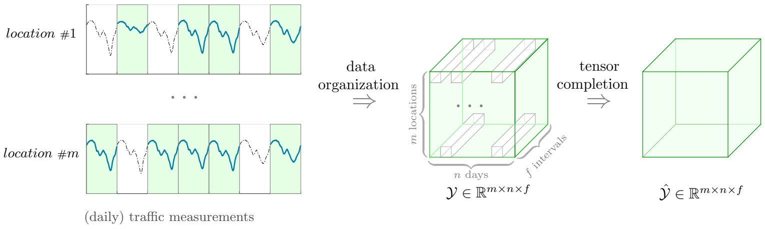
Tensor Methods for High-Dimensional Spatiotemporal Data Modeling

Xinyu Chen (Polytechnique Montreal, University of Montreal)

Spatiotemporal data modeling is an emerging research area due to the availability of data sources and the development of machine learning approaches. Spatiotemporal data are by nature time series in the form of various algebraic structures (e.g., matrix/tensor) and associated with complicated spatiotemporal data patterns. Typically, these data involve multiple data behaviors including high-dimensionality, multidimensionality, sparsity, uncertainty, nonstationarity, and nonlinearity, posing both methodological and practical challenges. The goal of my research is to develop data-driven machine learning approaches (e.g., matrix/tensor methods) for modern spatiotemporal data modeling with theoretical and statistical foundations. My work focuses on answering some scientific questions: (**Q1**) **How to make accurate reconstruction from sparse spatiotemporal traffic data?** (**Q2**) **How to perform efficient forecasting on the spatiotemporal traffic data with missing values?** (**Q3**) **How to discover dynamic patterns from spatiotemporal data with time-varying system behaviors?** The goal is to formulate these questions appropriately from a machine learning perspective.

Q1. Spatiotemporal Traffic Data Imputation

Motivation. With the development of intelligent transportation systems, large quantities of traffic flow data are collected on a continuous basis from various sources, such as loop detectors, cameras, and floating vehicles. These data capture the underlying states and dynamics of transportation networks and the whole system becomes beneficial to many traffic operation and management applications, including route planning, traffic signal control, travel time estimation, and traffic flow forecasting. However, a common drawback that undermines the use of such spatiotemporal data is the missing data problem, which may result from various operational issues such as sensor failure and network communication problems, leading to data corruption and missing values. To make full use of the incomplete spatiotemporal data, a critical research question is how to provide robust estimates of these missing values.



location #1
location #m
(daily) traffic measurements
data organization
m locations
n days
f intervals
$\mathcal{Y} \in \mathbb{R}^{m \times n \times f}$
tensor completion
$\hat{\mathcal{Y}} \in \mathbb{R}^{m \times n \times f}$

Figure 1: Illustration of tensor completion for partially observed traffic measurements.

Methodologies & Contributions. For modeling spatiotemporal traffic data, it is critical to take into account spatiotemporal dependencies and patterns. Tensor decomposition is a classical technique for capturing the multidimensional structural dependencies (Kolda and Bader, 2009). In my research, I have developed a Bayesian probabilistic tensor decomposition model for reconstructing missing data in a spatiotemporal setting and demonstrated that tensor data representation can characterize the spatiotemporal data generation process, in the meanwhile providing an appropriate solution to the traffic data imputation task as shown in Figure 1 (Chen, He, Chen, Lu and Wang, 2019; Chen, He and Sun, 2019).

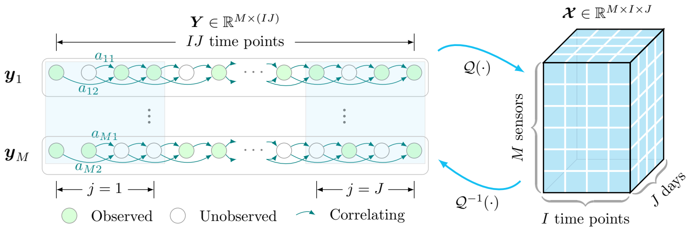
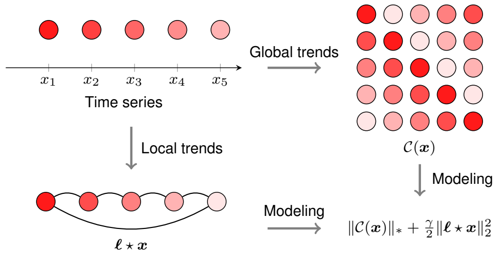


Figure 2: Illustration of LATC framework for spatiotemporal traffic data imputation.

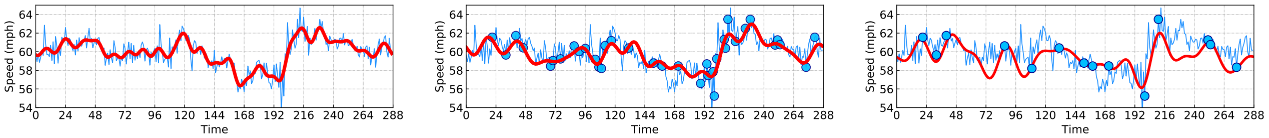


x_1
x_2
x_3
x_4
x_5
Time series
Global trends
$\mathcal{C}(\boldsymbol{x})$
Local trends
$\boldsymbol{\ell} \star \boldsymbol{x}$
Modeling
Modeling
$\|\mathcal{C}(\boldsymbol{x})\|_* + \frac{\gamma}{2}\|\boldsymbol{\ell} \star \boldsymbol{x}\|_2^2$

Figure 3: Illustration of the proposed LCR model for traffic time series imputation.

To fully utilize the low-rankness property and characterize the global and local trends in the partially observed traffic data, I have developed a low-rank autoregressive tensor completion (LATC, see Figure 2) model (Chen, Lei, Saunier and Sun, 2022), in which the autoregressive time series process is integrated into a low-rank model to capture temporal dynamics in spatiotemporal traffic data. The proposed LATC model reinforces both global and local trends modeling in a low-rank framework, in the meanwhile demonstrating superior imputation performance over some baseline models on several traffic datasets with complicated missing data scenarios.

Although these tensor models allow one to accurately reconstruct missing values in spatiotemporal traffic datasets, they



(a) Fully observed.

(b) 90% missing values.

(c) 95% missing values.

Figure 4: Traffic time series imputation on the freeway traffic speed data in Portland, USA. The blue and red curves refer to the ground truth time series and the reconstructed time series by LCR, respectively. Blue circles indicate the partial observations.

usually suffer from high complexity and expensive computational cost. In my most recent study, to handle the large-scale traffic data imputation, I introduced a circular Laplacian kernel and used it to define the temporal regualrization for characterizing local trends in traffic time series. Following that definition, the Laplacian temporal regularization can be formulated as a circular convolution operation. Based on it, I developed a low-rank completion approach—Laplacian convolutional representation (LCR, see Figure 3)—by modeling global trends as the nuclear norm of circulant matrix and modeling local trends by Laplacian temporal regularization (Chen, Cheng, Saunier and Sun, 2022).[1] According to the properties of circulant matrix and circular convolution, the LCR model invokes a fast implementation through the fast Fourier transform, showing accurate imputation results on some real-world traffic datasets (e.g., the reconstructed time series in Figure 4) and strong generalization ability to high-dimensional and large problems.

Significance. In the past five years, I have proposed several low-rank tensor models for spatiotemporal data imputation, including conventional tensor decomposition (Chen, He and Wang, 2018), Bayesian tensor decomposition (Chen, He, Chen, Lu and Wang, 2019; Chen, He and Sun, 2019), Low-rank tensor completion (Chen, Yang and Sun, 2020), and low-rank autoregressive tensor models (Chen, Chen, Saunier and Sun, 2021; Chen, Lei, Saunier and Sun, 2022). On the top of these works, I have developed a GitHub project—**transdim** (i.e., machine learning for **trans**portation **d**ata **im**putation and prediction)[2]—with Prof. Lijun Sun (McGill University) and Prof. Nicolas Saunier (Polytechnique Montreal) as a benchmark platform for modeling sparse traffic data. So far, the **transdim** project has attracted a lot of attention (1,000+ stars & 270+ forks) with the publicly available datasets and the Python implementation of state-of-the-art models. The whole research aims to build advanced low-rank machine learning frameworks for spatiotemporal data modeling. In addition, our project brings fundamental research advances to the general field of spatiotemporal modeling and promotes its application to other scientific areas.

Q2. Spatiotemporal Traffic Data Forecasting

Motivation. With recent advances in sensing technologies, large-scale and multidimensional time series data—in particular spatiotemporal data—are collected on a continuous basis from various types of sensors and applications. Making prediction on these time series, such as forecasting urban traffic states and regional air quality, serves as a foundation to many real-world applications and benefits many scientific fields. However, given the complex spatiotemporal dependencies in these datasets, making efficient and reliable predictions for real-time applications has been a long-standing and fundamental research challenge.

Despite the vast body of literature on time series analysis from many scientific areas, three emerging issues in modern sensor technologies have posed challenges to classical modeling frameworks. First, modern time series are often large-scale, collected from a large number of locations/sensors simultaneously. One intuitive example is that the highway traffic Performance Measurement System (PeMS) in California, USA, an open intelligent transportation system, consists of more than 35,000 detectors, and it has been gathering traffic flow and speed information with 30-second time resolution since 1999 (Chen, Petty, Skabardonis, Varaiya and Jia, 2001), resulting in high-dimensional data. Second, modern time series generated by advanced sensing technologies are usually multidimensional with different attributes. Third, most existing time series models require complete time series data as input, while the missing data problem is almost inevitable due to various factors in real-world time series datasets. Taken together, it has become a critical challenge to perform reliable forecasting on large-scale time series data in the presence of missing values.

Methodologies & Contributions. In my research, I have proposed a new Bayesian temporal factorization (BTF) framework which can effectively handle both missing data problem and large-scale/high-dimensional properties in modern spatiotemporal data (Chen and Sun, 2022). Vector autoregressive (VAR) process was integrated into the temporal matrix factorization (TMF, see Figure 5) framework under a spatiotemporal set-

[1] **Chen, X.**, Cheng, Z., Saunier, N. and Sun, L. (2022), 'Laplacian convolutional representation for traffic time series imputation'. (Under review at the first round, *IEEE Transactions on Signal Processing*)

[2] https://github.com/xinychen/transdim

ting. For multidimensional time series, there is a higher-order generalization of TMF to tensor data in a fully Bayesian treatment, namely, Bayesian temporal tensor factorization (BTTF). The BTF framework is fully Bayesian and gives a flexible solution to ensure model accuracy while avoiding overfitting on both matrix and tensor data.

Following the above study, I have proposed to address nonstationarity and sparsity issues by reinforcing temporal modeling through differenced VAR processes (Chen, Zhang, Zhao, Saunier and Sun, 2022). To solve the optimization problem of the proposed nonstationary TMF model, here is an alternating minimization algorithm in which the complicated subproblem about latent temporal factors—the generalized Sylvester equation involving both partially observed matrix factorization and differenced VAR—is solved through the conjugate gradient method. Our experiments on two Uber movement speed datasets[3] (featured as high-dimensional, sparse, and nonstationary) demonstrated the superiority of nonstationary TMF over state-of-the-art baseline models.

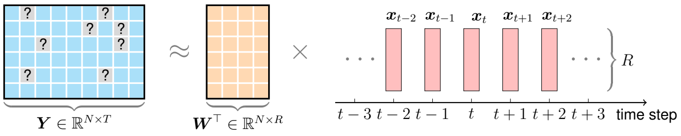

?
$Y \in \mathbb{R}^{N\times T}$
$\approx$
$W^\top \in \mathbb{R}^{N\times R}$
$\times$
x_{t-2} x_{t-1} x_t x_{t+1} x_{t+2}
R
$t-3$ $t-2$ $t-1$ t $t+1$ $t+2$ $t+3$
time step

Figure 5: Illustration of temporal matrix factorization.

Significance. This research aims to provide a flexible machine learning framework for spatiotemporal forecasting in the presence of missing values. Despite the low-rankness property underlying data, temporal modeling through a certain technique is essential for spatiotemporal data analysis because temporal correlations are the basis of time series modeling. Therefore, these research works provide valuable insight into real-world time series forecasting. To make our work reproducible, I created a GitHub project—**tracebase**—for time series forecasting on high-dimensional and sparse traffic data.[4]

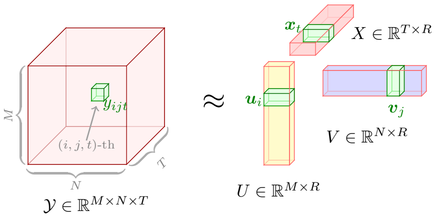

y_{ijt}
(i,j,t)-th
M
N
T
$\mathcal{Y} \in \mathbb{R}^{M\times N\times T}$
$\approx$
x_t
$X \in \mathbb{R}^{T\times R}$
u_i
v_j
$V \in \mathbb{R}^{N\times R}$
$U \in \mathbb{R}^{M\times R}$

Figure 6: Illustration of tensor factorization.

Q3. Pattern Discovery from Spatiotemporal Data

Motivation. Dynamic mechanisms that drive nonlinear systems are universally complex and obscure. Straightforwardly, one can investigate the behavior of a system by discovering the patterns from its observations. In practice, when we take observations from a real-world complex system, spatiotemporal data are one of the most widely encountered form relating to space and time and showing the characteristics of time series. With the remarkable development of sensing technologies, tremendous spatiotemporal data are accessible for analysis such as trajectories of human mobility (Zheng, 2015). In the literature, various spatiotemporal data were characterized by time series models. Leveraging time series models not only allows one to analyze spatiotemporal data but also makes it possible to discover inherent spatial and temporal patterns from the data over space and time.

Methodologies & Contributions. In my research, I have revisited the classical spatiotemporal data analysis problem and introduced a novel method to discover spatial and temporal patterns from time-varying systems in a data-driven fashion (Chen, Zhang, Chen, Saunier and Sun, 2022).[5] I have developed a fully time-varying reduced-rank VAR model which allows one to discover interpretable dynamic modes from spatiotemporal data. The proposed model draws strong connections between time-varying VAR and tensor factorization that address the over-parameterization issue and reveal spatial and temporal patterns in the latent spaces. For empirical analysis, I have demonstrated the model capability of the time-varying reduced-rank VAR for discovering interpretable patterns from extensive spatiotemporal datasets, including fluid dynamics (see Figure 7), sea surface temperature, USA surface temperature (see Figure **??**), and NYC taxi trips (see Figure 9). The evaluation results show that the latent variables in the tensor factorization of our model can reveal both spatial and temporal patterns. Time-varying system behaviors underlying these spatiotemporal data can be clearly identified by our model.

Significance. This research presents a time-varying reduced-rank VAR model for discovering interpretable modes from time series, providing insights into modeling real-world spatiotemporal systems.

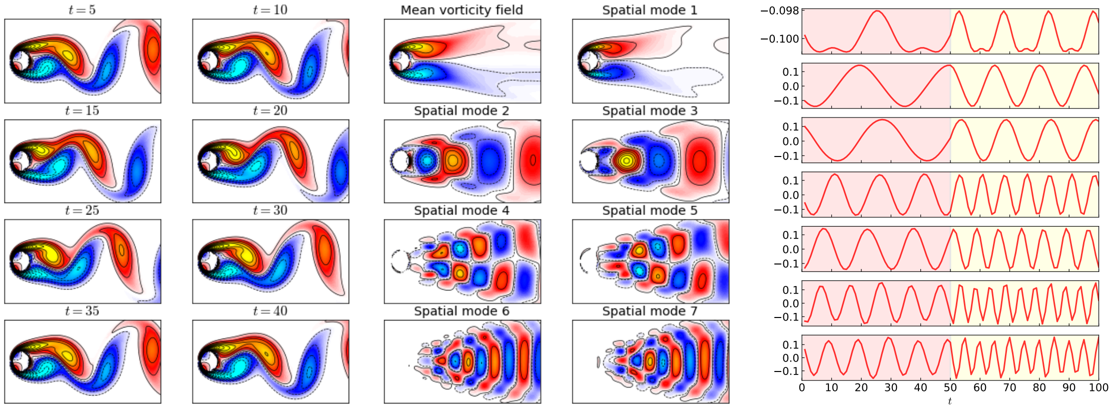
References

Chen, C., Petty, K., Skabardonis, A., Varaiya, P. and Jia, Z. (2001), 'Freeway performance measurement system: mining loop detector data', *Transportation Research Record* **1748**(1), 96–102.

[3] https://movement.uber.com/

[4] https://github.com/xinychen/tracebase

[5] **Chen, X.**, Zhang, C., Chen, X., Saunier, N. and Sun, L. (2022). 'Discovering dynamic patterns from spatiotemporal data with time-varying low-rank autoregression'. *IEEE Transactions on Knowledge and Data Engineering*. Accepted for publication.

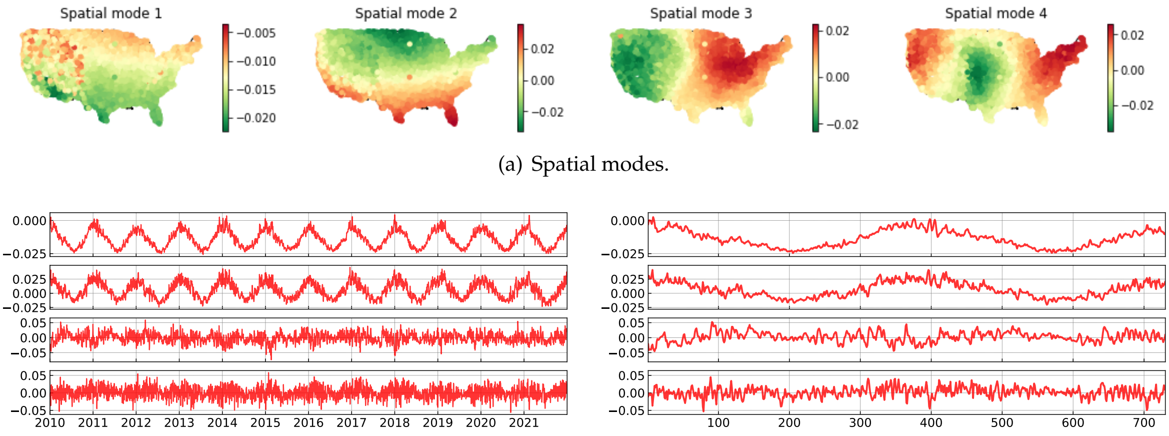


(a) Fluid flow (original data.)

(b) Fluid flow (original data.)

(c) Fluid flow (original data.)

Figure 7: Fluid flow and spatial/temporal modes to demonstrate the model. (a) Heatmaps (snapshots) of the fluid flow at times $t = 5, 10, \ldots, 40$. It shows that the snapshots at times $t = 5$ and $t = 35$ are even same, and the snapshots at times $t = 10$ and $t = 40$ are also even same, allowing one to infer the seasonality as 30 for the first 50 snapshots. (b) Mean vorticity field and spatial modes of the fluid flow. Spatial modes are plotted by the columns of $\boldsymbol{W}$ in which seven panels correspond to the rank $R = 7$. Note that the colorbars of all modes are on the same scale. (c) Temporal modes of the fluid flow in $\boldsymbol{X}$. Seven panels correspond to the rank $R = 7$.



(a) Spatial modes.

(b) Temporal modes over the 12-year period.

(c) Temporal modes over the two-year period.

Figure 8: Model application and results on the daily temperature data in the United States Mainland. (a) Geographical distribution of four spatial modes of the temperature data achieved by our model. (b) Four temporal modes (over the 12-year period) of the temperature data achieved by our model. From the top panel to the bottom panel, we have the temporal modes 1-4, respectively. (c) Four temporal modes (during the two years (730 days) from 2010 to 2011) of the temperature data achieved by our model.

Chen, X., Chen, Y., Saunier, N. and Sun, L. (2021), 'Scalable low-rank tensor learning for spatiotemporal traffic data imputation', *Transportation Research Part C: Emerging Technologies* **129**, 103226.

Chen, X., Cheng, Z., Saunier, N. and Sun, L. (2022), 'Laplacian convolutional representation for traffic time series imputation', *arXiv preprint arXiv:2212.01529* .

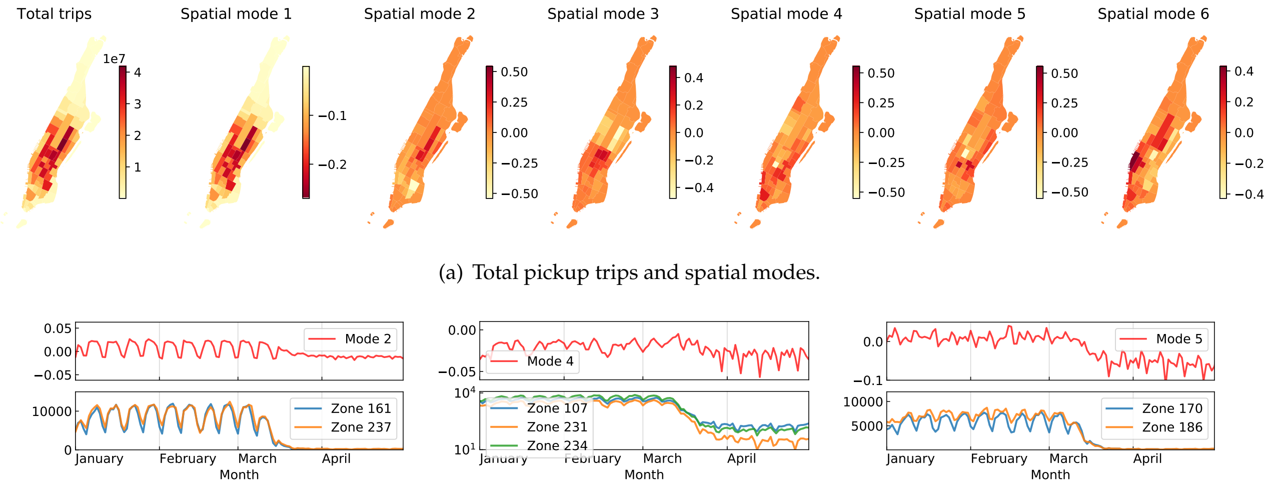
Chen, X., He, Z., Chen, Y., Lu, Y. and Wang, J. (2019), 'Missing traffic data imputation and pattern discovery with a bayesian augmented tensor factorization model', *Transportation Research Part C: Emerging Technologies* **104**, 66–77.

Chen, X., He, Z. and Sun, L. (2019), 'A bayesian tensor decomposition approach for spatiotemporal traffic data imputation', *Transportation research part C: emerging technologies* **98**, 73–84.

Chen, X., He, Z. and Wang, J. (2018), 'Spatial-temporal traffic speed patterns discovery and incomplete data recovery via svd-combined tensor decomposition', *Transportation research part C: emerging technologies* **86**, 59–77.

Chen, X., Lei, M., Saunier, N. and Sun, L. (2022), 'Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation', *IEEE Transactions on Intelligent Transportation Systems* **23**(8), 12301–12310.

Chen, X. and Sun, L. (2022), 'Bayesian temporal factorization for multidimensional time series prediction',



(a) Total pickup trips and spatial modes.

(b) Temporal mode 2 and taxi trips.

(c) Temporal mode 4 and taxi trips.

(d) Temporal mode 5 and taxi trips.

Figure 9: NYC taxi pickup trips and their spatial and temporal modes achieved by our model. We zoom in the temporal modes in the first four months of 2020. These modes reveal the total trip reduction due to the COVID-19 pandemic since March 2020. (a) Total trips and spatial modes revealed by $\boldsymbol{W}$. (b-d) refer to temporal mode 2, 4, 5, respectively; note that the bottom panels of these temporal modes uncover the trip time series of certain taxi zones.

IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4659–4673.

Chen, X., Yang, J. and Sun, L. (2020), ‘A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation’, *Transportation Research Part C: Emerging Technologies* **117**, 102673.

Chen, X., Zhang, C., Chen, X., Saunier, N. and Sun, L. (2022), ‘Discovering dynamic patterns from spatiotemporal data with time-varying low-rank autoregression’, *arXiv preprint arXiv:2211.15482* .

Chen, X., Zhang, C., Zhao, X.-L., Saunier, N. and Sun, L. (2022), ‘Nonstationary temporal matrix factorization for multivariate time series forecasting’, *arXiv preprint arXiv:2203.10651* .

Kolda, T. G. and Bader, B. W. (2009), ‘Tensor decompositions and applications’, *SIAM review* **51**(3), 455–500.

Zheng, Y. (2015), ‘Trajectory data mining: an overview’, *ACM Transactions on Intelligent Systems and Technology (TIST)* **6**(3), 1–41.