Missing Data Prediction in Multi-source Time Series with Sensor Network Regularization

Weiwei Shi*, Yongxin Zhu*, Xiao Pan*, Philip S. Yu†, Bin Liu*, and Yufeng Chen‡

*School of Microelectronics, Shanghai Jiao Tong University, China
*Shijiazhuang Tiedao University, China
†Department of Computer Science, University of Illinois at Chicago, USA
‡Shandong Power Supply Company of State Grid, China
{iamweiweishi,zhuyongxin}@sjtu.edu.cn
smallpx@stdu.edu.cn, psyu@uic.edu, liubinpoem@gmail.com,
chenyufeng169@sina.com

Abstract. Raw multi-source time series usually contain missing values, which can hardly meet the requirement of precise analysis. To address the problem, we propose a novel method named Matrix factorization with Sensor Network Regularization (MSNR). In the paper, we consider two sensors in a sensor network to be ***correlated*** if one sensor has a strong correlation with the other one. On the contrary, if one sensor has a weak correlation with another one, we regard the two sensors as ***uncorrelated***. The proposed method aims to predict missing data by minimizing the difference between a sensor and its correlated sensors or to maximize the difference between a sensor and its uncorrelated sensors via matrix factorization. In the process of matrix factorization, we impose the sensor network regularization terms to constrain the objective functions of matrix factorization. To treat the correlated or uncorrelated sensors differently, we further improve the objective functions by incorporating similarity functions. Extensive experiments on real-world data sets demonstrate that the proposed approach MSNR can effectively improve the performance of missing data prediction in multi-source time series, even when the missing ratio is as high as 90 percent.

Keywords: missing data prediction, matrix factorization, multi-source time series

1 Introduction

Multi-source time series are ubiquitous in many real-world applications, such as electric equipment monitoring, weather forecasting, environment state monitoring, security surveillance, and so on [1, 2]. In most applications, multiple sensors are used to generate time series data, and they usually share one common goal. In this paper, the sensors sharing one common goal are treated as a sensor network. Unfortunately, the raw time series in a sensor network usually contain missing values due to the harsh working conditions or uncontrollable factors. A

large collection of data mining and statistical methods have been proposed to predict the missing values of time series [3]. Grabocka et. al. proposed a matrix factorization method to classify multiple time series. Their work aims at extracting latent factors based on observed entries [4]. However, these methods either focus on predicting the missing data in the time series from one single source or could not effectively deal with the missing data prediction problem of the time series from multiple sources. In this paper, aiming at solving the above problems, we propose MSNR, a matrix factorization with sensor network regularization method, that utilizes the correlation information among the different sensors in a sensor network to improve the accuracy of missing data prediction in multi-source time series. Moreover, to treat the correlated or uncorrelated sensors differently, we further improve the sensor network regularization terms of the objective function by incorporating similarity functions. The experimental results reveal that our proposed method shows superior performance to the state-of-the-art algorithms.

2 Proposed Methods

2.1 Low Rank Matrix Factorization

Let X be the multi-source time series collected from N different data sources, and the jth entity in time series data from the ith source can be denoted as X_{ij} for $i = \{1, 2, 3, \ldots, N\}$, $j = \{1, 2, 3, \ldots, M\}$. As the original matrix X might contain a great number of missing values, we only need to factorize the observed entities in X. Hence, we have a optimization problem based on Singular Value Decomposition (SVD):

$$\min_{S,V} \frac{1}{2}\|W \circ (X - SV^T)\|_F^2 + \frac{\lambda_1}{2}\|S\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2, \tag{1}$$

where $S \in \mathbb{R}^{N*L}$, $V \in \mathbb{R}^{M*L}$ with $L < min(N, M)$, $\lambda_1, \lambda_2 > 0$, W is an indicator matrix, and $\circ$ denotes the Hadamard product. Two regularization terms $\|S\|_F^2$ and $\|V\|_F^2$ are added in order to avoid overfitting. Gradient based approaches can be applied to find a minimum due to their effectiveness and simplicity [5].

2.2 Model 1: Correlated Sensors based Regularization

In a sensor network, although the different sensors are assigned different tasks, they usually share one common goal and there might exist strong correlation among some of the sensors. If one sensor has a strong correlation with another one, we call the two sensors are *correlated*.

As S denotes the latent sensor matrix and there might be strong correlation among correlated sensors, we propose the first missing data prediction model based on matrix factorization technique with the following optimization problem:

$$\begin{aligned}\min_{S,V} \mathcal{L}(X, S, V) = \frac{1}{2}\|W \circ (X - SV^T)\|_F^2 + \frac{\lambda_1}{2}\|S\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2 + \\ \frac{\alpha}{2}\sum_{i=1}^{N}\|S_i - \frac{\sum_{c \in C(i)} H(i,c) * \rho_{i,c} S_c}{\sum_{c \in C(i)} H(i,c)}\|_F^2.\end{aligned} \tag{2}$$

where α is the penalty factor and $\alpha > 0$, $C(i)$ denotes the set of the correlated sensors of the ith sensor and $|C(i)|$ is the total number of these correlated sensors. The included scaling factor $\rho_{i,c}$ aims at matching the scale difference between the ith sensor and the cth sensor. In the first model, we incorporate one sensor network regularization term, i.e., the correlated sensors based regularization term. Concretely, if the correlated sensors are $C(i)$, we deduce that the state of the ith sensor is correlated to the average state of $C(i)$. The function $H(i, c)$ measures the similarity between the ith sensor and the cth sensor. From this improved regularization item, we know that if the cth sensor is very correlated to the ith sensor, the value of $H(i, c)$ will be large, i.e, it contributes more to the state of the ith sensor.

2.3 Model 2: Uncorrelated Sensors based Regularization

The first model we propose imposes a correlated sensors based regularization term to constrain the matrix factorization. From the opposite view, if one sensor has a weak correlation with another one, we call the two sensors are *uncorrelated*. And we also employ another sensor network regularization term, i.e., the uncorrelated sensors based regularization term, to build the second model. Since uncorrelated sensors share weak correlation, we attempt to add one constrain term to maximize the distance between the ith sensor and its uncorrelated sensors. Consequently, the optimization problem in Equation (2) is updated as:

$$\begin{aligned}\min_{S,V} \mathcal{L}'(X, S, V) = \frac{1}{2}\|W \circ (X - SV^T)\|_F^2 + \frac{\lambda_1}{2}\|S\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2 - \\ \frac{\alpha'}{2}\sum_{i=1}^{N}\|S_i - \frac{\sum\limits_{c' \in C'(i)} H(i, c') * \rho_{i,c'} S_{c'}}{\sum\limits_{c' \in C'(i)} H(i, c')}\|_F^2.\end{aligned} \tag{3}$$

where α' is the penalty factor and $\alpha' > 0$, $C'(i)$ denotes the set of the uncorrelated sensors of the ith sensor.

2.4 Similarity Function

The proposed regularization terms in Equation (2) and Equation (3) require a function H to measure the similarity between two sensors, which is a key component of the proposed method. Vector Space Similarity (VSS) is applied to measure the similarity between two sensors i and c:

$$H_{VSS}(i, c) = \frac{\sum\limits_{j \in \boldsymbol{o}_i \cap \boldsymbol{o}_c} X_{ij} \cdot X_{cj}}{\sqrt{\sum\limits_{j \in \boldsymbol{o}_i \cap \boldsymbol{o}_c} X_{ij}^2}\sqrt{\sum\limits_{j \in O_i \cap O_c} X_{cj}^2}}, \tag{4}$$

where $\boldsymbol{o}_i$ and $\boldsymbol{o}_c$ is the subset of $\boldsymbol{x}_i$ and $\boldsymbol{x}_c$. The entities in $\boldsymbol{o}_i$ and $\boldsymbol{o}_c$ are observed.

Another way to measure the similarity between two sensors i and c is based on Gaussian Kernel (GK):

$$H_{GK}(i, c) = exp(-\frac{\sum\limits_{j \in \boldsymbol{o}_i \cap \boldsymbol{o}_c} (X_{ij} - X_{cj})^2}{2\sigma^2}). \tag{5}$$

Another commonly used function Pearson Correlation Coefficient (PCC) is employed to take the different scales between two sensors into consideratio. The details of PCC could be found in [6].

Dynamic Time Warping (DTW) is a well-known technique to compare two time series with different length. The strategy is to find a warping path W that minimize the warping cost. This path and the relevant details can be found using dynamic programming [7]. To make it consistent that a larger value of H means that sensors i and c are more correlated, the reciprocal of DTW is employed as the similarity function:

$$H_{DTW}(i,c) = \frac{1}{DTW(\boldsymbol{o}_i, \boldsymbol{o}_c)}. \tag{6}$$

Furthermore, to better reveal the necessity of incorporating similarity functions, a constant function (CF) $H_{CF}(i,c) = C$ is also employed as the baseline function in the paper.

3 Experiments

In this section, to demonstrate the effectiveness of the proposed method MSNR, we conduct extensive experiments on two real-world data sets, which include the Motes data set [8] and the Diagnostic Gases data set [6].

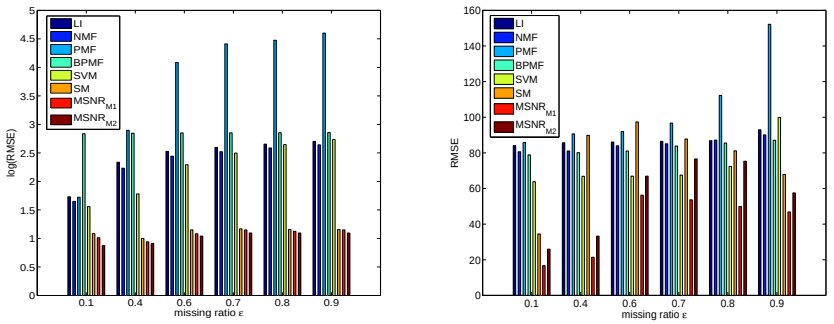
To evaluate all the methods fairly, we incrementally simulate the data missing of the two data sets with an increasing missing ratio. For example, to increase the missing ratio from 0.10 to 0.15, we randomly move 5% of the total data from the observed data set to the missing data set. In this way, the subsequent missing data set always contains the missing data of the previous one.

From the Equation (2) and (3), we know that the constant value C will not change the value of the equations. Thus, C could be simply set as 1. Besides, the parameters λ_1 and λ_2 are both set equal to λ in this paper. The parameter $|C(i)|$ determines how many correlated or uncorrelated sources should be incorporated into the optimization functions. To make the paper more concise, we denote the method MSNR based on Model 1 as MSNR_{M1} and use MSNR_{M2} to indicate the method MSNR based on Model 2. We compare the proposed method MSNR with many baseline methods in predicting the values of the missing samples in multi-source time series. The comparison methods used include: Linear Interpolation (LI), Non-negative Matrix Factorization (NMF) [9], Probabilistic Matrix Factorization (PMF) [10], Bayesian PMF (BPMF) [11], Support Vector Machine (SVM) [6] and Simplified MSNR (SM), which is a simplified version of MSNR with $\alpha = 0$. To evaluate the performance of the proposed method, root mean squared error (RMSE) is used to measure the prediction quality [6]:

$$RMSE = \sqrt{\frac{\sum_{i,j}(1 - W_{ij})(X_{ij} - \hat{X}_{ij})^2}{\sum_{i,j}(1 - W_{ij})}}, \tag{7}$$

where X_{ij} is the observed value and $\hat{X}_{ij}$ is the corresponding predicted value. W is the indicator matrix.

3.1 Experimental Results



(a) Motes (N=54, M=14000) (b) Diagnostic Gases (N=5, M=500)

Fig. 1. Performance of the proposed method MSNR and baseline methods.

Fig. 1 shows the experimental results of the proposed methods and baseline methods on the Motes and the Diagnostic Gases data sets. The logarithm of RMSE is shown in the vertical axis to present the experimental results more clearly. First, we take the first model MSNR_{M1} into consideration. We can see that MSNR_{M1} consistently outperforms the other baseline methods. For instance, as for the Motes data set, when the missing ratio ϵ is equal to 0.4, MSNR_{M1} achieves the lowest RMSE 2.56, which is about 85% lower than PMF. Even when the missing ratio exceeds 60%, the RMSE of MSNR_{M1} is still within a reasonable range. As for the Diagnostic Gases data set, the proposed method also obtains the best performance, while the other baseline methods show barely satisfactory results even when the missing ratio is as low as 0.1.

Moreover, given the second model MSNR_{M2}, we can observe that the RMSE of MSNR_{M2} is generally a bit lower than MSNR_{M1} for the Motes data set. However, as for the Diagnostic Gases data set, the performance of MSNR_{M2} is not as good as that of MSNR_{M1}. As the Motes data set generates from 54 sensors, the Motes data set has a much higher chance of containing uncorrelated sensors. Thus MSNR_{M2} shows better performance for the Motes data set. On the contrary, the Diagnostic Gases data set is collected from only five sensors. As a consequence, it is much more important for the Diagnostic Gases data set to find the correlated sensors, thus MSNR_{M1} shows superior performance. Nevertheless, as MSNR_{M1} and MSNR_{M2} show better performance than the baseline methods, they are both alternative models in the proposed method.

Furthermore, it is noteworthy that the only difference between the proposed method and SM is whether the sensor network regularization terms are incorporated or not. The SM method shows larger RMSE than both MSNR_{M1} and MSNR_{M2} based on the experimental results in the Fig. 1. As a consequence, we may safely deduce that the incorporated network regularization terms mainly contribute to the superior performance of the proposed method.

Table 1. Performance of MSNR_{M1} with different similarity functions and missing ratio ϵ.

	ϵ	VSS	GK	PCC	DTW	CF
Motes data set	0.1	2.48	2.49	**2.41**	2.48	2.50
	0.4	2.61	2.62	**2.55**	2.62	2.59
	0.6	2.94	2.93	**2.90**	2.97	2.93
	0.7	3.08	3.01	2.98	**2.95**	3.09
	0.8	3.07	3.13	2.97	**2.86**	3.06
	0.9	3.06	3.16	2.98	**2.96**	3.09
Diagnostic Gases data set	0.1	40.47	24.66	**16.72**	24.93	56.45
	0.4	52.02	51.16	**21.40**	72.26	54.02
	0.6	62.20	85.82	**56.29**	92.23	56.44
	0.7	55.51	49.92	63.19	**33.40**	74.21
	0.8	80.84	99.06	91.03	**58.66**	91.01
	0.9	63.47	46.84	99.34	**42.50**	62.58

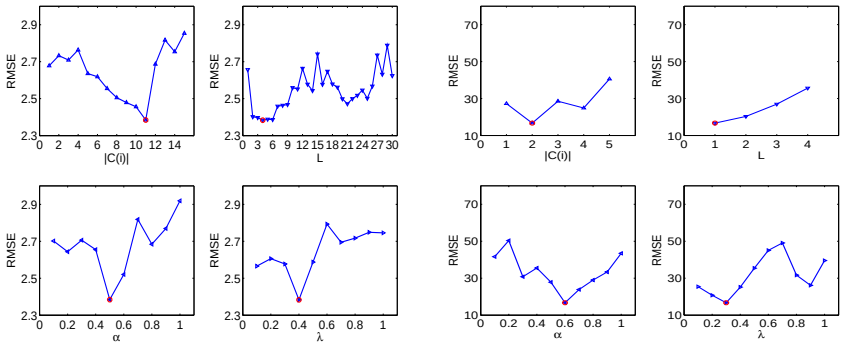
3.2 Similarity Functions Impact Discussion

Due to the lack of space, we only give the impact discussion of the similarity functions in the first model MSNR_{M1}. Similar results are observed for the second model. The similarity function H aims at finding the set of correlated sensors $C(i)$ or the set of uncorrelated sensors $C'(i)$. H directly determines which sensors are correlated or uncorrelated with the ith sensor and the weights of the sensor network regularization terms. Thus, we mainly focus on the analysis of the similarity functions in this subsection. As Table 1 shows, when the missing ratio is below 0.6, PCC obtains lower RMSE for both of the two data sets. PCC takes the different scales among various sensors into consideration, which might contribute to its better performance. However, DTW achieves far superior performance to the other functions when the missing ratio exceeds 0.6. We deduce that DTW can better measure the similarity between two time series when the missing ratio is high, as it utilizes all the observed entities in the raw time series. Nevertheless, both PCC and DTW are alternative similarity functions in the proposed method. In addition, we observe that the constant function CF shows barely satisfactory results, which further demonstrates the necessity and importance of employing an appropriate similarity function.

3.3 Parameters Impact Discussion

In this subsection, we also only give the analyses of the parameters of the first model MSNR_{M1}, which is shown in Fig. 2.

Firstly, the parameter $|C(i)|$ denotes the total number of the correlated sensors with the ith sensor, which plays a very important role in the proposed method. Taking the Motes data set for example, when $|C(i)|$ is set as 4, the RMSE is equal to 2.76. However, when $|C(i)|$ is equal to 11, the method achieves the lowest RMSE 2.38, which is reduced by about 14%. Specially, the optimum RMSE is about 58% lower than the worst RMSE for the Diagnostic Gases data



(a) Motes (N=54, M=14000) (b) Diagnostic Gases (N=5, M=500)

Fig. 2. Impact of Parameters.

set. We deduce that an oversized $|C(i)|$ will bring in more noise while too small a $|C(i)|$ will be not enough to constrain the matrix factorization. Thus, an appropriate value of $|C(i)|$ is of great importance in the proposed method.

Then, the impact of the dimension of L on the performance is also shown in the figure. On the whole, the RMSE is consistently below a reasonable value. Concretely, for the Motes data set, the RMSE is equal to 2.65 when L is set as 1, while the lowest RMSE 2.38 is obtained when L is equal to 4. Nevertheless, based on the experimental results, we may safely set $L = 4$ and $L = 1$ for the Motes data set and the Diagnostic Gases data set respectively. Hence, the dimension of the latent factors L also plays an important part in the proposed method.

Next, the impact of α on the performance is presented. α controls how much information of the sensor network should be incorporated into the optimization problem. In general, as Fig. 2 shows, RMSE not only shows little variation but also is consistently below a relatively low value for all of the different α values. We can observe that the best performance is achieved when α is equal to 0.5 and 0.6 for the two data sets respectively. We deduce that too small an α would greatly decrease the influence of the sensor regularization term on the matrix factorization. On the other hand, if we employ too large an α, the sensor regularization term would dominate the learning processes. So, an appropriate coefficient α could further improve the performance of the proposed method.

Finally, the penalty coefficient λ is optimized. Based on the experimental results, we can reasonably set $\lambda = 0.4$ and $\lambda = 0.3$ for the Motes data set and the Diagnostic Gases data set respectively.

4 Conclusion

In this paper, we have proposed a novel method MSNR for predicting the missing data in the time series from multiple sources. The method incorporates sensor network regularization terms to minimize the difference between one sensor and its correlated sensors or to maximize the distance between one sensor and

its uncorrelated sensors during matrix factorization. As expected, the proposed method MSNR exhibits higher precision in terms of lower RMSE than classical methods and state-of-the-art matrix factorization based approaches. We deduce that the incorporated network regularization terms mainly contribute to the superior prediction result.

Acknowledgment

This paper is sponsored in part by the National High Technology and Research Development Program of China (863 Program, 2015AA050204), State Grid Science and Technology Project (520626140020, 14H100000552, SGCQDK00PJJS1400020), State Grid Corporation of China, the National Research Foundation, Prime Ministers Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, and the National Natural Science Foundation of China (No.61373032).

References

1. Yongjie Cai, Hanghang Tong, Wei Fan, and Ping Ji. Fast mining of a network of coevolving time series. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 298–306.
2. Nicolas Méger, Christophe Rigotti, and Catherine Pothier. Swap randomization of bases of sequences for mining satellite image times series. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 190–205. Springer, 2015.
3. Shin-Fu Wu, Chia-Yung Chang, and Shie-Jue Lee. Time series forecasting with missing values. In *2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)*, pages 151–156, 2015.
4. Josif Grabocka, Alexandros Nanopoulos, and Lars Schmidt-Thieme. Classification of sparse time series via supervised matrix factorization, 2012.
5. Hao Ma, Dengyong Zhou, Chao Liu, Michael R. Lyu, and Irwin King. Recommender systems with social regularization. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 287–296. ACM, 2011.
6. Weiwei Shi, Yongxin Zhu, Jinkui Zhang, Xiang Tao, Gehao Sheng, Yong Lian, Guoxing Wang, and Yufeng Chen. Improving power grid monitoring data quality: An efficient machine learning framework for missing data prediction. In *HPCC*, pages 417–422. IEEE, 2015.
7. Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.
8. Madden Samuel. Intel lab data. `http://db.csail.mit.edu`.
9. Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
10. Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1257–1264. Curran Associates, Inc., 2008.
11. Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887. ACM, 2008.