Bayesian Temporal Factorization Models

Lijun Sun

Joint work with Xinyu Chen & Zhaocheng He (SYSU)

McGill University

February 20, 2019

Outline

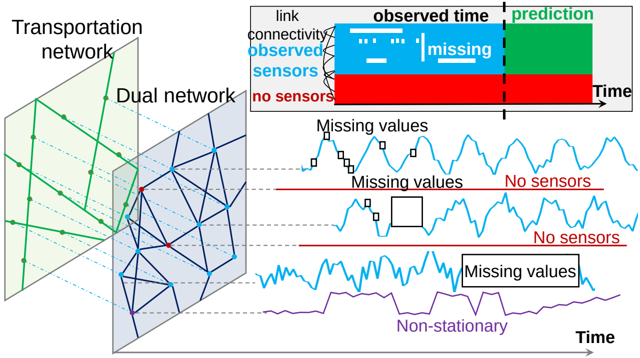
1. Motivation
2. Solution for multi-dimensional data imputation
3. Bayesian temporal factorization
4. Experiment and results
5. Conclusion

Outline

Traffic data from sensor networks

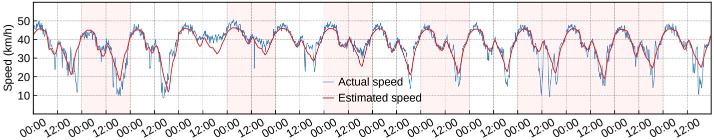
Spatiotemporal mobility/traffic data
(complex yet recurring spatiotemporal patterns; very useful in ITS)

However: large-scale, high-dimensional, incomplete, nonlinear, non-stationary, heterogeneous



Real-world applications

- Missing data imputation.
  - Sensor failure or network communication problem;
  - limited reports from crowdsourcing systems.



An example of traffic speed time series with missing values during two weeks, where the estimated traffic speed is achieved by BGCP[1] and red panels indicate missing values.

- Short-term traffic forecasting with incomplete observations.

[1] X. Chen, Z. He, L. Sun, 2019. A Bayesian tensor decomposition model for spatiotemporal traffic data imputation. Transp. Res. Part C, 98: 73-84.

Outline

Multi-dimensional traffic data

location $\times$ time series:

(or similar representation) is by nature a matrix.

origin $\times$ destination $\times$ time series:

(or similar representation) is by nature a third-order (higher-order) tensor.

Missing data problem: an intuitive example

Traffic volume matrix with row being location and column being time window, and its element is passing volume per 15 min:

$$Y = \begin{bmatrix} ? & 99 & 449 & 517 \\ ? & ? & 412 & ? \\ 192 & ? & 697 & 687 \\ 185 & ? & 699 & 657 \\ 164 & 58 & ? & ? \end{bmatrix}$$

Q: How to estimate the ? entries in this matrix?

Solution: matrix factorization technique

Find latent factors: Given a data matrix $Y \in \mathbb{R}^{m \times f}$, the ultimate goal is to find two matrices $W \in \mathbb{R}^{m \times r}$ and $X \in \mathbb{R}^{f \times r}$ with dimension r for approximating Y by $WX^\top$, e.g.,

$$Y = \begin{bmatrix} ? & 99 & 449 & 517 \\ ? & ? & 412 & ? \\ 192 & ? & 697 & 687 \\ 185 & ? & 699 & 657 \\ 164 & 58 & ? & ? \end{bmatrix}$$

$$\approx \begin{bmatrix} 15 & 10 \\ 7 & 12 \\ 12 & 20 \\ 9 & 11 \\ 8 & 18 \end{bmatrix} \begin{bmatrix} 4 & 5 & 11 & 20 \\ 7 & 1 & 28 & 23 \end{bmatrix} = \begin{bmatrix} 130 & 85 & 445 & 530 \\ 112 & 47 & 413 & 416 \\ 188 & 80 & 692 & 700 \\ 183 & 66 & 687 & 663 \\ 158 & 58 & 592 & 574 \end{bmatrix}$$

Solution: matrix factorization technique

Bayesian matrix factorization

Factorization for any data matrix $Y \in \mathbb{R}^{m \times f}$ is $y_{it} \approx \sum_{s=1}^{r} w_{is} x_{ts}$, We assume that the observation follows Gaussian assumption[2],

$$y_{it} \sim \mathcal{N}\left(\sum_{s=1}^{r} w_{is} x_{ts}, \tau^{-1}\right), (i,t) \in \Omega, \tag{1}$$

and further place conjugate (Gaussian-Wishart and Gamma) priors on model parameters, i.e.,

$$\begin{aligned} \boldsymbol{w}_i &\sim \mathcal{N}\left(\boldsymbol{\mu}_w, \Lambda_w^{-1}\right), i = 1, 2, ..., m, \\ \boldsymbol{x}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}_x, \Lambda_x^{-1}\right), t = 1, 2, ..., f, \\ \tau &\sim \text{Gamma}(\alpha, \beta). \end{aligned} \tag{2}$$

[2] R. Salakhutdinov, A. Mnih, 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo, ICML.

Tensor factorization technique

CANDECOMP/PARAFAC (CP) factorization

Definition

Given a third-order tensor $\mathcal{Y} \in \mathbb{R}^{m \times n \times f}$, the idea of CP factorization[a] is to approximate $\mathcal{Y}$ using a low-rank structure as follows

$$\mathcal{Y} \approx \sum_{s=1}^{r} \boldsymbol{u}_s \circ \boldsymbol{v}_s \circ \boldsymbol{x}_s \quad \text{or} \quad y_{ijt} \approx \sum_{s=1}^{r} u_{is} v_{js} x_{ts}$$

where $U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}, X \in \mathbb{R}^{f \times r}$ are factor matrices, and the symbol $\circ$ means vector outer product.

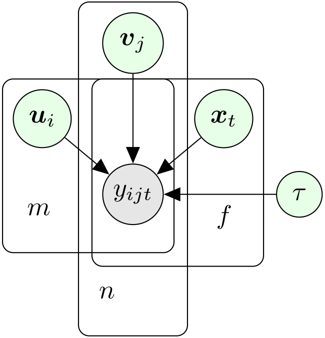
[a]Kolda, T. G., Bader, B. W., 2009. Tensor decompositions and applications. SIAM Review, 51(3):455-500.

Tensor factorization technique

Probabilistic/Bayesian CP factorization

Suppose that the observation follows Gaussian distribution,

$$y_{ijt} \sim \mathcal{N}\left(\sum_{s=1}^{r} u_{is} v_{js} x_{ts}, \tau^{-1}\right), (i, j, t) \in \Omega.$$



Tensor factorization technique

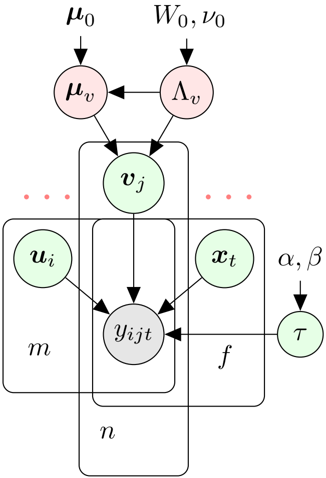
Probabilistic/Bayesian CP factorization

Suppose that the observation follows Gaussian distribution,

$$y_{ijt} \sim \mathcal{N}\left(\sum_{s=1}^{r} u_{is} v_{js} x_{ts}, \tau^{-1}\right), (i,j,t) \in \Omega.$$

In the Bayesian setting, we place conjugate priors on model parameters, i.e.,

$$\begin{aligned}
\boldsymbol{u}_i &\sim \mathcal{N}\left(\boldsymbol{\mu}_u, \Lambda_u^{-1}\right), i = 1, 2, ..., m, \\
\boldsymbol{v}_j &\sim \mathcal{N}\left(\boldsymbol{\mu}_v, \Lambda_v^{-1}\right), j = 1, 2, ..., n, \\
\boldsymbol{x}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}_x, \Lambda_x^{-1}\right), t = 1, 2, ..., f, \\
\tau &\sim \text{Gamma}(\alpha, \beta).
\end{aligned}$$



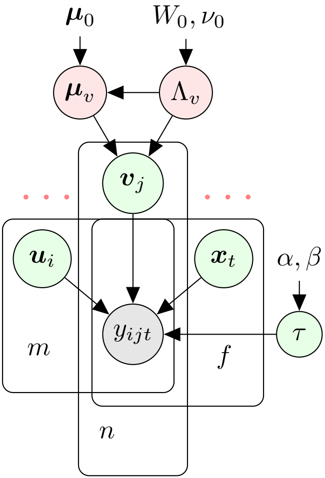
Tensor factorization technique

Model inference

- posterior $\propto$ likelihood $\times$ prior

$$p(\boldsymbol{v}_j \mid \mathcal{Y}, U, X, \tau, \boldsymbol{\mu}_v, \Lambda_v)$$
$$\propto p(\mathcal{Y} \mid U, \boldsymbol{v}_j, X, \tau) p(\boldsymbol{v}_j \mid \boldsymbol{\mu}_v, \Lambda_v)$$

- Tensor-based approaches have been widely applied in missing data imputation tasks. Key idea is to borrow information across dimensions.
- However, result is invariant to the shuffle subjects on each dimension. In particular, temporal dynamics is not fully taken into account.



Outline

1. Motivation
2. Solution for multi-dimensional data imputation
3. Bayesian temporal factorization
4. Experiment and results
5. Conclusion

Why do we apply temporal modeling?

Basic ideas

- Source data (from transportation systems)
  - is time-evolving;
  - consists of time series.
- Time series analysis
  - classical matrix/tensor models: learn discrete temporal factors;
  - temporal matrix/tensor models: capture time-evolving trends and patterns.
- Multi-dimensional forecasting
  - classical matrix/tensor models: similarity based estimation;
  - temporal regularized models: temporal pattern based prediction.

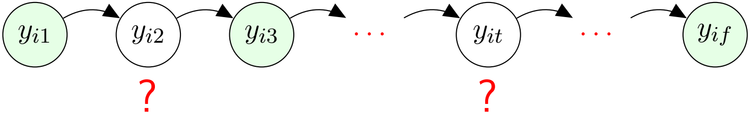
Why do we apply temporal modeling?

Basic ideas

- Source data (from transportation systems)
  - is time-evolving;
  - consists of time series.
- Time series analysis
  - classical matrix/tensor models: learn discrete temporal factors;
  - temporal matrix/tensor models: capture time-evolving trends and patterns.
- Multi-dimensional forecasting
  - classical matrix/tensor models: similarity based estimation;
  - temporal regularized models: temporal pattern based prediction.
- Integrate temporal modeling (e.g., AR) into matrix/tensor models may be a better alternative.

BTMF: Bayesian Temporal Matrix Factorization

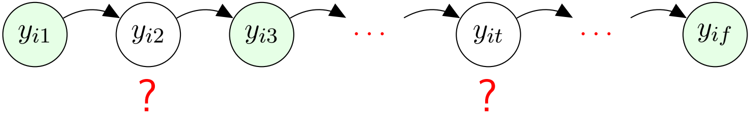
- Temporal modeling for incomplete time series:



Almost all time series related models cannot deal with such case!

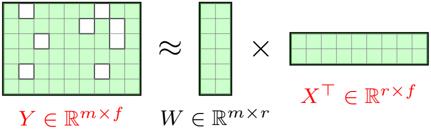
BTMF: Bayesian Temporal Matrix Factorization

- Temporal modeling for incomplete time series:

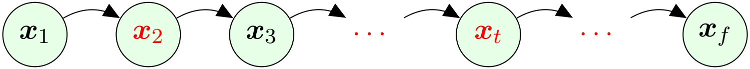


Almost all time series related models cannot deal with such case!

MF explained:



- Temporal modeling for latent temporal factors ($X^\top$):



BTMF: Bayesian Temporal Matrix Factorization

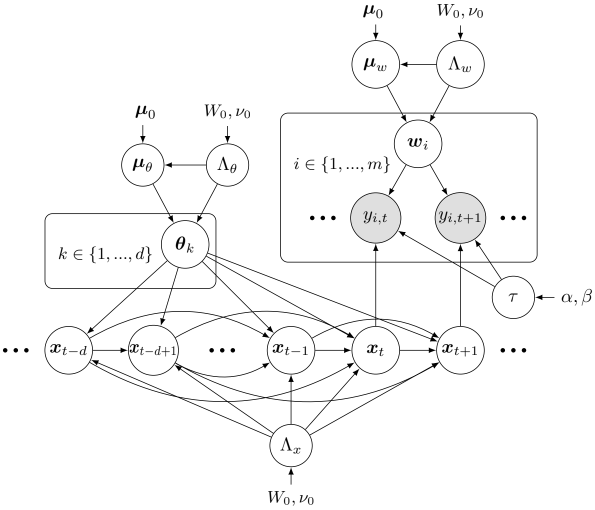
Given an incomplete matrix $Y \in \mathbb{R}^{m\times f}$, a Gaussian assumption over observations and conjugate priors (Gaussian-Wishart and Gamma) of model parameters are configured as follows

$$
\begin{aligned}
y_{it} &\sim \mathcal{N}\left(\sum_{s=1}^{r} w_{is}x_{ts}, \tau^{-1}\right), (i,t) \in \Omega, \\
\boldsymbol{w}_i &\sim \mathcal{N}\left(\boldsymbol{\mu}_w, \Lambda_w^{-1}\right), i = 1, 2, ..., m, \\
\boldsymbol{x}_t &\sim \mathcal{N}\left(\tilde{\boldsymbol{x}}_t, \Lambda_x^{-1}\right), t = 1, 2, ..., f, \\
\tau &\sim \text{Gamma}(\alpha, \beta),
\end{aligned}
$$

where temporal factors are constrained by an auto-regressive (AR) model, i.e.,

$$
\tilde{\boldsymbol{x}}_t = \begin{cases} \boldsymbol{0}, & \text{if } t \in \{1, 2, ..., h_d\}, \\ \sum_{k=1}^{d} \boldsymbol{\theta}_k \circledast \boldsymbol{x}_{t-h_k}, & \text{otherwise.} \end{cases}
$$

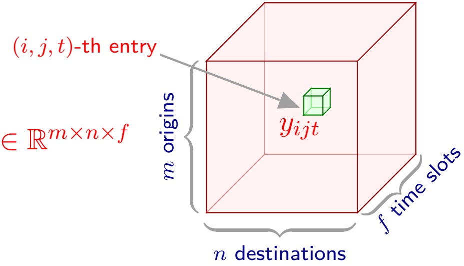
Graphical model



BTTF: Bayesian Temporal Tensor Factorization

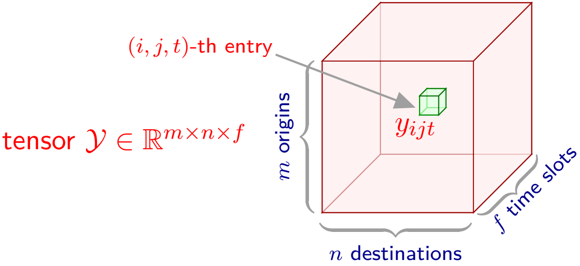
- Tensor representation (e.g., third-order trip volume data)

tensor $\mathcal{Y} \in \mathbb{R}^{m \times n \times f}$

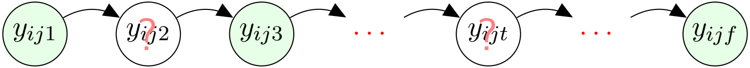


BTTF: Bayesian Temporal Tensor Factorization

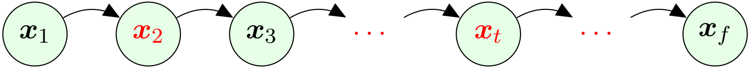
- Tensor representation (e.g., third-order trip volume data)



- Instead of: Temporal modeling for incomplete time series:



- Apply: Temporal modeling for latent temporal factors:



BTTF: Bayesian Temporal Tensor Factorization

Similar to BTMF, but on an incomplete tensor $\mathcal{Y} \in \mathbb{R}^{m \times n \times f}$

$$
\begin{aligned}
y_{ijt} &\sim \mathcal{N}\left(\sum_{s=1}^{r} u_{is} v_{js} x_{ts}, \tau^{-1}\right), (i, j, t) \in \Omega, \\
\boldsymbol{u}_i &\sim \mathcal{N}\left(\boldsymbol{\mu}_u, \Lambda_u^{-1}\right), i = 1, \ldots, m \\
\boldsymbol{v}_j &\sim \mathcal{N}\left(\boldsymbol{\mu}_v, \Lambda_v^{-1}\right), j = 1, \ldots, n, \\
\boldsymbol{x}_t &\sim \mathcal{N}\left(\tilde{\boldsymbol{x}}_t, \Lambda_x^{-1}\right), t = 1, 2, ..., f, \\
\tau &\sim \text{Gamma}(\alpha, \beta),
\end{aligned}
$$

where temporal factors are constrained by an auto-regressive (AR) model, i.e.,

$$
\tilde{\boldsymbol{x}}_t = \begin{cases} \mathbf{0}, & \text{if } t \in \{1, 2, ..., h_d\}, \\ \sum_{k=1}^{d} \boldsymbol{\theta}_k \circledast \boldsymbol{x}_{t-h_k}, & \text{otherwise.} \end{cases}
$$

Outline

Traffic speed data[3]

Data set:

- 214 road segments in Guangzhou, China;
- 10-minute time window;
- 61 days (from August 1 to September 30, 2016);
- original missing rate of this data set is 1.29%.

Experiment setting:

- Missing data imputation: random missing & non-random (fiber) missing.
- Rolling traffic forecasting: use historical data (controlling its missing rate) to predict traffic during next 5 days in a rolling manner.

[3] Publicly available: https://doi.org/10.5281/zenodo.1205228

Traffic speed data

Table: MAPE scores of the missing data imputation.

	random missing		non-random missing	
	20%	40%	20%	40%
BTMF	**0.0733**$_{80}$	**0.0750**$_{80}$	**0.1021**$_{10}$	0.1043$_{10}$
BPMF[4]	0.0946$_{80}$	0.0974$_{80}$	0.1027$_{10}$	0.1039$_{10}$
TRMF[5]	0.1024$_{10}$	0.1041$_{10}$	0.1034$_{10}$	0.1089$_{5}$
MF (ALS)	0.1024$_{10}$	0.1042$_{10}$	0.1035$_{10}$	0.1089$_{5}$
BGCP (ours)	0.0842$_{80}$	0.0833$_{80}$	**0.1021**$_{10}$	**0.1025**$_{10}$
HaLRTC[6]	0.0815	0.0887	0.1046	0.1088

[4] R. Salakhutdinov, A. Mnih, 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo, ICML.

[5] H.-F. Yu, N. Rao, I. S. Dhillon, 2016. Temporal regularized matrix factorization for high-dimensional time series prediction, NIPS.

[6] J. Liu, P. Musialski, P. Wonka, J. Ye, 2013. Tensor completion for estimating missing values in visual data. IEEE PAMI Trans., 35(1): 208-220.

Traffic speed data

Table: MAPE scores of the rolling traffic forecasting.

	–	random missing		non-random missing	
	0%	20%	40%	20%	40%
BTMF	0.0740_{80}	**0.0787_{80}**	**0.0827_{80}**	**0.1038_{20}**	**0.1139_{10}**
BPMF-AR	0.0899_{80}	0.0959_{80}	0.1013_{80}	0.1088_{20}	0.1174_{10}
BPMF-GP	0.0844_{80}	0.0999_{80}	0.1175_{80}	0.1045_{20}	0.1149_{10}
TRMF	0.0995_{80}	0.1149_{10}	0.1164_{10}	0.1149_{10}	0.1269_{5}
MF-AR	0.0884_{80}	0.1143_{10}	0.1163_{10}	0.1146_{10}	0.1266_{5}
MF-GP	0.0724_{80}	0.1110_{10}	0.1141_{10}	0.1109_{10}	0.1242_{5}
LSTM	0.0902	-	-	-	-
single-GP	0.1220	-	-	-	-

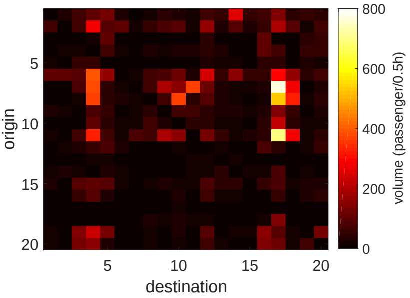
Transit demand data

Data set:

- 20 metro stations (most crowded) in Guangzhou, China;
- 30-minute time window;
- 62 days (from July 1 to August 31, 2017);

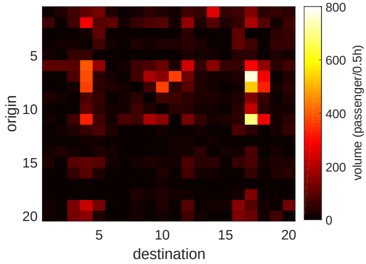
Experiment setting:

- Rolling traffic forecasting: use historical data to predict volume during the final week in a rolling manner.

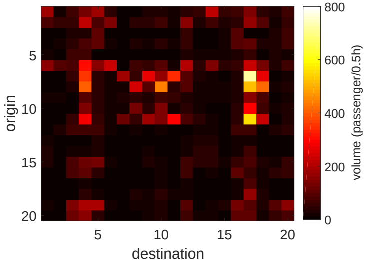


An OD volume matrix of the time window of 7:30-8:00 at August 25, 2017 (Friday).

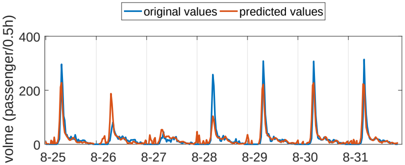
Transit demand data



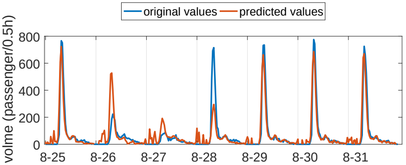
Original OD volume matrix.



Predicted OD volume matrix.



Volume curves of the #(6,17) OD pair.



Volume curves of the #(7,17) OD pair.

Outline

Contributions

- Develop a temporal modeling framework that enable us to achieve accurate prediction even for incomplete time series sequences, and on the other hand improve missing data imputation performance.
- Incorporate time series information among varying spatial locations.
- Apply temporal modeling in latent space rather than to incomplete observations directly.
- Characterize uncertainty and randomness in the real-world observations with a fully Bayesian treatment.

Future directions: tensor factorization + time series

- Long-range and short-range traffic forecasting with incomplete observations.
- Multi-resolution problems.
- Non-stationary time series data.
- Apply more powerful temporal models.
- Build generative models for heterogeneous data (e.g., no longer Gaussian distribution).

Future directions: tensor factorization + deep learning

Emerging models:

- Deep factorization model applying variational auto-encoder (see Deng et al., 2017[7]).
- Graph convolutional matrix completion (see Berg et al., 2018[8]).
- Integration with physical models (i.e., traffic flow models).

[7] Z. Deng, R. Navarathna, P. Carr, et al., 2017. Factorized Variational Autoencoders for Modeling Audience Reactions to Movies, CVPR.

[8] R. Berg, T. N. Kipf, M. Welling, 2018. Graph Convolutional Matrix Completion, KDD.

Questions?

Thank You!

For more information
`lijun.sun[at]mcgill.ca`
`http://smart.lab.mcgill.ca`