Low-Rank Autoregressive Tensor Completion for Spatiotemporal Traffic Data Imputation

Xinyu Chen
Polytechnique Montreal
Montreal, Quebec, Canada
chenxy346@gmail.com

Mengying Lei
McGill University
Montreal, Quebec, Canada
mengying.lei@mail.mcgill.ca

Nicolas Saunier
Polytechnique Montreal
Montreal, Quebec, Canada
nicolas.saunier@polymtl.ca

Lijun Sun
McGill University
Montreal, Quebec, Canada
lijun.sun@mcgill.ca

ABSTRACT

Spatiotemporal traffic time series (e.g., traffic volume/speed) collected from sensing systems are often incomplete with considerable corruption and large amounts of missing values, preventing users from harnessing the full power of the data. Missing data imputation has been a long-standing research topic and critical application for real-world intelligent transportation systems. A widely applied imputation method is low-rank matrix/tensor completion; however, the low-rank assumption only preserves the global structure while ignores the strong local consistency in spatiotemporal data. In this paper, we propose a low-rank autoregressive tensor completion (LATC) framework by introducing *temporal variation* as a new regularization term into the completion of a third-order (sensor × time of day × day) tensor. The third-order tensor structure allows us to better capture the global consistency of traffic data, such as the inherent seasonality and day-to-day similarity. To achieve local consistency, we design the temporal variation by imposing an autoregressive model for each time series with coefficients as learnable parameters. Different from previous spatial and temporal regularization schemes, the minimization of temporal variation can better characterize temporal generative mechanisms beyond local smoothness, allowing us to deal with more challenging scenarios such as “blackout" missing. To solve the optimization problem in LATC, we introduce an alternating minimization scheme that estimates the low-rank tensor and autoregressive coefficients iteratively. We conduct extensive numerical experiments on several real-world traffic data sets, and our results demonstrate the effectiveness of LATC in diverse missing scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Temporal reasoning**; **Regularization**.

KEYWORDS

Spatiotemporal traffic data, missing data imputation, low-rank tensor completion, truncated nuclear norm, autoregressive time series model

ACM Reference Format:
Xinyu Chen, Mengying Lei, Nicolas Saunier, and Lijun Sun. 2021. Low-Rank Autoregressive Tensor Completion for Spatiotemporal Traffic Data Imputation. In *MileTS '21: 7th KDD Workshop on Mining and Learning from Time Series, August 14th, 2021, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Spatiotemporal traffic data collected from various sensing systems (e.g. loop detectors and floating cars) serve as the foundation to a wide range of applications and decision-making processes in intelligent transportation systems. The emerging “big” data is often large-scale, high-dimensional, and incomplete, posing new challenges to modeling spatiotemporal traffic data. Missing data imputation is one of the most important research questions in spatiotemporal data analysis, since accurate and reliable imputation can help various downstream applications such as traffic forecasting and traffic control/management.

The key to missing data imputation is to efficiently characterize and leverage the complex correlations across both spatial and temporal dimensions [1]. Specifically, traffic state data (e.g., speed and flow) is individual sensor-based with a fixed temporal resolution. This allows us to summarize spatiotemporal traffic state data in the form of matrix (e.g., sensor × time) or tensor (e.g., sensor × time of day × day) [2], and low-rank matrix/tensor completion becomes a natural solution to solve the imputation problem. Over the past decade, extensive effort has been made on developing low-rank models through principle component analysis, matrix/tensor factorization (with predefined rank) and nuclear norm minimization (see e.g., [2, 4, 6]). However, the default low-rank structure (e.g., nuclear norm) purely relies on the algebraic property of the data, which is invariant to permutation in the spatial and temporal dimensions. In other words, with the low-rank assumption alone, we essentially overlook the strong “local” spatial and temporal consistency in the data. To this end, some recent studies have tried to encode such “local” consistency by introducing total/quadratic variation and graph regularization as a “smoothness” prior into low-rank factorization

MileTS '21, August 14th, 2021, Singapore

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

models [1, 7, 9, 10] and imposing time series dynamics on the temporal latent factor in the factorization framework [3, 8, 11]. However, these studies essentially adopt a bilinear/multilinear factorization model, which requires a predefined rank as a hyperparameter.

In this paper, we propose a low-rank autoregressive tensor completion (LATC) framework to impute missing values in spatiotemporal traffic data. For each completed time series, we define temporal variation as the accumulated sum of autoregressive errors. To model the low-rankness property, we use truncated nuclear norm [5] as an effective approximation to avoid the rank determination problem in factorization models. The final objective function of LATC consists of two components, i.e., the truncated nuclear norm of the completed tensor and the temporal variation defined on the unfolded time series matrix. The combination allows us to effectively characterize both global patterns and local consistency in spatiotemporal traffic data. The overall contribution of this work is twofold:

1) We integrate the autoregressive time series process into a low-rank tensor completion model to capture both global and local trends in spatiotemporal traffic data.
2) We conduct extensive numerical experiments on four traffic data sets. Imputation results show the superiority and advantage of LATC over recent state-of-the-art models.

2 METHODOLOGY

To ensure both global consistency and local consistency, LATC framework takes into account both low-rank tensor completion and autoregressive process. For any partially observed data matrix $Y\in\mathbb{R}^{M\times(IJ)}$ consisting of M spatial sensors, I time points per day, and J days in spatiotemporal setting, the minimization problem can be formulated as follows,

$$\begin{aligned}\min_{\boldsymbol{\mathcal{X}},Z,A}\ & \|\boldsymbol{\mathcal{X}}\|_{r,*}+\frac{\lambda}{2}\|Z\|_{A,\mathcal{H}}\\ \text{s.t. } & \begin{cases}\boldsymbol{\mathcal{X}}=\mathcal{Q}(Z),\\ \mathcal{P}_{\Omega}(Z)=\mathcal{P}_{\Omega}(Y),\end{cases}\end{aligned}\tag{1}$$

where $\mathcal{P}_{\Omega}(\cdot)$ denotes the orthogonal projection supported on the observed set Ω, which holds the following definition for the element $y_{m,n},\forall(m,n)$ from Y:

$$[\mathcal{P}_{\Omega}(Y)]_{m,n}=\begin{cases}y_{m,n}, & \text{if } (m,n)\in\Omega,\\ 0, & \text{otherwise.}\end{cases}$$

In this model, by defining the forward tensorization operator $\mathcal{Q}(\cdot)$, it is possible to generate a third-order tensor as $\boldsymbol{\mathcal{X}}=\mathcal{Q}(Z)\in\mathbb{R}^{M\times I\times J}$. In contrast, the resulted tensor can also be converted into the original matrix by $Z=\mathcal{Q}^{-1}(\boldsymbol{\mathcal{X}})\in\mathbb{R}^{M\times(IJ)}$ where $\mathcal{Q}^{-1}(\cdot)$ denotes the inverse operator of $\mathcal{Q}(\cdot)$. Relying on these definitions, there are two critical components in the objective:

- Truncated nuclear norm $\|\boldsymbol{\mathcal{X}}\|_{r,*}$ (r denotes the integer-wise truncation) on the tensor $\boldsymbol{\mathcal{X}}$ serves as the fundamentals of the whole model for capturing global low-rank patterns and imputing missing values, and it takes the form:

$$\|\boldsymbol{\mathcal{X}}\|_{r,*}=\sum_{h=1}^{3}\alpha_h\|\boldsymbol{\mathcal{X}}_{(h)}\|_{r,*}\tag{2}$$

for $\boldsymbol{\mathcal{X}}\in\mathbb{R}^{M\times I\times J}$ with $\sum_{h=1}^{3}\alpha_h=1$.

- Temporal variation of a time series matrix Z with a coefficient matrix $A\in\mathbb{R}^{M\times d}$ and a time lag set $\mathcal{H}=\{h_1,\cdots,h_d\}$ is defined as:

$$\|Z\|_{A,\mathcal{H}}=\sum_{m,t}(z_{m,t}-\sum_{i}a_{m,i}z_{m,t-h_i})^2,\tag{3}$$

which quantifies the total squared errors when fitting each time series $\boldsymbol{z}_m\in\mathbb{R}^{IJ}$ through an individual aoturegressive model with coefficient $\boldsymbol{a}_m\in\mathbb{R}^{d}$. In the objective, λ is a weight parameter which controls the trade-off between truncated nuclear norm and temporal variation.

As mentioned above, the formulation of LATC can ensure both global consistency and local consistency by combining truncated nuclear norm minimization with temporal variation minimization. Fig. 1 shows that Y can be reconstructed with both low-rank patterns and time series dynamics because the constraint in the optimization problem, i.e., $\boldsymbol{\mathcal{X}}=\mathcal{Q}(Z)$, is closely related to the partially observed matrix Y.

3 EXPERIMENTS

In this section, we evaluate the proposed LATC model on several real-world traffic data sets with different missing patterns.

3.1 Traffic Data Sets

We use the following spatiotemporal traffic sets for our experiments.

- **(G)**: Guangzhou urban traffic speed data set.[1] This data set contains traffic speed collected from 214 road segments over two months (from August 1 to September 30, 2016) with a 10-minute resolution (i.e., 144 time intervals per day) in Guangzhou, China. The prepared data is of size 214 × 8784 in the form of multivariate time series matrix (or tensor of size 214 × 144 × 61).
- **(H)**: Hangzhou metro passenger flow data set.[2] This data set provides incoming passenger flow of 80 metro stations over 25 days (from January 1 to January 25, 2019) with a 10-minute resolution in Hangzhou, China. We discard the interval 0:00 a.m. 6:00 a.m. with no services, and only consider the remaining 108 time intervals of a day. The prepared data is of size 80 × 2700 in the form of multivariate time series (or tensor of size 80 × 108 × 25).
- **(S)**: Seattle freeway traffic speed data set.[3] This data set contains freeway traffic speed from 323 loop detectors with a 5-minute resolution (i.e., 288 time intervals per day) over the first four weeks of January, 2015 in Seattle, USA. The prepared data is of size 323 × 8064 in the form of multivariate time series (or tensor of size 323 × 288 × 28).
- **(P)**: Portland highway traffic volume data set.[4] This data set is collected from highways in the Portland-Vancouver Metropolitan region, which contains traffic volume from 1156 loop detectors with a 15-minute resolution (i.e., 96 time intervals per day) in January, 2021. The prepared data is of size 1156 × 2976 in the form of multivariate time series matrix (or tensor of size 1156 × 96 × 31).

[1] https://doi.org/10.5281/zenodo.1205229
[2] https://tianchi.aliyun.com/competition/entrance/231708/information
[3] https://github.com/zhiyongc/Seattle-Loop-Data
[4] https://portal.its.pdx.edu/home

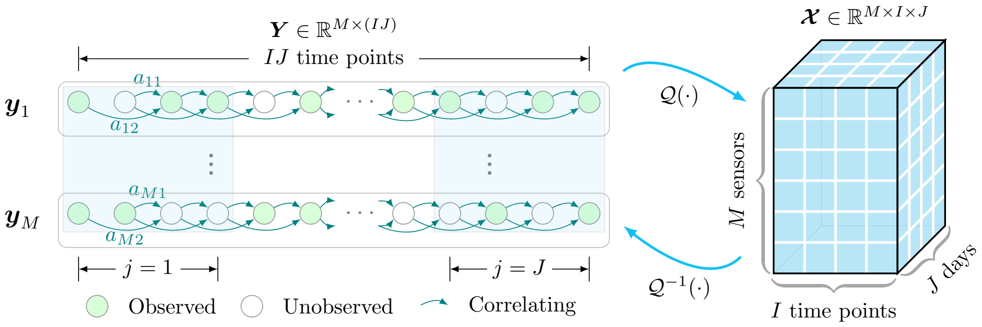


Figure 1: Illustration of the proposed LATC framework for spatiotemporal traffic data imputation with time lags $\mathcal{H} = \{1, 2\}$. Each time series $\boldsymbol{y}_m, \forall m \in \{1, 2, \ldots, M\}$ is modeled by the autoregressive coefficients $\{a_{m1}, a_{m2}\}$.

3.2 Missing Data Generation

In this work, we take into account three missing data patterns as shown in Fig. 2, i.e., random missing (RM), non-random missing (NM), and blackout missing (BM). RM and NM data are generated by referring to the existing work [2]. According to the mechanism of RM and NM data, we mask certain amount of observations as missing values (e.g., 30%, 70%, 90%), and the remaining partial observations are input data for learning a well-behaved model. BM pattern is different from RM and NM patterns, which masks observations of all spatial sensors/locations as missing values with certain window length. BM is a challenging scenario with complete column-wise missing where we set the missing rate in our experiments to 30%.

To assess the imputation performance, we use the actual values of the masked missing entries as the ground truth to compute MAPE and RMSE:

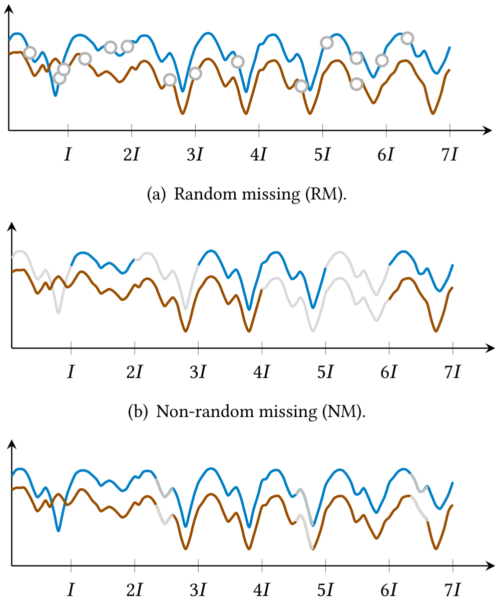
$$\begin{aligned} \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100, \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \end{aligned} \tag{4}$$

where y_i and $\hat{y}_i$ are actual values and imputed values, respectively.

3.3 Baseline Models

For comparison, we take into account the following baseline:

- Low-Rank Tensor Completion with Truncation Nuclear Norm minimization (LRTC-TNN, [4]). This is a low-rank completion model in which truncated nuclear norm minimization can help maintain the most important low-rank patterns.
- Bayesian Temporal Matrix Factorization (BTMF, [3]). This is a fully Bayesian temporal factorization framework which builds the correlation of temporal dynamics on latent factors by vector autoregressive process.
- Smooth PARAFAC Tensor Completion (SPC, [10]). This is a tensor decomposition based completion model with total variation smoothness constraints.



(a) Random missing (RM).

(b) Non-random missing (NM).

(c) Blackout missing (BM).

Figure 2: Illustration of three missing data patterns for spatiotemporal traffic data. In these graphics, two curves correspond to two different time series. (a) Data are missing at random. Small circles indicate the missing values. (b) Data are missing continuously during a few time periods. Segments in gray indicate missing values. (c) No sensors are available (i.e., blackout) over a certain time window.

3.4 Results

In Table 1, despite the truncated nuclear norm built on tensor, the results also show the advantage of temporal variation built on the multivariate time series matrix. Due to the temporal modeling, temporal variation can improve the imputation performance for missing traffic data imputation. Table 1 shows the overall imputation performance of LATC and baseline models on the four selected traffic data sets with various missing scenarios. Of these results, NM and BM data seem to be more difficult to reconstruct with all these imputation models than RM data. In most cases, LATC outperforms other baseline models. Comparing LATC with LRTC-TNN shows the advantage of temporal variation, i.e., temporal modeling with autoregressive process has positive influence for improving the imputation performance. For volume data sets (H) and (P), the relative errors are quite high because some volume values are close to 0 or relatively small and estimating these values would accumulate relatively large relative errors.

Figs. 3, 4, and 5 show some imputation examples with different missing scenarios that achieved by LATC. In these examples, we can see explicit temporal dependencies underlying traffic time series data. For all missing scenarios, LATC can achieve accurate imputation and learn the true signals from observations even with severe missing data (e.g., NM/BM data). In Fig. 3, it shows that the time series signal of passenger flow is not complex. By referring to Table 1, we can see that LRTC-TNN without temporal variation outperforms the proposed LATC model on Hangzhou metro passenger flow data, and this demonstrates that not all multivariate time series imputation cases require temporal modeling, for some cases that the signal does not show strong temporal dependencies, purely low-rank model can also provide accurate imputation.

Table 1: Performance comparison (in MAPE/RMSE) of LATC and baseline models for RM, NM, and BM data imputation. The number next to the BM denotes the window length.

	Missing	LATC	LRTC-TNN	BTMF	SPC
(G)	30%, RM	**5.71**/**2.54**	6.99/3.00	7.54/3.27	7.37/5.06
	70%, RM	**7.22**/**3.18**	8.38/3.59	8.75/3.73	8.91/4.44
	90%, RM	**9.11**/**3.86**	9.55/4.05	10.02/4.21	10.60/4.85
	30%, NM	9.63/4.09	9.61/**4.07**	10.32/4.33	**9.13**/5.29
	70%, NM	10.37/4.35	**10.36**/**4.34**	11.36/4.85	11.15/5.17
	30%, BM-6	**9.23**/**3.91**	9.45/3.97	12.43/7.04	11.14/5.13
(H)	30%, RM	19.12/24.97	**18.87**/**24.90**	22.37/28.66	19.82/26.21
	70%, RM	20.25/28.25	**20.07**/**28.13**	25.65/32.23	21.02/31.91
	90%, RM	24.32/**34.44**	**23.46**/35.84	31.51/46.24	24.97/49.68
	30%, NM	**19.93**/**47.38**	19.94/50.12	25.61/77.00	27.46/68.56
	70%, NM	24.30/47.30	**23.88**/**45.06**	34.50/70.11	46.86/98.81
	30%, BM-6	21.93/28.64	**21.40**/**27.83**	52.15/57.61	22.49/37.53
(S)	30%, RM	**4.90**/**3.16**	4.99/3.20	5.91/3.72	5.92/3.62
	70%, RM	**5.96**/**3.71**	6.10/3.77	6.47/3.98	7.38/4.30
	90%, RM	**7.47**/**4.51**	8.08/4.80	8.17/4.81	9.75/5.31
	30%, NM	7.11/4.33	**6.85**/**4.21**	9.26/5.36	8.87/4.99
	70%, NM	9.46/5.42	**9.23**/**5.35**	10.47/6.15	11.32/5.92
	30%, BM-12	**9.44**/**5.36**	9.52/5.41	14.33/13.60	11.30/5.84
(P)	30%, RM	17.46/**15.89**	**17.27**/16.08	18.22/19.14	21.29/56.73
	70%, RM	**19.56**/**18.70**	19.99/18.73	19.96/22.21	24.35/43.32
	90%, RM	23.47/22.74	**22.90**/**22.68**	23.90/25.71	28.45/39.65
	30%, NM	**18.90**/**18.84**	19.59/18.91	19.55/20.38	26.96/60.33
	70%, NM	24.67/31.74	30.26/60.85	**23.86**/**26.74**	33.42/47.34
	30%, BM-4	**24.04**/**23.52**	31.74/74.42	27.85/25.68	31.01/60.33

Best results are highlighted in bold fonts.

4 CONCLUSION

Spatiotemporal traffic data imputation is of great significance in data-driven intelligent transportation systems. Fortunately, for analyzing and modeling traffic data, there are some fundamental features such as low-rank properties and temporal dynamics that can be taken into account. In this work, the proposed LATC model builds both low-rank structure (i.e., truncated nuclear norm) and time series autoregressive process on certain data representations. By doing so, numerical experiments on some real-world traffic data sets show the advantages of LATC over other low-rank models.

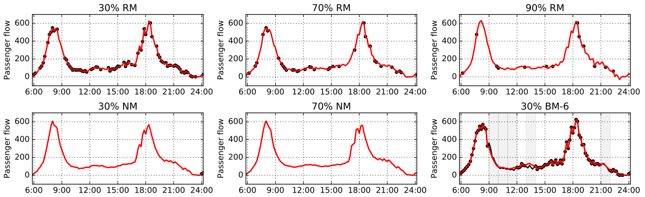


Figure 3: Imputed values by LATC for Hangzhou metro passenger flow data. This example corresponds to metro station #3 and the 4th day of the data set. Black dots/curves indicate the partially observed data, gray rectangles indicate blackout missing, while red curves indicate the imputed values.

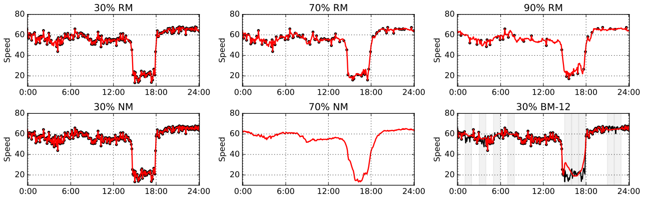


Figure 4: Imputed values by LATC for Seattle freeway traffic speed data. This example corresponds to detector #3 and the 7th day of the data set.

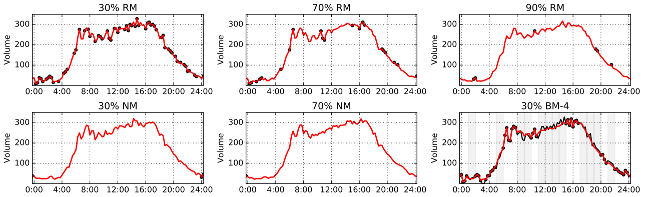


Figure 5: Imputed values by LATC for Portland traffic volume data. This example corresponds to detector #3 and the 8th day of the data set.

ACKNOWLEDGMENTS

This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Fonds de recherche du Quebec – Nature et technologies (FRQNT), and the Canada Foundation for Innovation (CFI). X. Chen and M. Lei would like to thank the Institute for Data Valorisation (IVADO) for providing the PhD Excellence Scholarship to support this study.

REFERENCES

[1] Mohammad Taha Bahadori, Qi Rose Yu, and Yan Liu. 2014. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in Neural Information Processing Systems*. 3491–3499.

[2] Xinyu Chen, Zhaocheng He, and Lijun Sun. 2019. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transportation Research Part C: Emerging Technologies* 98 (2019), 73 – 84. https://doi.org/10.1016/j.trc.2018.11.003

[3] X. Chen and L. Sun. 2021. Bayesian Temporal Factorization for Multidimensional Time Series Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3066551

[4] Xinyu Chen, Jinming Yang, and Lijun Sun. 2020. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. *Transportation Research Part C: Emerging Technologies* 117 (2020), 102673.

[5] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. 2013. Fast and Accurate Matrix Completion via Truncated Nuclear Norm Regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 9 (2013), 2117–2130.

[6] Lei Li, James McCann, Nancy S Pollard, and Christos Faloutsos. 2009. Dynammo: Mining and summarization of coevolving sequences with missing values. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 507–516.

[7] Nikhil Rao, Hsiang-Fu Yu, Pradeep Ravikumar, and Inderjit S Dhillon. 2015. Collaborative Filtering with Graph Information: Consistency and Scalable Methods.. In *NIPS*, Vol. 2. Citeseer, 7.

[8] Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. 2019. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems*. 4838–4847.

[9] Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, and Jaime G Carbonell. 2010. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *SIAM International Conference on Data Mining*. 211–222.

[10] Tatsuya Yokota, Qibin Zhao, and Andrzej Cichocki. 2016. Smooth PARAFAC decomposition for tensor completion. *IEEE Transactions on Signal Processing* 64, 20 (2016), 5423–5436.

[11] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. 2016. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in Neural Information Processing Systems*. 847–855.