network design problem. Transp. Res. Rec. 2497 (1), 124–136.
Correia, G., Viegas, J., 2011. Carpooling and carpool clubs: Clarifying concepts and assessing value enhancement possibilities through a stated preference web survey in Lisbon, Portugal. Transport. Res. Part A: Pol. Pract. 45 (2), 81–90.
de Dios Ortuzar, J., 1983. Nested logit models for mixed-mode travel in urban corridors. Transport. Res. A: General 17 (4), 283–299.
de Lathauwer, L., de Moor, B., Vandewalle, J., 2000. A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21 (4), 1253–1278.
Dias, F.F., Lavieri, P.S., Garikapati, V.M., Astroza, S., Pendyala, R.M., Bhat, C.R., 2017. A behavioral choice model of the use of car-sharing and ride-sourcing services. Transportation 44 (6), 1307–1323.
Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., & Leisch, M. F., 2009. Package 'e1071'. R Software package, available at http://cran.rproject.org/web/packages/e1071/index.html.
Dong, Y., Wang, S., Li, L., Zhang, Z., 2018. An empirical study on travel patterns of internet based ride-sharing. Transport. Res. C: Emerg. Technol. 86, 1–22.
Dunlavy, D., Kolda, T., Acar, E., 2011. Temporal link prediction using matrix and tensor factorizations. ACM Trans. Knowl. Discovery Data 5 (2), 1–27.
Ferguson, E., 1997. The rise and fall of the American carpool: 1970–1990. Transportation 24 (4), 349–376.
Fu, K., Meng, F., Ye, J., Wang, Z., 2020. Compacteta: A fast inference system for travel time prediction. In: In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3337–3345.
George, E., McCulloch, R., 1997. Approaches for Bayesian variable selection. Statistica Sinica 7 (2), 339–373.
Giuliano, G., 1992. Transportation demand management: promise or panacea? J. Am. Plan. Assoc. 58 (3), 327–335.
Gliebe, J., Koppelman, F., 2002. A model of joint activity participation between household members. Transportation 29 (1), 49–72.
Han, Y., Moutarde, F., 2016. Analysis of large-scale traffic dynamics in an urban transportation network using non-negative tensor factorization. Int. J. Intell. Transp. Syst. Res. 14 (1), 36–49.

Hensher, D., Ton, T., 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. Transport. Res. E: Logist. Transport. Rev. 36 (3), 155–172.
Hunt, J., McMillan, J., 1997. Stated-preference examination of attitudes toward carpooling to work in Calgary. Transp. Res. Rec. 1598 (1), 9–17.
Kamargianni, M., Dubey, S., Polydoropoulou, A., Bhat, C., 2015. Investigating the subjective and objective factors influencing teenagers' school travel mode choice – an integrated choice and latent variable model. Transport. Res. A: Pol. Pract. 78, 473–488.
Ke, J., Zheng, H., Yang, H., Chen, X., 2017. Short-term forecasting of passenger demand under on-demand ride services: a spatio-temporal deep learning approach. Transport. Res. Part C: Emerg. Technol. 85, 591–608.
Kim, Y., Choi, S., 2007. Nonnegative Tucker decomposition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA.
Kolda, T., Bader, B., 2009. Tensor decompositions and applications. SIAM Rev. 51 (3), 455–500.
Lemp, J., Kockelman, K., Damien, P., 2010. The continuous cross-nested logit model: Formulation and application for departure time choice. Transport. Res. B: Methodol. 44 (5), 646–661.
Li, W., Pu, Z., Li, Y., Ban, X., 2019. Characterization of ridesplitting based on observed data: a case study of Chengdu, China. Transport. Res. C: Emerg. Technol. 100, 330–353.
Li, Y., Fu, K., Wang, Z., Shahabi, C., Ye, J., Liu, Y., 2018. Multi-task representation learning for travel time estimation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1695–1704).
McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.
Miller, E.J., Mohammadian, A., 2003. An empirical investigation of household vehicle type choice decision. In: The 82nd Annual Transportation Research Board Meeting, Washington, DC.
Miller, E., Roorda, M., Carrasco, J., 2005. A tour-based model of travel mode choice. Transportation 32 (4), 399–422.
Morency, C., 2007. The ambivalence of ridesharing. Transportation 34 (2), 239–253.
Neoh, J., Chipulu, M., Marshall, A., 2017. What encourages people to carpool? An evaluation of factors with meta-analysis. Transportation 44 (2), 423–447.
Poirier, D., 1996. A Bayesian analysis of nested logit models. J. Econometr. 75 (1), 163–181.
Ripley, B., Venables, W., Ripley, M.B., 2016. Package 'nnet'. R package version, 7, 3-12.
Schaller, B., 2018. The New Automobility: Lyft, Uber and the Future of American Cities. Schaller Consulting. New York: Schaller Consulting.
Shashua, A., & Hazan, T., 2005. Non-negative tensor factorization with applications to statistics and computer vision. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 792–799, Bonn, Germany.
Small, K., 1987. A discrete choice model for ordered alternatives. Econometrica: J. Econometr. Soc. 55 (2), 409–424.
Su, Q., Zhou, L., 2012. Parking management, financial subsidies to alternatives to drive alone and commute mode choices in Seattle. Reg. Sci. Urban Econ. 42 (1–2), 88–97.
Sun, L., Erath, A., Cai, M., 2018. A hierarchical mixture modeling framework for population synthesis. Transport. Res. B: Methodol. 114, 199–212.
Sun, L., Axhausen, K., 2016. Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transport. Res. Part B: Methodol. 91, 511–524.
Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y., Li, F., 2013. A tensor-based method for missing traffic data completion. Transport. Res. Part C: Emerg. Technol. 28, 15–27.
Tan, H., Wu, Y., Shen, B., Jin, P., Ran, B., 2016. Short-term traffic prediction based on dynamic tensor completion. IEEE Trans. Intell. Transp. Syst. 17 (8), 2123–2133.
Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press.
Tucker, L., 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31 (3), 279–311.
Vannieuwenhoven, N., Vandebril, R., Meerbergen, K., 2012. A new truncation strategy for the higher-order singular value decomposition. SIAM J. Scientific Computing 34 (2), A1027–A1052.
Wang, X., Agatz, N., Erera, A., 2017. Stable matching for dynamic ride-sharing systems. Transport. Sci. 52 (4), 850–867.
Wang, Z., Fu, K., Ye, J., 2018. Learning to estimate the travel time. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 858–866.
Wang, Y., Wang, S., Wang, J., Wei, J., Wang, C., 2020. An empirical study of consumers' intention to use ride-sharing services: using an extended technology acceptance model. Transportation 47 (1), 397–415.
Wets, G., Vanhoof, K., Arentze, T., Timmermans, H., 2000. Identifying decision structures underlying activity patterns: an exploration of data mining algorithms. Transp. Res. Rec. 1718 (1), 1–9.
Yang, Y., Dunson, D., 2016. Bayesian conditional tensor factorizations for high-dimensional classification. J. Am. Stat. Assoc. 111 (514), 656–669.
Zeileis, A., Hothorn, T., Hornik, K., 2008. Model-based recursive partitioning. J. Computat. Graph. Statist. 17 (2), 492–514.
Zhang, H., Chen, P., Zheng, J., Zhu, J., Yu, G., Wang, Y., Liu, H.X., 2019. Missing data detection and imputation for urban ANPR system using an iterative tensor decomposition approach. Transport. Res. Part C: Emerg. Technol. 107, 337–355.
Zhang, Y., Xie, Y., 2008. Travel mode choice modeling with support vector machines. Transp. Res. Rec. 2076 (1), 141–150.
Zhu, Z., Chen, X., Xiong, C., Zhang, L., 2018. A mixed Bayesian network for two-dimensional decision modeling of departure time and mode choice. Transportation 45 (5), 1499–1522.
Zhu, Z., Mardan, A., Zhu, S., Yang, H., 2021. Capturing the interaction between travel time reliability and route choice behavior based on the generalized Bayesian traffic model. Transport. Res. B: Methodol. 143, 48–64.
Zhu, Z., Peng, B., Xiong, C., Zhang, L., 2016. Short-term traffic flow prediction with linear conditional Gaussian Bayesian network. J. Adv. Transport. 50 (6), 1111–1123.
Zhu, Z., Qin, X., Ke, J., Zheng, Z., Yang, H., 2020. Analysis of multi-modal commute behavior with feeding and competing ridesplitting services. Transport. Res. A: Pol. Pract. 132, 713–727.
Zhu, Z., Li, X., Liu, W., Yang, H., 2019a. Day-to-day evolution of departure time choice in stochastic capacity bottleneck models with bounded rationality and various information perceptions. Transport. Res. E: Logist. Transport. Rev. 131, 168–192.
Zhu, Z., Zhu, S., Zheng, Z., Yang, H., 2019b. A generalized Bayesian traffic model. Transport. Res. C: Emerg. Technol. 108, 182–206.