Variational Inference for Extreme Spatio-Temporal Matrix Completion

Charul Paliwal
IIITD, India
charuli@iiitd.ac.in

Pravesh Biyani
IIITD, India
praveshb@iiitd.ac.in

Abstract

Missing data is a common problem in real-world sensor data collection. The performance of various approaches to impute data degrade rapidly in the extreme scenarios of low data sampling and noisy sampling, a case present in many real-world problems in the field of traffic sensing and environment monitoring, etc. However, jointly exploiting the spatiotemporal and periodic structure, which is generally not captured by classical matrix completion approaches, can improve the imputation performance of sensor data in such real-world conditions. We present a Bayesian approach toward spatiotemporal matrix completion wherein we estimate the underlying temporarily varying subspace using a Variational Bayesian technique. We jointly couple the low-rank matrix completion with the state space autoregressive framework along with a penalty function on the slowly varying subspace to model the temporal and periodic evolution in the data. We also propose a robust version of the above formulation, which improves the performance of imputation in the presence of outliers. Results demonstrate that the proposed method outperforms the recent state-of-the-art methods for real-world traffic and air pollution data. We demonstrate that fusing the subspace evolution over days can improve the imputation performance with even 15% of the data sampling.

1 Introduction

Data acquired from both static and moving sensors contains missing data due to sensor malfunction, irregularity in sensor measurements, etc. Additionally, moving sensors scheme uses relatively fewer sensors resulting in high data "gaps" in both spatial and temporal dimensions [1]. This motivates the problem of extreme matrix completion, where the percentage of data sampled may be as low as 15%. Thereby, a natural question to ask is: how to fill the high percentage missing entries within a reasonable error range? Can we leverage additional periodic information in the matrix completion framework to estimate the high percentage of missing entries. Also, in addition to the missing entries, sensor measurements can be contaminated with outliers emerging from the sensor malfunctioning, communication errors, or impulse noises. The occurrence of outliers in the measurements can further degrade the performance of data imputation. However, unlike the missing entries, the location and the value of outliers are unknown, which makes the problem more challenging. Therefore, the other question to ask is: how to estimate the missing data while detecting the noisy outliers? The answer to the above questions lies in exploiting the underlying structure available in the data. For instance, both the air pollution [2] and the traffic [3] data exhibit joint spatial and temporal correlation as well as periodicity in daily patterns, thereby generating redundancy that can be potentially exploited by performing an intelligent spatio-temporal extrapolation.

In this paper, we propose a Variational Bayesian Filtering with Subspace Information (VBFSI) to estimate the data in the case of extremely sparsely sampled data. We observe that even with a fraction of sampled observed data, we can estimate the remaining measurements with reasonable

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

accuracy. This work exploits the spatiotemporal and periodic pattern in the measurements and perform extreme matrix completion with reasonable accuracy. First, we enforce a low-rank structure to the spatiotemporal data. Second, we enforce the state-space model on the temporal embeddings to capture the temporal evolution in the data. Third, we enforce the subspace estimate of the matrix to be close to previously learned subspace distribution using the Mahalanobis distance. Exploiting the prior subspace over days results in a considerable reduction of the number of measurements required to estimate the matrix [4, 5] thereby boosting the performance in case of a low sampling rate. We use the Variational Bayes approach to update the parameters in an iterative fashion. In our work, the subspace distribution is chosen to allow automatic relevance determination, and unlike the matrix or tensor completion methods, the algorithm parameters such as rank, noise powers need not be specified or tuned. We compare the performance of the proposed algorithm with state-of-the-art algorithms on many real-world traffic and air quality datasets. The result shows that modeling the subspace evolution leads to improvement in performance even when a small percentage of random measurements are available for the purpose of imputation. A likely impact of our method is that cities with a low sensing budget can perform random drive-by sampling of the urban environment, and the suggested matrix completion framework can provide a reasonably accurate imputation leading to better decision making.

1.1 Related work

The traditional matrix completion framework is not applicable for time series data imputation, as it does not take into account the ordering among the temporal embeddings [6]. Autoregressive model can model temporal embeddings to capture the temporal evolution in the time series data [6–8]. However, these models fail to capture the prior subspace information that can be exploited to capture periodicity. Exploiting the 3-way pattern in the data using tensor completion-based frameworks can improve the imputation performance [9–11]. However, the temporal evolution and subspace evolution is not modeled in the traditional tensor completion frameworks. We propose to incorporate the periodicity in the matrix completion framework while modeling the temporal and subspace evolution over days. Variational Bayesian approaches are proposed for matrix/tensor completion and robust PCA [9, 10, 12–14]. A state-space model to capture the temporal evolution is also proposed in [7, 8]. However, these approaches do not explicitly model the evolution of the subspace to capture the periodicity in the data. Also, the proposed matrix and tensor completion methods are not evaluated for extreme matrix completion and extreme matrix completion in the presence of outliers.

2 Variational Bayesian Filtering with Subspace Information (VBFSI)

Let $\mathbf{X} \in \mathbb{R}^{n \times t}$ be the data matrix for a day d, where n and t denote the number of spatial locations and time stamps respectively. The low rankness in the data can be imposed as

$$\mathcal{L}_1 = \min_{\mathbf{U},\mathbf{V}} ||\mathbf{P}_\Omega \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^T)||_F \tag{1}$$

where $\mathbf{U} \in \mathcal{R}^{n \times r}$ and $\mathbf{V} \in \mathcal{R}^{t \times r}$ and r =rank($\mathbf{X}$) $<< min(n, t)$. For sampling percentage of p, let Ω denote the sampled data containing $p \times n \times t$ samples. $\mathbf{P}_\Omega$ is the indicator matrix where $P_{ij} = 1 \ \forall (i, j) \in \Omega$. Temporal evolution is captured by regularizing the columns of $\mathbf{V}$ to follow an autoregressive model and regularize $\mathbf{U}$ to capture the periodicity over days.

$$\mathcal{R}(\mathbf{V}) = \sum_{i=1}^{t} ||\mathbf{v}_i - \mathbf{F}\mathbf{v}_{i-1}||, \ \mathcal{R}(\mathbf{U}) = \eta \sum_{i=1}^{n} (\mathbf{u}_i - \mathbf{u}_i^{d-1})^T (\mathbf{\Xi}_i^{\mathbf{U}^{d-1}})^{-1} (\mathbf{u}_i - \mathbf{u}_i^{d-1}) \tag{2}$$

here $\mathcal{R}(\mathbf{U})$ corresponds to the Mahalanobis distance between each row vector of $\mathbf{U}^{d-1}$ (subspace estimate of previous day) and $\mathbf{U}$ (current subspace estimate for a given day d). $\mathbf{\Xi}^{\mathbf{U}^{d-1}}$ denotes the covariance matrix of $\mathbf{U}^{d-1}$. Here η controls the effect of prior subspace ($\mathbf{U}^{d-1}$, $\mathbf{\Xi}_i^{d-1}$) in the estimation of $\mathbf{U}$. $\mathbf{u}_{i\cdot}$ denotes the i-th row of $\mathbf{U}$. The entries of $\mathbf{X}_\Omega$ can be defined as:

$$\boldsymbol{p}(\mathbf{X}_\Omega \mid \mathbf{U}, \mathbf{V}, \beta) = \prod_{(i,j) \in \Omega} \mathcal{N}(X_{ij} \mid \mathbf{u}_{i\cdot} \mathbf{v}_{j\cdot}^T, \beta^{-1}), \ \boldsymbol{p}(\beta) = \beta^{-1} \tag{3}$$

here β is the noise precision. The prior on the noise is assumed to be non-informative Jeffrey's prior.

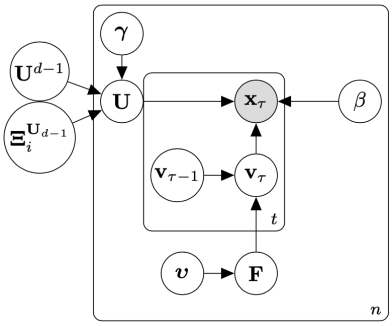
Columns of $\mathbf{U}$ are enforced with a sparsity profile using precision γ_i to automate the rank.

$$\boldsymbol{p}(\mathbf{U} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{r} \mathcal{N}(\mathbf{u}_i \mid \mathbf{u}_i^{d-1}, \gamma_i^{-1}\mathbf{I}_n) \tag{4}$$

When γ_i are driven to a large value then the column mean will be $\mathbf{u}_i^{d-1}$, and we prune these columns and in turn, reducing the rank thereby modelling the low rank in the Bayesian framework, referred as Automatic Rank Determination [12]. Further, the autoregressive regularization can be modelled as

$$\boldsymbol{p}(\mathbf{V} \mid \mathbf{F}) = \mathcal{N}(\mathbf{v}_1; \boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1) \prod_{\tau=2}^{t} \mathcal{N}(\mathbf{v}_\tau \mid \mathbf{F}\mathbf{v}_{\tau-1}, \mathbf{I}_r) \tag{5}$$

Figure 1: VBFSI model



$\mathbf{F}$ is assigned multivariate Gaussian priors with column-specific precisions υ. Precision variables $\boldsymbol{\gamma}$ and $\boldsymbol{\upsilon}$ are selected to have non-informative Jeffrey's priors

$$\boldsymbol{p}(\mathbf{F} \mid \boldsymbol{\upsilon}) = \prod_{i=1}^{r} \mathcal{N}(\mathbf{f}_i \mid 0, \upsilon_i^{-1}\mathbf{I}_r),\ \boldsymbol{p}(\gamma_i) = \frac{1}{\gamma_i}, \qquad \boldsymbol{p}(\upsilon_i) = \frac{1}{\upsilon_i} \tag{6}$$

The overall joint distribution for spatio-temporal matrix completion can be expressed as

$$\boldsymbol{p}(\mathbf{X}_\Omega, \mathbf{U}, \mathbf{V}, \mathbf{F}, \beta, \boldsymbol{\gamma}, \boldsymbol{\upsilon}) = \boldsymbol{p}(\mathbf{X}_\Omega|\mathbf{U}, \mathbf{V}, \beta)\boldsymbol{p}(\mathbf{U}|\boldsymbol{\gamma})\boldsymbol{p}(\mathbf{V}|\mathbf{F})\boldsymbol{p}(\mathbf{F}|\boldsymbol{\upsilon})\boldsymbol{p}(\beta)\boldsymbol{p}(\boldsymbol{\upsilon})\boldsymbol{p}(\boldsymbol{\gamma}) \tag{7}$$

The Full Bayesian graphical model for spatio-temporal matrix completion is shown in Fig 1. The posterior distribution $p(\mathcal{H}|\mathbf{x}_\Omega)$ is intractable as $p(\mathbf{x}_\Omega)$ needs to be marginalized over all parameters ($\mathcal{H}$). Therefore, we utilize the mean-field approximation and provide the Variational Bayesian update steps for all the parameters in the Appendix.

3 Robust VBFSI (RVBFSI)

RVBFSI estimates the missing data while detecting the noisy outliers. For robust matrix completion, we model $\mathbf{X} = \mathbf{U}\mathbf{V}^T + \mathbf{E} + \mathbf{N}$, where $\mathbf{E}$ denotes the sparse outlier matrix and $\mathbf{N}$ is the dense error matrix. The low rankness defined in the Eq. 1 is modified as

$$\mathcal{L}_1 = \min_{\mathbf{U},\mathbf{V}} ||\mathbf{P}_\Omega \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{E})||_F \tag{8}$$

The regularization on $\mathbf{U}$ and $\mathbf{V}$ follows the Eq. 2. Each entry of sparse outlier matrix E_{ij} is assigned a precision α_{ij}.where the α_{ij} have the non informative prior.This works similar to the ARD where instead of a column of the matrix, each entry of the matrix is assigned with precision. Whenever $\alpha_{i,j}$ is driven to a large value $E_{i,j} \to 0$ thereby enforcing sparsity.

$$\boldsymbol{p}(\mathbf{E}|\boldsymbol{\alpha}) = \prod_{i|((i,j)\in\Omega)} \prod_{j|((i,j)\in\Omega)} \mathcal{N}(E_{ij} \mid 0, \alpha_{ij}^{-1}), \qquad \boldsymbol{p}(\alpha_{ij}) = {\alpha_{ij}}^{-1} \tag{9}$$

4 Experimentation

Dataset and Parameters : We use traffic speed and air quality (PM 2.5) data for performance evaluation. Data (DT): Delhi traffic speed data [8]. Data (GT): Guangzhou urban traffic speed data [15]. Data (PT): Pems traffic speed data [16]. Data (CA): China Air Quality data [17]. Rank determination is automatic. We only tune the parameter η. We grid search the best η for different sampling percentages. Then, we fit exponential model for η vs. p. The initial subspace $\mathbf{U}^0$ is calculated using the eight days average, then we run the algorithm in an online fashion for the next 30 days.

Baseline Algorithms: We compare VBFSI with state of the art matrix and tensor imputation methods. **VBSF** [8], **VMC** [18], **TRMF** [6], **BTMF** [19], **BCPF** [10] and **TRLRF** [20]. We compare RVBFSI with the following Robust imputation methods. **RVBSF** [8], **RegL**$_1$ [21], **BRTF** [14].

	p %	VBFSI	VBSF	VMC	BCPF	TRLRF	TRMF	BTMF
Data:DT	5%	**0.156 / 4.387**	0.782 / 22.03	0.998 / 28.18	0.164 / 4.6	0.901 / 25.45	0.183 / 5.146	0.157 / 4.394
	15%	**0.135 / 3.796**	0.162 / 4.552	0.97 / 27.39	0.147 / 4.137	0.682 / 19.25	0.151 / 4.24	0.137 / 3.836
	50%	0.117 / 3.289	0.119 / 3.344	0.131 / 3.687	**0.115 / 3.23**	0.171 / 4.785	0.121 / 3.409	0.119 / 3.342
	75%	0.11 / 3.086	0.11 / 3.099	0.117 / 3.28	**0.109 / 3.076**	0.13 / 3.642	0.117 / 3.262	0.115 / 3.224
Data:PT	5%	**0.144 / 8.608**	1 / 60.08	0.998 / 59.974	0.175 / 10.494	0.94 / 56.47	0.161 / 9.571	0.151 / 9.084
	15%	**0.111 / 6.625**	0.179 / 10.7	0.974 / 58.52	0.147 / 8.836	0.807 / 48.49	0.139 / 8.264	0.118 / 7.06
	50%	**0.084 / 5.056**	0.097 / 5.79	0.087 / 5.213	0.091 / 5.431	0.168 / 10.07	0.093 / 5.511	0.1 / 6.02
	75%	0.081 / 4.841	0.081 / 4.854	**0.069 / 4.135**	0.081 / 4.848	0.097 / 5.833	0.083 / 4.951	0.097 / 5.81
Data:GT	5%	0.159 / 6.384	1 / 40.31	0.993 / 40.03	0.158 / 6.346	0.863 / 34.76	0.184 / 6.614	**0.131 / 5.244**
	15%	0.121 / 4.854	0.148 / 5.91	0.382 / 15.33	0.138 / 5.541	0.492 / 19.8	0.162 / 5.845	**0.11 / 4.43**
	50%	**0.088 / 3.547**	0.112 / 4.501	0.09 / 3.616	0.097 / 3.902	0.112 / 4.503	0.128 / 4.624	0.095 / 3.801
	75%	**0.079 / 3.189**	0.1 / 4.027	0.081 / 3.247	0.088 / 3.515	0.091 / 3.652	0.12 / 4.303	0.092 / 3.712
Data:CA	5%	0.439 / 32.44	1 / 76.562	0.998 / 76.464	0.435 / 32.672	0.978 / 74.949	0.434 / 33.915	**0.414 / 31.431**
	15%	0.35 / 25.964	0.396 / 29.762	0.986 / 75.582	**0.341 / 25.471**	0.936 / 72.046	0.369 / 28.735	0.344 / 25.94
	50%	0.222 / 16.466	0.23 / 17.172	**0.213 / 15.892**	0.237 / 17.646	0.731 / 57.059	0.235 / 18.274	0.248 / 18.483
	75%	0.198 / 14.648	0.2 / 14.675	**0.171 / 12.636**	0.209 / 15.603	0.472 / 36.816	0.197 / 15.249	0.223 / 16.556

Table 1: MRE/RMSE scores for data imputation. The best two results are bold and underlined.

	o=5%				o=10%			
p %	10%	25%	50%	75%	10%	25%	50%	75%
RVBFSI	**0.167 / 4.672**	**0.14 / 3.91**	**0.126 / 3.544**	**0.119 / 3.337**	**0.17 / 4.78**	**0.14 / 3.925**	**0.128 / 3.573**	**0.118 / 3.314**
RVBSF	0.196 / 5.527	0.154 / 4.313	0.132 / 3.696	0.124 / 3.485	0.227 / 6.403	0.16 / 4.492	0.132 / 3.728	0.124 / 3.487
VBFSI	0.188 / 5.277	0.177 / 4.972	0.17 / 4.778	0.158 / 4.427	0.208 / 5.86	0.194 / 5.443	0.184 / 5.173	0.175 / 4.899
VBSF	0.305 / 8.583	0.201 / 5.656	0.177 / 4.98	0.166 / 4.65	0.391 / 11.01	0.237 / 6.65	0.198 / 5.561	0.184 / 5.16
VMC	0.995 / 28.05	0.798 / 22.4	0.368 / 10.35	0.306 / 8.605	0.996 / 28.08	0.937 / 26.41	0.562 / 15.8	0.44 / 12.38
BCPF	0.193 / 5.422	0.164 / 4.597	0.146 / 4.101	0.14 / 3.927	0.205 / 5.763	0.179 / 5.017	0.155 / 4.363	0.148 / 4.146
TRLRF	0.973 / 27.45	0.956 / 26.95	0.934 / 26.34	1.144 / 32.24	0.977 / 27.55	0.969 / 27.33	0.997 / 28.1	1.447 / 40.71
TRMF	0.382 / 10.76	0.419 / 11.78	0.265 / 7.426	0.217 / 6.05	0.565 / 15.91	0.56 / 15.75	0.338 / 9.492	0.291 / 8.139
BTMF	0.226 / 5.828	0.218 / 5.604	0.221 / 5.615	0.218 / 5.591	0.308 / 7.79	0.303 / 7.681	0.304 / 7.794	0.303 / 7.764
RegL_1	0.521 / 14.66	0.489 / 13.78	0.192 / 5.429	0.132 / 3.718	0.699 / 19.69	0.498 / 14.01	0.242 / 6.825	0.155 / 4.387
BRTF	0.243 / 6.829	0.232 / 6.522	0.138 / 3.888	0.131 / 3.706	0.218 / 6.133	0.2 / 5.609	0.152 / 4.285	0.145 / 4.121

Table 2: MRE/RMSE for outlier corrupted data (DT), outlier percentage (o) is 5% and 10%.

Performance Comparison The performance comparison of VBFSI with the current state of the art methods is shown in Table 1. VBFSI outperforms VBSF for all the datasets. The performance of VBSF is comparable to VBFSI for higher sampling. While for lower sampling, the performance of VBSF degrades. Incorporating even the noisy prior subspace information in the architecture can reduce the sampling complexity of the matrix by logarithmic factor [4]. Therefore, for low sampling VBFSI performs better than VBSF since we have incorporate the prior information using η. VMC experience a similar trend, where the performance is comparable for higher sampling and degrades for low sampling. VBFSI outperforms VMC for almost all the cases for traffic data (DT, PT, GT). However, for the air quality data (CA), the performance of VMC is better than VBFSI for a higher sampling percentage because VMC can capture the nonlinearity in the data. For low sampling percentage, VBFSI performance is comparable to BTMF in most of the cases. However, for higher sampling percentage, VBFSI outperforms BTMF. One of the disadvantage of BTMF and TRMF is that rank is not tuned automatically and uses gibbs sampling which is relatively slower than the Variational Bayesian approach for parameter estimation [22].

To compare the performance of VBFSI and RVBFSI in the case of outliers, we artificially add the outliers in the Data: DT. We randomly add 5% and 10% of the outliers in the total sampled data, i.e., the number of outliers is $o \times p$ fraction of the overall data. The entries corrupted with outliers are uniformly distributed between $[-\sigma, \sigma]$, where σ is set as 100. Performance comparison of RVBFSI and VBFSI with other imputation methods are shown in Table 2. RVBFSI outperforms all imputation methods and robust imputation methods RVBSF, RegL_1 and BRTF significantly.

5 Conclusion

In this paper, we propose VBFSI for the imputation of spatiotemporal matrices that works even for extreme matrix completion. VBFSI simultaneously models the low rank, temporal evolution, and periodic evolution in one framework to capture the structure in the spatiotemporal data. We incorporate the prior subspace in our model to capture the periodic evolution in the data. We also propose a Robust VBFSI for missing data imputation in the presence of outliers. It is observed that RVBFSI performs better than the other imputation methods in the presence of the outliers.

References

[1] A. Anjomshoaa, F. Duarte, D. Rennings, T. J. Matarazzo, P. deSouza, and C. Ratti, "City scanner: Building and scheduling a mobile sensing platform for smart city services," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4567–4579, 2018.

[2] P. D. Sampson, A. A. Szpiro, L. Sheppard, J. Lindström, and J. D. Kaufman, "Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data," *Atmospheric Environment*, vol. 45, no. 36, pp. 6593–6606, 2011.

[3] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[4] A. Eftekhari, D. Yang, and M. B. Wakin, "Weighted matrix completion and recovery with prior subspace information," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4044–4071, 2018.

[5] X. Zhang, W. Cui, and Y. Liu, "Matrix completion with prior subspace information via maximizing correlation," *arXiv preprint arXiv:2001.01152*, 2020.

[6] H.-F. Yu, N. Rao, and I. S. Dhillon, "Temporal regularized matrix factorization for high-dimensional time series prediction," in *Advances in neural information processing systems*, 2016, pp. 847–855.

[7] J. Luttinen, "Fast variational Bayesian linear state-space model," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 305–320.

[8] C. Paliwal, U. Bhatt, P. Biyani, and K. Rajawat, "Traffic estimation and prediction via online variational bayesian subspace filtering," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2021.

[9] M. T. Asif, N. Mitrovic, J. Dauwels, and P. Jaillet, "Matrix and tensor based methods for missing data estimation in large traffic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1816–1825, 2016.

[10] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian cp factorization of incomplete tensors with automatic rank determination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1751–1763, 2015.

[11] X. Chen, Z. He, and L. Sun, "A bayesian tensor decomposition approach for spatiotemporal traffic data imputation," *Transportation research part C: emerging technologies*, vol. 98, pp. 73–84, 2019.

[12] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse bayesian methods for low-rank matrix estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 3964–3977, 2012.

[13] L. Yang, J. Fang, H. Duan, H. Li, and B. Zeng, "Fast low-rank bayesian matrix completion with hierarchical gaussian prior models," *IEEE Transactions on Signal Processing*, 2018.

[14] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, and S.-I. Amari, "Bayesian robust tensor factorization for incomplete multiway data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 736–748, 2015.

[15] . Z. H. Xinyu Chen, Yixian Chen, "Urban traffic speed dataset of guangzhou, china [data set].zenodo." 2018. [Online]. Available: http://doi.org/10.5281/zenodo.1205229

[16] . X. C. Yixian Chen, "A large scale pems traffic speed dataset (version v1) [data set].zenodo." 2020. [Online]. Available: http://doi.org/10.5281/zenodo.3939793

[17] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proceedings of the 21th SIGKDD conference on Knowledge Discovery and Data Mining*, August 2015. [Online]. Available: http://research.microsoft.com/apps/pubs/?id=246398

[18] G. Ongie, R. Willett, R. D. Nowak, and L. Balzano, "Algebraic variety models for high-rank matrix completion," in *International Conference on Machine Learning*, 2017, pp. 2691–2700.

[19] X. Chen and L. Sun, "Bayesian temporal factorization for multidimensional time series prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[20] L. Yuan, C. Li, D. Mandic, J. Cao, and Q. Zhao, "Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9151–9158.

[21] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi, "Practical low-rank matrix approximation under robust l 1-norm," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1410–1417.

[22] T. Salimans, D. Kingma, and M. Welling, "Markov chain monte carlo and variational inference: Bridging the gap," in *International Conference on Machine Learning*, 2015, pp. 1218–1226.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[24] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

A Appendix

A.1 Variational Bayesian Filtering with Subspace Information

A.1.1 Variational Bayesian Inference for Spatiotemporal Matrix Completion

We utilize the mean-field approximation, wherein the posterior distribution of parameters $\boldsymbol{\theta} := \{\mathbf{U}, \mathbf{V}, \mathbf{F}, \beta, \boldsymbol{\gamma}, \boldsymbol{\upsilon}\}$ is factorized into a set of conditionally independent components. It is expressive as it captures the marginal density of the parameters. The main advantage of this assumption is that the optimization takes the form of coordinate ascent where the posterior distribution of each parameter can be found by taking expectation of all the other parameters in an iterative manner. The posterior distribution of parameters is factorized as:

$$\boldsymbol{p}(\boldsymbol{\theta} \mid \mathbf{x}_\Omega) = q_{\mathbf{U}}(\mathbf{U}) q_{\mathbf{V}}(\mathbf{V}) q_{\mathbf{F}}(\mathbf{F}) q_{\boldsymbol{\upsilon}}(\boldsymbol{\upsilon}) q_\beta(\beta) q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}). \tag{10}$$

The posterior distribution of all the parameters is determined by minimizing the Kullback–Leibler divergence of $\boldsymbol{p}(\boldsymbol{\theta}|\mathbf{x}_\Omega)$ from $q(\boldsymbol{\theta})$, usually via an alternating minimization approach [23].

We use variational inference to estimate the posterior distribution of $q_{\mathbf{U}}$, $q_{\mathbf{V}}$, $q_{\mathbf{F}}$, $q_{\boldsymbol{\upsilon}}$, q_β, and $q_{\boldsymbol{\gamma}}$ for sampled data $\mathbf{Z} = \mathbf{P}_\Omega(\mathbf{X})$. The updates for posterior distribution of parameters are similar to the updates derived in [8, 12].
The posterior distribution for a row of $\mathbf{U}$ is given by

$$q_{\mathbf{u}_i} = \mathcal{N}(\mathbf{u}_i \mid \boldsymbol{\mu}_i^{\mathbf{U}}, \boldsymbol{\Xi}_i^{\mathbf{U}}) \tag{11}$$

The updates of mean and covariance of $\mathbf{U}$ are derived as

$$\begin{aligned}(\boldsymbol{\Xi}_i^{\mathbf{U}})^{-1} = \hat{\gamma}_i \mathbf{I}_r + \hat{\beta} \sum_{\tau|(i,\tau)\in\Omega} (\boldsymbol{\mu}_\tau^{\mathbf{V}} (\boldsymbol{\mu}_\tau^{\mathbf{V}})^T + \boldsymbol{\Xi}_{\tau,\tau}^{\mathbf{V}}) \\ + \eta (\boldsymbol{\Xi}_i^{\mathbf{U}_{d-1}})^{-1}\end{aligned} \tag{12}$$

$$\boldsymbol{\mu}_i^{\mathbf{U}} = \boldsymbol{\Xi}_i^{\mathbf{U}} (\hat{\beta} \sum_{\tau|(i,\tau)\in\Omega} \boldsymbol{\mu}_\tau^{\mathbf{V}} Z_{i\tau} + \eta (\boldsymbol{\Xi}_i^{\mathbf{U}_{d-1}})^{-1} \boldsymbol{\mu}_i^{\mathbf{U}_{d-1}}) \tag{13}$$

The mean and covariance for the Posterior Distribution of $\mathbf{V}$ are as follows:

$$q_{\mathbf{V}}(\mathbf{V}) = \mathcal{N}(\vec{\mathbf{V}} \mid \boldsymbol{\mu}^{\mathbf{V}}, \boldsymbol{\Xi}^{\mathbf{V}}) \tag{14}$$

$$\boldsymbol{\mu}^{\mathbf{V}} = \boldsymbol{\Xi}^{\mathbf{V}} \begin{bmatrix} \hat{\beta} \sum_{i|(i,1)\in\Omega} Z_{i,1} \boldsymbol{\mu}_i^{\mathbf{U}} + \boldsymbol{\Lambda}_1^{-1} \boldsymbol{\mu}_1 \\ \hat{\beta} \sum_{i|(i,2)\in\Omega} Z_{i,2} \boldsymbol{\mu}_i^{\mathbf{U}} \\ \vdots \\ \hat{\beta} \sum_{i|(i,t)\in\Omega} Z_{i,t} \boldsymbol{\mu}_i^{\mathbf{U}} \end{bmatrix} \tag{15}$$

$$\left[\boldsymbol{\Xi}^{\mathbf{V}}\right]^{-1} = \hat{\beta}\text{Diag}\left(\boldsymbol{\Xi}^{\mathbf{U}}_{(1)}, \ldots, \boldsymbol{\Xi}^{\mathbf{U}}_{(t)}\right) + + \begin{bmatrix} \boldsymbol{\Lambda}_1^{-1} & -\hat{\mathbf{F}} & \ldots & 0 \\ -\hat{\mathbf{F}} & \mathbf{I}_r + \boldsymbol{\Sigma}^{\mathbf{F}} & -\hat{\mathbf{F}} & \ldots \\ \vdots & \vdots & & \vdots \\ \ldots & 0 & -\hat{\mathbf{F}} & \mathbf{I}_r \end{bmatrix} \tag{16}$$

The direct inversion of the dense matrix $\boldsymbol{\Xi}^{\mathbf{V}}$ would be computationally costly. The block-tridiagonal structure $(\boldsymbol{\Xi}^{\mathbf{V}})^{-1}$ can be exploited to carry out the updates for $\boldsymbol{\Xi}^{\mathbf{V}}$ in an efficient manner using LDL decomposition [7, 8].

The updates of the posterior distribution of $\mathbf{F}$ are given by

$$q_{\mathbf{f}_i} = \mathcal{N}(\mathbf{f}_i \mid \boldsymbol{\mu}_i^{\mathbf{F}}, \boldsymbol{\Xi}_i^{\mathbf{F}}) \tag{17}$$

$$\boldsymbol{\mu}_i^{\mathbf{F}} = [\boldsymbol{\Xi}_i^{\mathbf{F}}(\boldsymbol{\mu}_\tau(\boldsymbol{\mu}_{\tau-1})^T + \boldsymbol{\Xi}^{\mathbf{V}}_{\tau,\tau-1})]_i \tag{18a}$$

$$(\boldsymbol{\Xi}_i^{\mathbf{F}})^{-1} = \text{Diag}\,(\hat{\boldsymbol{\upsilon}}) + \sum_{\tau=1}^{t-1}(\boldsymbol{\mu}_\tau(\boldsymbol{\mu}_{\tau-1})^T + \boldsymbol{\Xi}^{\mathbf{V}}_{\tau,\tau-1}) \tag{18b}$$

The posterior distribution for hyperparameters $\{\beta, \boldsymbol{\gamma}, \mathbf{v}\}$ are given by

$$q_\beta(\beta) = \text{Ga}(\beta; a^\beta, b^\beta) \tag{19a}$$

$$q_{\gamma_i}(\gamma_i) = \text{Ga}(\gamma_i; a_i^\gamma, b_i^\gamma) \tag{19b}$$

$$q_{\upsilon_i}(\upsilon_i) = \text{Ga}(\upsilon_i; a_i^\upsilon, b_i^\upsilon) \tag{19c}$$

where $\text{Ga}(x, a, b)$ denotes the Gamma pdf with parameters a and b. The updates for $\{\beta, \boldsymbol{\gamma}, \mathbf{v}\}$ are given by

$$\hat{\upsilon}_i = \frac{r}{\sum_{k=1}^{r}\left([\boldsymbol{\mu}_k^{\mathbf{F}}]_i^2 + [\boldsymbol{\Xi}_k^{\mathbf{F}}]_{ii}\right)} \tag{20a}$$

$$\hat{\gamma}_i = \frac{n+t}{\sum_{k=1}^{n}\left([\boldsymbol{\mu}_k^{\mathbf{U}}]_i^2 + [\boldsymbol{\Sigma}_k^{\mathbf{U}}]_{ii}\right) + \sum_{k=1}^{t}\left([\boldsymbol{\mu}_k^{\mathbf{V}}]_i^2 + [\boldsymbol{\Sigma}_k^{\mathbf{V}}]_{ii}\right)} \tag{20b}$$

$$\hat{\beta} = \frac{p\,n\,t}{\parallel \mathbf{Z} - P_\Omega(\mathbf{U}\,\mathbf{V}^T) \parallel_F^2} \tag{20c}$$

We update the mean, covariance of $\mathbf{U}, \mathbf{V}, \mathbf{F}$ and the hyperparameters $\boldsymbol{\gamma}, \boldsymbol{\upsilon}, \beta$ iteratively as shown in Algorithm 1

Algorithm 1 VBSFI

Input: $\boldsymbol{\Xi}^{\mathbf{U}_{d-1}}, \boldsymbol{\mu}^{\mathbf{U}_{d-1}}, \mathbf{P}_\Omega(\mathbf{X})$
Initialization:$\boldsymbol{\gamma}, \beta, \boldsymbol{\upsilon}, \boldsymbol{\Xi}^{\mathbf{U}}, \boldsymbol{\mu}^{\mathbf{U}}, \boldsymbol{\Xi}^{\mathbf{V}}, \boldsymbol{\mu}^{\mathbf{V}}, \boldsymbol{\Xi}^{\mathbf{F}}, \boldsymbol{\mu}^{\mathbf{F}}, \mathbf{Z}$

1: **while** $X_{conv} < 10^{-5}$ **do**
2: $\mathbf{X}_{\mathbf{old}} = \hat{\mathbf{X}}$
Compute $\mathbf{V}, \mathbf{F}, \boldsymbol{\upsilon}, \beta$ using (21, 22, 24, 26a, 26c)
Compute $\mathbf{U}, \boldsymbol{\gamma}, \beta$ using (18, 19, 26b, 26c)
$\hat{\mathbf{X}} = \boldsymbol{\mu}^{\mathbf{U}}(\boldsymbol{\mu}^{\mathbf{V}})^T$
$X_{conv} = \frac{\left\|\hat{\mathbf{X}} - \mathbf{X}_{\mathbf{old}}\right\|_F}{\|\mathbf{X}_{\mathbf{old}}\|_F}$
3: **end while**
4: **Output**: $\hat{\mathbf{X}}$

A.2 Robust Variational Bayesian Filtering with Subspace Information (RVBFSI)

A.2.1 Bayesian Model

The conditional distribution of generating the entries of $\mathbf{X}_\Omega$ can be defined as

$$\boldsymbol{p}(\mathbf{X}_\Omega \mid \mathbf{U}, \mathbf{V}, \mathbf{E}, \beta) = \prod_{(i,j)\in\Omega} \mathcal{N}(X_{ij} \mid (\mathbf{u}_i.\mathbf{v}_j.^T + E_{ij}), \beta^{-1}) \tag{21}$$

Columns of $\mathbf{U}$,$\mathbf{V}$ and $\mathbf{F}$ and precision variables $\boldsymbol{\gamma}, \beta, \upsilon$ follows same the prior distribution defined in (6-10).
Each entry of sparse outlier matrix E_{ij} is assigned a precision α_{ij}.

$$\boldsymbol{p}(\mathbf{E}|\boldsymbol{\alpha}) = \prod_{i|((i,j)\in\Omega)} \prod_{j|((i,j)\in\Omega)} \mathcal{N}(E_{ij} \mid 0, \alpha_{ij}^{-1}) \tag{22}$$

where the α_{ij} have the non informative prior

$$\boldsymbol{p}(\alpha_{ij}) = \frac{1}{\alpha_{ij}} \tag{23}$$

This works similar to the ARD where instead of column of the matrix, each entry of the matrix is assigned with a precision. Whenever $\alpha_{i,j}$ is driven to a large value , the $E_{i,j} \to 0$ thereby enforcing sparsity. The overall joint distribution for Robust Spatio-Temporal Matrix Completion is expressed as

$$\begin{aligned}\boldsymbol{p}(\mathbf{X}_\Omega, \mathbf{U}, \mathbf{V}, \mathbf{F}, \mathbf{E}, \beta, \boldsymbol{\gamma}, \upsilon, \alpha) = \boldsymbol{p}(\mathbf{X}_\Omega|\mathbf{U}, \mathbf{V}, \mathbf{E}, \beta)\boldsymbol{p}(\mathbf{U}|\boldsymbol{\gamma}) \\ \times \boldsymbol{p}(\mathbf{V}|\mathbf{F})\boldsymbol{p}(\mathbf{F}|\boldsymbol{\upsilon})\boldsymbol{p}(\mathbf{E}|\boldsymbol{\alpha})\boldsymbol{p}(\beta)\boldsymbol{p}(\boldsymbol{\upsilon})\boldsymbol{p}(\boldsymbol{\gamma})\end{aligned} \tag{24}$$

The full bayesian model for the Robust Spatio-Temporal Matrix Completion is depicted in Fig. 2.

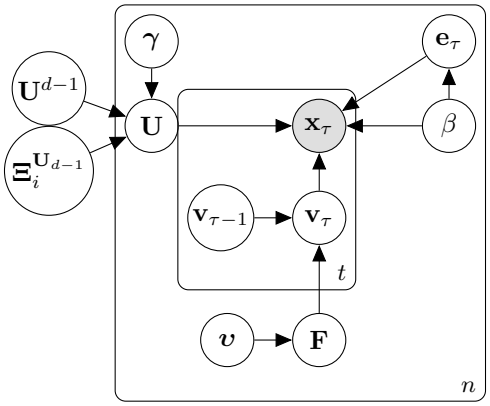


Figure 2: Robust Variational Bayesian Filtering with Subspace Information

A.2.2 Variational Bayesian Inference

Approximate posterior distrubution of parameters $\boldsymbol{\theta}_R := \{\mathbf{U}, \mathbf{V}, \mathbf{F}, \mathbf{E}, \beta, \boldsymbol{\gamma}, \boldsymbol{\upsilon}, \boldsymbol{\alpha}\}$ are derived using Variational Inference.

We utilize the mean-field approximation, wherein the posterior distribution of parameters $\boldsymbol{\theta}_R$ is factorized as:

$$q_{\mathbf{U}}(\mathbf{U})q_{\mathbf{V}}(\mathbf{V})q_{\mathbf{F}}(\mathbf{F})q_{\mathbf{E}}(\mathbf{E})q_{\boldsymbol{\upsilon}}(\boldsymbol{\upsilon})q_\beta(\beta)q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}).$$

The posterior distribution of $q_{\mathbf{U}}$, $q_{\mathbf{V}}$, $q_{\mathbf{F}}$, $q_{\boldsymbol{\upsilon}}$, q_β, and $q_{\boldsymbol{\gamma}}$ takes the same form for $\mathbf{Z} = \mathbf{P}_\Omega(\mathbf{X} - \mathbf{E})$ as shown in (17-26c). The posterior distribution for $\mathbf{E}$ take the following form $\forall(i,j) \in \Omega$.

$$q(E_{ij}) = \mathcal{N}(E_{ij} \mid \mu_{ij}^{\mathbf{E}}, \Xi_{ij}^{\mathbf{E}}) \tag{25}$$

$$\boldsymbol{\Xi}_{i,j}^{\mathbf{E}} = \frac{1}{\hat{\beta} + \hat{\alpha_{i,j}}} \tag{26}$$

$$\boldsymbol{\mu}_{i,j}^{\mathbf{E}} = \hat{\beta}\, \boldsymbol{\Xi}_{i,j}^{\mathbf{E}} (X_{i,j} - \boldsymbol{\mu}_{i.}^{A} (\boldsymbol{\mu}_{j.}^{B})^{T}) \tag{27}$$

$$\hat{\alpha}_{i,j}^{new} = \frac{1 - \hat{\alpha}_{i,j}^{old}\, \boldsymbol{\Xi}_{i,j}^{\mathbf{E}}}{(\boldsymbol{\mu}_{i,j}^{\mathbf{E}})^{2}} \tag{28}$$

$\hat{\alpha}_{i,j}^{new}$ is the fixed-point update for $\boldsymbol{\alpha}$. This is used in the sparse bayesian learning that leads to much faster convergence and enhanced sparsity [12,24]. For robust estimation of entries in the presence of outliers, we update the mean, covariance of $\mathbf{U}, \mathbf{V}, \mathbf{F}, \mathbf{E}$ and the hyperparameters $\boldsymbol{\gamma}, \boldsymbol{\upsilon}, \beta, \boldsymbol{\alpha}$ iteratively as shown in Algorithm 2.

Algorithm 2 RVBSFI

Input: $\boldsymbol{\Xi}^{\mathbf{U}_{d-1}}, \boldsymbol{\mu}^{\mathbf{U}_{d-1}}, \mathbf{P}_{\Omega}(\mathbf{X})$
Initialization:$\boldsymbol{\gamma}, \beta, \boldsymbol{\upsilon}, \boldsymbol{\Xi}^{\mathbf{U}}, \boldsymbol{\mu}^{\mathbf{U}}, \boldsymbol{\Xi}^{\mathbf{V}}, \boldsymbol{\mu}^{\mathbf{V}}, \boldsymbol{\Xi}^{\mathbf{F}}, \boldsymbol{\mu}^{\mathbf{F}}, \boldsymbol{\mu}^{\mathbf{E}}, \boldsymbol{\Xi}^{\mathbf{E}}$,
$\boldsymbol{\alpha}, \mathbf{Z} = \mathbf{P}_{\Omega}(\mathbf{X} - \mathbf{E})$

1: **while** $X_{conv} < 10^{-5}$ **do**
2: $\mathbf{X}_{\mathbf{old}} = \hat{\mathbf{X}}$
Compute $\mathbf{V}, \mathbf{F}, \boldsymbol{\upsilon}, \beta$ using (21, 22, 24, 26a, 26c)
Compute $\mathbf{U}, \boldsymbol{\gamma}, \mathbf{E}, \boldsymbol{\alpha}, \beta$ (18, 19, 26b, 26c)
Compute $\mathbf{E}, \boldsymbol{\alpha}, \beta$ using (28-29, 26c)
$\hat{\mathbf{X}} = \boldsymbol{\mu}^{\mathbf{U}} (\boldsymbol{\mu}^{\mathbf{V}})^{T}$
$X_{conv} = \frac{\left\| \hat{\mathbf{X}} - \mathbf{X}_{\mathbf{old}} \right\|_{F}}{\left\| \mathbf{X}_{\mathbf{old}} \right\|_{F}}$
3: **end while**
4: **Output**: $\hat{\mathbf{X}}$