2024 IEEE 20th International Conference on Automation
Science and Engineering (CASE)
August 28 - September 1, 2024. Bari, Italy

2024 IEEE 20th International Conference on Automation Science and Engineering (CASE) | 979-8-3503-5851-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/CASE59546.2024.10711621

Dynamic Feature Extraction and Prediction for High Dimensional Time Series with Seasonality

Bingyuan Li, Yang Wang, Yang Zhao, Lishuai Li, Yining Dong*

Abstract— **In this paper, a novel reduced-dimensional seasonal autoregressive modeling algorithm with a canonical correlation analysis objective (RDSAR-CCA) is developed for dynamic feature extraction and prediction in high-dimensional time series with seasonality. The proposed algorithm estimates a seasonal reduced-dimensional dynamic model for extracting and modeling the latent dynamics within the data. This approach facilitates dynamic latent variable (DLV) analysis in high-dimensional seasonal time series. The DLVs extracted by the proposed algorithm are orthogonal and ranked in descending order of predictability, which simplifies interpretation and enhances visualization. The effectiveness and superiority of the proposed RDSAR-CCA algorithm are evaluated on the passenger flow data from the Shenzhen metro system.**

I. INTRODUCTION

Analysis and applications on high dimensional time series are drawing increasing attention in many domains. Time series data obtained from dynamic systems often exhibit high dimensionality, typically characterized by strong cross-correlations and auto-correlations [1]–[3]. However, the dynamic of interest in time series is often in a latent space with low dimension [4]. Owing to the strong cross-correlations among variables, a set of lower-dimensional latent variables could represent the original high-dimensional time series.

Traditional reduced-dimensional latent methods such as principal component analysis (PCA), canonical correlation analysis (CCA), and partial least squares (PLS) are wildly used for data analytics [5]–[7]. These methods only consider static modeling and ignore the auto-correlations or dynamics of time series. Moreover, dynamic factor models (DFM) methods extract dynamic latent factors [8], [9], considering time dependence among data. When extracting latent factors, these DFM methods minimize mean squared reconstruction error rather than mean squared prediction error.

Predictable feature analysis (PFA) aims to extract predictable latent variables from high-dimensional time series by minimizing the mean squared prediction error [10]. This method builds a vector autoregressive model on the original high dimensional time series, leading to numerical stability issues when the time series are highly colinear. Moreover, slow feature analysis (SFA) based methods [11]–[13] try to extract “slowly varying” time series. However, ‘slowly varying’ is not necessary for a predictable time series. There exist predictable time series without slow variations.

Recently, Dong *et al.* introduced the dynamic-inner principal component analysis (DiPCA) [14] and dynamic-inner canonical correlation analysis (DiCCA) [4], [15] for latent dynamic modeling. These methods extract low dimensional dynamic latent variables (DLVs) from high dimensional time series and model them as univariate autoregressive processes. To handle the interactions between DLVs, Dong *et al.* [16] developed an improved algorithm that models DLVs with vector autoregressive processes. Later, Qin [17], [18] developed a latent vector autoregressive modeling algorithm with a CCA objective (LaVAR-CCA) for extracting fully interacting DLVs, whose probabilistic version and state space generalization are developed in [19] and [20]. However, these methods extract DLVs of explicit dynamics without considering seasonality. High dimensional time series data from many dynamic systems, such as climate, transportation, and electricity consumption, often demonstrate typical seasonality due to physical principles and usage patterns. The actual latent structure of these dynamic systems contains various seasonal patterns. These methods fail to capture seasonal patterns, reducing the efficiency of extracting and predicting DLVs.

In this study, we propose a novel reduced-dimensional seasonal autoregressive modeling algorithm with a CCA objective (RDSAR-CCA). The proposed algorithm extracts the lower-dimensional DLVs by maximizing the correlation between DLVs and their self-predictions, thus capturing the most significant dynamic patterns in high-dimensional time series. The key of the proposed algorithm lies in its use of seasonal autoregressive (SAR) models as the inner model, which effectively captures and explicitly models the seasonal patterns within time series data. The DLVs extracted by RDSAR-CCA are uncorrelated to each other and ranked in descending order of predictability, offering clear interpretation and enhanced visualization. By constructing a seasonal latent model that closely aligns with the actual latent structure of the high-dimensional time series, RDSAR-CCA significantly improves prediction accuracy. The passenger flow data from the Shenzhen metro system is utilized to demonstrate the superiority of the proposed RDSAR-CCA algorithm.

II. REDUCED-DIMENSIONAL SAR WITH A CCA OBJECTIVE

Time series models aim to explore the temporal dependence in data. Since high-dimensional time series, especially in climate, transportation, and electricity consumption, often exhibit significant seasonal patterns, we propose a DLV

Bingyuan Li, Yang Wang, Lishuai Li and Yining Dong are with the School of Data Science, City University of HongKong, Kowloon, HongKong. *Corresponding author. Yining Dong (e-mail: yining.dong@cityu.edu.hk).

Yang Zhao is with the School of Public Health (Shenzhen), Sun Yat-sen University, Guangzhou, China.

modeling method to capture and model the seasonal latent patterns in high-dimensional time series. In this section, RDSAR-CCA is developed, which can be applied to the time series of seasonal patterns and achieve practical DLV analysis.

A. Objective function

RDSAR-CCA aims to represent the strong dynamics in data through the latent principal time series $\{t_k\}_{k=1}^{N+h}$, which is explicitly modeled by a SAR model, noted as $\mathrm{SAR}(p, P, l)$:

$$\phi(\mathcal{B})\boldsymbol{\Phi}(\mathcal{B})t_k = \gamma_k \tag{1}$$

where $\phi(\mathcal{B}) = 1 - \sum_{i=1}^{p} \phi_i \mathcal{B}^i$, $\boldsymbol{\Phi}(\mathcal{B}) = 1 - \sum_{j=1}^{P} \Phi_j \mathcal{B}^{jl}$, l defines the length of season and $\mathcal{B}$ is the backward shift operator, which means $\mathcal{B}t_k = t_{k-1}$. The residual γ_k is uncorrelated in time and with zero mean. The vector of model parameters is denoted as $\boldsymbol{\varphi} = [\phi_1 ... \phi_p, \Phi_1 ... \Phi_P]^T$. The dynamic process (1) can be rewritten as

$$t_k = \sum_{i=1}^{p} \phi_i t_{k-i} + \sum_{j=1}^{P} \Phi_j t_{k-jl} - \sum_{i=1}^{p}\sum_{j=1}^{P} \phi_i \Phi_j t_{k-i-jl} + \gamma_k. \tag{2}$$

Therefore, the self-prediction of t_k is given as

$$\hat{t}_k = (1 - \phi(\mathcal{B})\boldsymbol{\Phi}(\mathcal{B}))t_k \equiv G(\mathcal{B})t_k \tag{3}$$

where $G(\mathcal{B})$ is the the prediction transfer function.

Denote a time series $\{\boldsymbol{x}_k\}_{k=1}^{N+s}$ with m variables, such latent variable t_k can be formed by linearly combining the variables of $\boldsymbol{x}_k$:

$$t_k = \boldsymbol{x}_k^T \mathbf{w} \tag{4}$$

where $\mathbf{w}$ is the linear transformation vector. To exclude the initial transition period, we formulate the data matrix $\mathbf{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ ...\ \boldsymbol{x}_{h+N}]^T$ into a submatrix $\mathbf{X}_h = [\boldsymbol{x}_{h+1}\ \boldsymbol{x}_{h+2}\ ...\ \boldsymbol{x}_{h+N}]^T$, followed by defining its corresponding latent score vector $\mathbf{t}_h = [t_{h+1}\ t_{h+2}\ ...\ t_{h+N}]^T$ and predicted latent score vector $\hat{\mathbf{t}}_h = [\hat{t}_{h+1}\ \hat{t}_{h+2}\ ...\ \hat{t}_{h+N}]^T$, where $h = p + Pl$.

Mathematically, the objective function of RDSAR-CCA is developed as maximizing the correlation between $\mathbf{t}_h$ and its self-prediction $\hat{\mathbf{t}}_h$ to make sure $\mathbf{t}_h$ is best predicted. This objective function can be represented as minimizing the prediction error [17]:

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\varphi}} \quad & \left\|\mathbf{t}_h - \hat{\mathbf{t}}_h\right\|^2 \\ \text{s.t.} \quad & \left\|\mathbf{t}_h\right\|^2 = 1 \end{aligned} \tag{5}$$

B. Solution to RDSAR-CCA

To solve the objective function (5) more straightforwardly by transforming the objective function into a simple form, we perform the economy singular value decomposition (SVD) on $\mathbf{X}_h$:

$$\mathbf{X}_h = \mathbf{U}_h \mathbf{D} \mathbf{V}^T \tag{6}$$

For $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, we can get $\mathbf{U}$ from:

$$\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{D}^{-1}. \tag{7}$$

We let

$$\mathbf{t} = \mathbf{U}\mathbf{w}_u = \mathbf{X}\mathbf{w}, \mathbf{t}_h = \mathbf{U}_h\mathbf{w}_u = \mathbf{X}_h\mathbf{w}. \tag{8}$$

In this way, we can use $\mathbf{U}$ instead of data matrix $\mathbf{X}$ to obtain $\mathbf{w}_u$ that minimize objective function (5). From equation (6) and (8), we have the relationship between $\mathbf{w}$ and $\mathbf{w}_u$:

$$\mathbf{w} = \mathbf{V}\mathbf{D}^{-1}\mathbf{w}_u. \tag{9}$$

Therefore, $t_k = \boldsymbol{u}_k^T\mathbf{w}_u = \boldsymbol{x}_k^T\mathbf{w}$, where $\boldsymbol{u}_k^T$ means each row of the matrix $\mathbf{U}$ and $\mathcal{B}\boldsymbol{u}_k = \boldsymbol{u}_{k-1}$. The prediction residuals are equivalent to

$$t_k - \hat{t}_k = \boldsymbol{u}_k^T\mathbf{w}_u - (1 - \phi(\mathcal{B})\boldsymbol{\Phi}(\mathcal{B}))\boldsymbol{u}_k^T\mathbf{w}_u = \tilde{\boldsymbol{u}}_k^T\mathbf{w}_u \tag{10}$$

where $\tilde{\boldsymbol{u}}_k = \phi(\mathcal{B})\boldsymbol{\Phi}(\mathcal{B})\boldsymbol{u}_k$.

Similarly, we denote $\widetilde{\mathbf{U}}_h = [\tilde{\boldsymbol{u}}_{h+1}\ ...\ \tilde{\boldsymbol{u}}_{h+N}]^T$ and $\mathbf{U}_h = [\boldsymbol{u}_{h+1}\ ...\ \boldsymbol{u}_{h+N}]^T$. The prediction error in (5) can be represented in a straightforward form as:

$$\mathbf{t}_h - \hat{\mathbf{t}}_h = \widetilde{\mathbf{U}}_h\mathbf{w}_u \tag{11}$$

The objective function (5) is equivalent to

$$\begin{aligned} \min_{\mathbf{w}_u,\boldsymbol{\varphi}} \quad & \left\|\widetilde{\mathbf{U}}_h\mathbf{w}_u\right\|^2 \\ \text{s.t.} \quad & \left\|\mathbf{U}_h\mathbf{w}_u\right\|^2 = 1 \end{aligned} \tag{12}$$

Importantly, owing to the orthogonality of $\mathbf{U}_h$, $\mathbf{U}_h^T\mathbf{U}_h = \mathbf{I}$. The constraint further becomes: $\mathbf{w}_u^T\mathbf{U}_h^T\mathbf{U}_h\mathbf{w}_u = \left\|\mathbf{w}_u\right\|^2 = 1$. Finally, the objective function becomes:

$$\begin{aligned} \min_{\mathbf{w}_u,\boldsymbol{\varphi}} \quad & J = \left\|\widetilde{\mathbf{U}}_h\mathbf{w}_u\right\|^2 \\ \text{s.t.} \quad & \left\|\mathbf{w}_u\right\|^2 = 1 \end{aligned} \tag{13}$$

Applying Lagrange multiplier to (13):

$$L = \left\|\widetilde{\mathbf{U}}_h\mathbf{w}_u\right\|^2 + \lambda(1 - \left\|\mathbf{w}_u\right\|^2). \tag{14}$$

Setting the partial derivatives to zero,

$$\frac{\partial L}{\partial \mathbf{w}_u} = 2\widetilde{\mathbf{U}}_h^T\widetilde{\mathbf{U}}_h\mathbf{w}_u - 2\lambda\mathbf{w}_u = 0 \tag{15}$$

we have

$$\widetilde{\mathbf{U}}_h^T\widetilde{\mathbf{U}}_h\mathbf{w}_u = \lambda\mathbf{w}_u. \tag{16}$$

Premultiplying equation (16) by $\mathbf{w}_u^T$, we have

$$\mathbf{w}_u^T\widetilde{\mathbf{U}}_h^T\widetilde{\mathbf{U}}_h\mathbf{w}_u = \lambda\mathbf{w}_u^T\mathbf{w}_u. \tag{17}$$

From objective function (13), we get $\mathbf{w}_u^T\widetilde{\mathbf{U}}_h^T\widetilde{\mathbf{U}}_h\mathbf{w}_u = J$ and $\mathbf{w}_u^T\mathbf{w}_u = 1$. Therefore, $J = \lambda$ and then equation (16) becomes

$$\widetilde{\mathbf{U}}_h^T\widetilde{\mathbf{U}}_h\mathbf{w}_u = J\mathbf{w}_u. \tag{18}$$

Clearly, when $\mathbf{w}_u$ is the eigenvector corresponding to the smallest eigenvalue of $\widetilde{\mathbf{U}}_h^T\widetilde{\mathbf{U}}_h$, this objective function is minimized.

To get the solution of $\boldsymbol{\varphi}$ and $\mathbf{w}_u$, we first obtain the initial DLVs with an initialized $\mathbf{w}_u$. Then, $\boldsymbol{\varphi}$ can be estimated by building a SAR model on DLVs, and the order of SAR model (p, P) can be automatically determined with the Hyndman-Khandakar algorithm [21]. When $\boldsymbol{\varphi}$ is given, $\mathbf{w}_u$ can be solved by the eigen-decomposition problem (18). Iterate the above process until convergence to get the final $\mathbf{w}_u$ and φ. The transformation vector $\mathbf{w}$ can be obtained based on Equation (9).

Following the identical procedure, subsequent DLVs can be extracted and modeled by utilizing the deflated matrix $\mathbf{X}$.

$$\mathbf{X} := \mathbf{X} - \mathbf{t}\mathbf{p}^T \tag{19}$$

where $\mathbf{p} = \mathbf{X}^T\mathbf{t}/\mathbf{t}^T\mathbf{t}$ is the loading vector. The procedure of RDSAR-CCA is summarized in Algorithm 1.

Algorithm 1 RDSAR-CCA Algorithm

```
1) Standardize $\mathbf{X}$ to have a mean of zero and a variance of one.
2) Perform SVD according to Equation (6) and calculate $\mathbf{U}$ according to Equation (7)
3) Initialize $\mathbf{w}_u$ as $\frac{[1,1,...1]^T}{\sqrt{\text{rank}(\mathbf{V})}}$.
4) Extracting and modeling latent variables by iterating the following steps until convergence.
    Calculate $\mathbf{t} = \mathbf{U}\mathbf{w}_u$, $\mathbf{t}_h = \mathbf{U}_h\mathbf{w}_u$ and choose an proper SAR order $(p, P)$ with Hyndman-Khandakar algorithm on the initial DLVs.
    Estimate SAR's parameters $\varphi$
    Calculate $\widetilde{\mathbf{U}}_h$ and update $\mathbf{w}_u$ by solving the eigen-decomposition problem (18).
5) Deflation:
    $\mathbf{p} = \frac{\mathbf{X}^T\mathbf{t}}{\mathbf{t}^T\mathbf{t}}$
    $\mathbf{X} := \mathbf{X} - \mathbf{t}\mathbf{p}^T$
6) Go back to Step 4 to obtain the next latent variable.
```

C. RDSAR-CCA Model Relations

Assuming there are r latent variables extracted by RDSAR-CCA and the index j denotes the result of the j-th DLV extraction. The following matrices are defined as:

$$\mathbf{T} = [\mathbf{t}^{[1]}\ \mathbf{t}^{[2]}\ \cdots\ \mathbf{t}^{[r]}]$$
$$\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_r]$$
$$\mathbf{P} = [\mathbf{p}_1\ \mathbf{p}_2\ \cdots\ \mathbf{p}_r]$$

Moreover, recuring Equation (19), we have:

$$\mathbf{X}^{[r+1]} = \mathbf{X} - \sum_{j=1}^{r} \mathbf{t}^{[j]}\mathbf{p}_j^T = \mathbf{X} - \mathbf{T}\mathbf{P}^T \tag{20}$$

which is equivalent to:

$$\mathbf{X} = \mathbf{X}^{[r+1]} + \mathbf{T}\mathbf{P}^T \tag{21}$$

Based on the same geometric properties developed in [15], the relation from $\mathbf{X}$ to $\mathbf{T}$ is given as:

$$\mathbf{T} = \mathbf{X}\mathbf{R} \tag{22}$$

where $\mathbf{R} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}$. Denote $\boldsymbol{t}_k^T$ as each row of latent score matrix $\mathbf{T}$. Therefore, from Equation (22), we have:

$$\boldsymbol{t}_k = \mathbf{R}^T\boldsymbol{x}_k. \tag{23}$$

Moreover, the relation between the projection residual $\tilde{\boldsymbol{x}}_k$ and $\boldsymbol{x}_k$ is given as:

$$\boldsymbol{x}_k = \mathbf{P}\boldsymbol{t}_k + \tilde{\boldsymbol{x}}_k. \tag{24}$$

Denote $G^j(\mathcal{B})$ as the latent prediction transfer function for the j-th DLV. We can formulate them into a diagonal matrix: $\mathbf{G}(\mathcal{B}) = diag(G^1(\mathcal{B}), G^2(\mathcal{B}), ..., G^r(\mathcal{B}))$. The prediction of all r latent variables is given as:

$$\hat{\boldsymbol{t}}_k = \mathbf{G}(\mathcal{B})\boldsymbol{t}_k = \mathbf{G}(\mathcal{B})\mathbf{R}^T\boldsymbol{x}_k. \tag{25}$$

The prediction of $\boldsymbol{x}_k$ is obtained by $\mathbf{P}\hat{\boldsymbol{t}}_k$ and the corresponding one-step prediction error can be expressed as:

$$\mathbf{e}_k = \boldsymbol{x}_k - \mathbf{P}\hat{\boldsymbol{t}}_k = \boldsymbol{x}_k - \mathbf{P}\mathbf{G}(\mathcal{B})\boldsymbol{t}_k. \tag{26}$$

III. Case Study

This section aims to evaluate the efficacy of the proposed algorithm through experimentation on a dataset collected from the Shenzhen metro system in 2013. Shenzhen metro system had five lines and 118 stations. The dataset was collected from the automatic fare collection (AFC) system. In this study, we processed the original data into 15-minute interval passenger outflow data. The data covers 42 days, from October 14 to November 24, 2013, between 6:00 and 23:30 daily. Therefore, there are 66 time slots in one day and six weeks in the whole period. This time series data contains 2772 time slots as observations and 118 stations as variables.

To extract DLVs more finely, we use the k-shape clustering method [22] to divide passenger outflows into three groups by identifying the similarity of their dynamic patterns. There are 37 stations divided into Group 1, 52 in Group 2, and 29 in Group 3. Fig. 1 illustrates the overlapped passenger flow from November 11 to November 17. Group 1 showcases a significant morning peek of passenger outflow towards work, indicating that the included stations are around the business area. Group 2 shows a significant evening peek of passenger outflow towards home after work, indicating that the included stations are around the residential area. The passenger outflows showcase a significant daily pattern and are classified according to their dynamic patterns.

In this study, 2310 data samples from the first five weeks are used for extracting and modeling DLVs, while 462 samples from the last week are used for test. We apply both DiCCA [4] and RDSAR-CCA to the same passenger flow data. The autoregressive order of DiCCA is selected as 17, 19, and 15 for groups 1, 2, and 3, making the latent residuals of training data white. The order of RDSAR-CCA (p, P) is automatically determined by the Hyndman-Khandakar algorithm, which is based on akaike information criterion (AIC) [21]. As shown in Fig. 1, the daily trend is similar to the previous one. Therefore, the season length l is selected as 66 to capture the daily patterns in passenger outflow. The determined orders of each groups are $(5, 1, 66)$, $(7, 1, 66)$ and $(4, 1, 66)$.

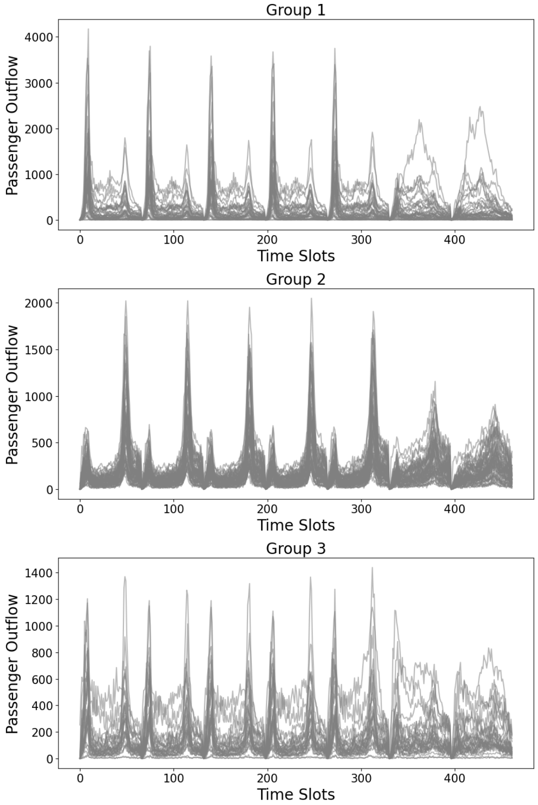

Fig. 1. The overlapped passenger flow during a week.

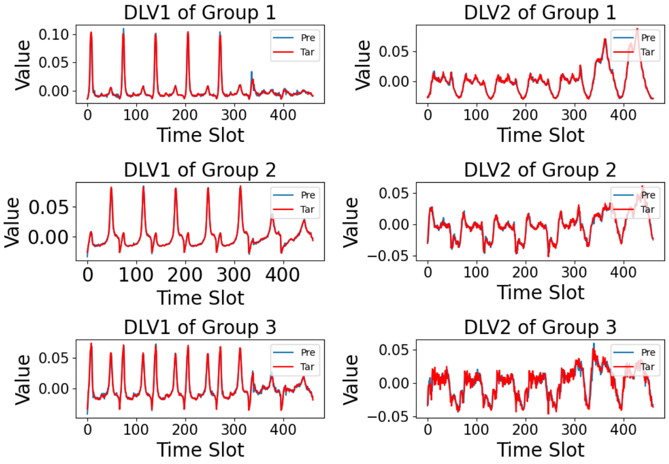

Fig. 2. The first two DLVs of each group.

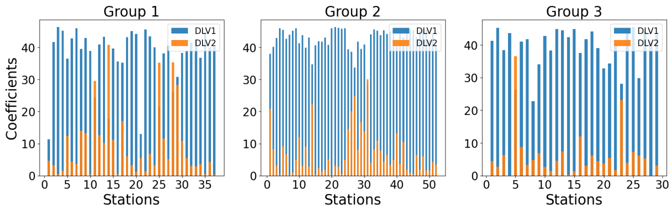

Fig. 3. The absolute value of the coefficients in $\mathbf{P}$ according to each station.

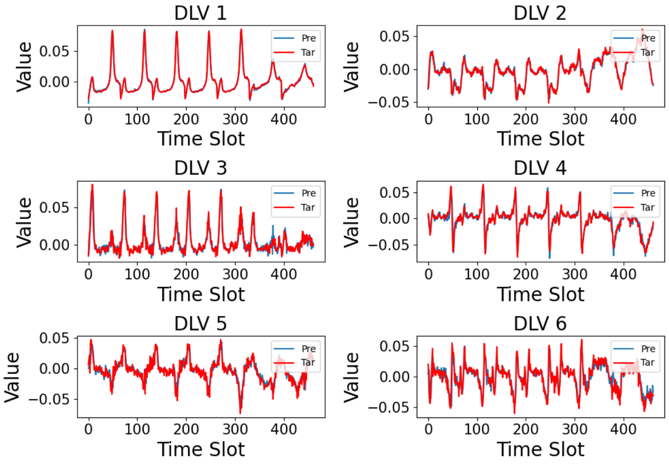

Fig. 4. The extracted DLVs of Group 2 during a week and their corresponding predictions.

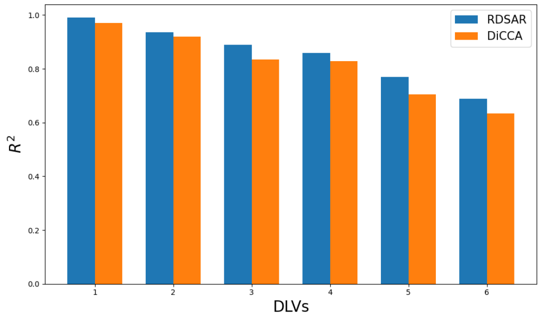

Fig. 5. R^2 value of each DLV in Group 2 on the testing dataset.

Passenger flow contains various underlying trends that exhibit different seasonality and dynamics, leading to complex modeling and challenging forecasting. These underlying trends are driven by specific dynamic patterns, which show strong dynamics. Therefore, we can extract these underlying trends by identifying DLVs of strong dynamics. RDSAR-CCA extracts DLVs by maximizing the correlations between the latent variable t_k and its self-prediction, representing the strong latent dynamics in data. Fig. 2 shows the first two DLVs extracted from each group. Compared with Fig. 1, the first DLV indicates the underlying trends during morning and evening rush hours, aligning with the peak performance of each group. While the second DLV indicates the underlying trends during non-rush hours. As shown in Fig. 3, the loading matrix $\mathbf{P}$ is used to show the contribution of DLVs to each station. Among these DLVs, DLV1 makes significant contributions to almost every station, thus representing the most common dynamic pattern in this group of stations.

Moreover, the passenger outflow exhibits significant seasonal patterns. RDSAR-CCA models each DLV by a SAR dynamic process, which matches the latent structure of seasonal high dimensional time series. For instance, Fig. 4 illustrates the six extracted DLVs of Group 2 and their corresponding predictions. The DLVs of explicit dynamics show good predictability and exhibit evident daily patterns,

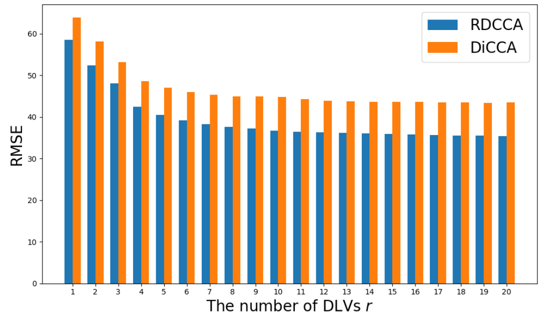

Fig. 6. RMSE(r) on testing dataset for Group 2.

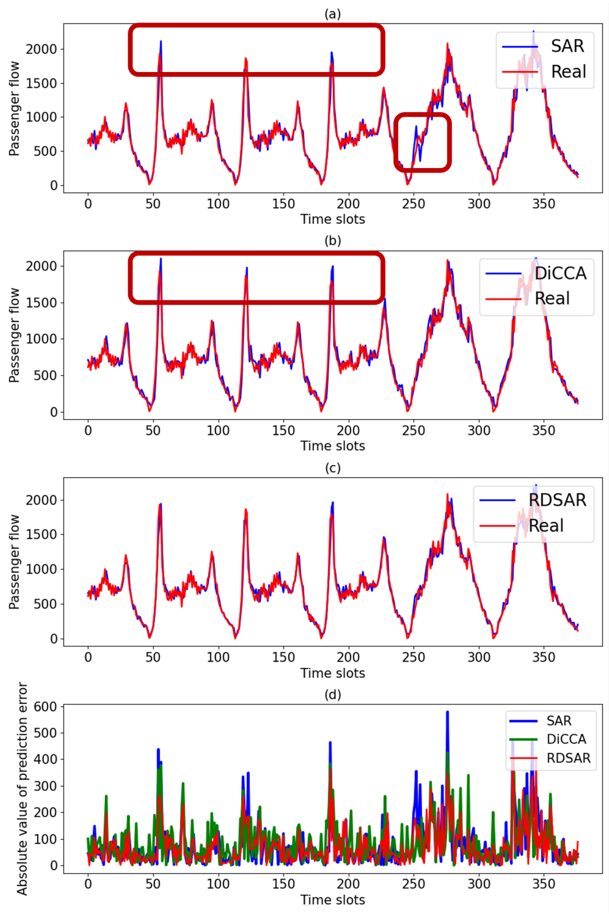

Fig. 7. Traffic flow prediction of station 'LaoJie' on the testing dataset.

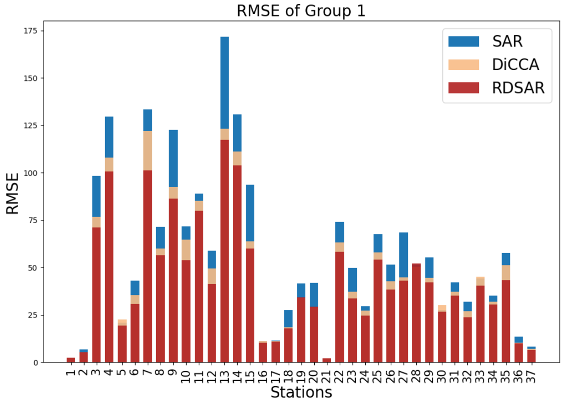

Fig. 8. Prediction error of stations in Group 1 on the testing dataset.

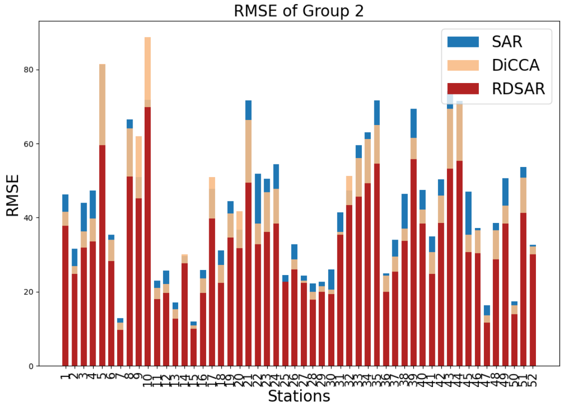

Fig. 9. Prediction error of stations in Group 2 on the testing dataset.

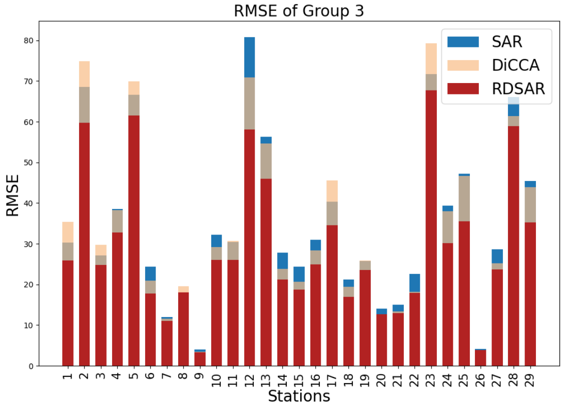

Fig. 10. Prediction error of stations in Group 3 on the testing dataset.

which align with the actual dynamic patterns of the data.

To verify that the DLVs are extracted in descending order of predictability, we calculate the R^2 value for each DLV. A higher R^2 value indicates better predictability. Fig. 5 illustrates the R^2 value of the six DLVs extracted by RDSAR-CCA and DiCCA on the testing data, indicating that RDSAR-CCA extracts DLVs in descending order of predictability and outperforms DiCCA in the one-step prediction of DLVs through considering seasonality during extraction and modeling. Moreover, we project the predicted DLVs $\hat{\boldsymbol{t}}_k$ to the original variable space through $\hat{\boldsymbol{x}}_k = \mathbf{P}\hat{\boldsymbol{t}}_k$ for passenger flow prediction of each station. The prediction error $\mathbf{e}_k$ is given in (26). We use root mean squared error (RMSE)

$$\mathrm{RMSE(r)} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\mathbf{e}_k^T\mathbf{e}_k} \tag{27}$$

to indicate the efficiency of the DLV predictive model on passenger flow prediction, where r means the number of DLVs that are used in computing $\mathbf{e}_k$. As shown in Fig. 6, the RMSE(r) is always lower than that of DiCCA on the testing data, indicating that RDSAR-CCA also achieves better prediction on the original passenger outflows. Besides, RMSE(r) decreases slowly when $r \geq 6$, indicating that after extracting 6 DLVs, the principle dynamics in data are well extracted.

For comparison, we directly apply SAR to each original group. Their order is the same as RDSAR-CCA. The prediction results of these methods in example station 'LaoJie' are shown in Fig. 7. Fig. 7 (a)-(c) show the prediction performance of these three methods, and Fig. 7 (d) shows the absolute value of the prediction error. From these charts, we can see that SAR and DiCCA always have a significant prediction error at peak times, while the prediction performance of RDSAR-CCA is superior to the other two methods for the majority of the time. The prediction errors of all the stations in each group are presented in Figs. 8, 9 and 10. The accuracy of SAR's predictions decreases due to making predictions directly on the complex dynamics of passenger outflow. DiCCA mismatches the seasonal latent structure and fails to capture seasonal patterns, making the prediction less accurate. RDSAR-CCA extracts DLVs of explicit dynamics that indicate various underlying trends in passenger outflow and models them using the SAR dynamic process. RDSAR-CCA builds the forecasting model on reduced-dimensional DLVs with explicit dynamics rather than the original time series, achieving better forecasting performance in both DLVs' one-step prediction and the original passenger flow prediction.

IV. Conclusion

In this study, a novel reduced-dimensional seasonal autoregressive modeling algorithm with a CCA objective is developed for dynamic feature extraction and prediction of high-dimensional time series with seasonality. The strong dynamics in high-dimensional time series are explicitly extracted and represented by the reduced-dimensional DLVs. The latent structure of the data is modeled through a SAR process, aligning with the actual dynamics in seasonal high-dimensional time series. Experiments on the passenger flow data from the Shenzhen metro system effectively demonstrate the superior performance of RDSAR-CCA in model prediction accuracy and dynamic feature extraction capability. The DLVs extracted by RDSAR-CCA reveal the underlying trends in original passenger flow and have an explicit dynamic model that considers the seasonal patterns in multivariate time series.

ACKNOWLEDGMENT

This work was financially supported by National Key Research and Development Program of China (2021YFA1003504); National Natural Science Foundation of China (2322816); Shenzhen-Hong Kong-Macao Science and Technology Project Fund (SGDX20210823103403028); City University of Hong Kong Project (7005889). The work described in this paper was partially supported by InnoHK initiative, The Government of the HKSAR, and Laboratory for AI-Powered Financial Technologies.

References

[1] S. J. Qin, Y. Dong, Q. Zhu, J. Wang, and Q. Liu, "Bridging systems theory and data science: A unifying review of dynamic latent variable analytics and process monitoring," *Annual Reviews in Control*, vol. 50, pp. 29–48, 2020.

[2] G. C. Reinsel, R. P. Velu, and K. Chen, *Multivariate reduced-rank regression: theory, methods and applications*. Springer Nature, 2022, vol. 225.

[3] S. J. Qin, Y. Liu, and Y. Dong, "Plant-wide troubleshooting and diagnosis using dynamic embedded latent feature analysis," *Computers & Chemical Engineering*, vol. 152, p. 107392, 2021.

[4] Y. Dong, Y. Liu, and S. J. Qin, "Efficient dynamic latent variable analysis for high-dimensional time series data," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4068–4076, 2019.

[5] G. Connor and R. A. Korajczyk, "Performance measurement with the arbitrage pricing theory: A new framework for analysis," *Journal of financial economics*, vol. 15, no. 3, pp. 373–394, 1986.

[6] S. J. Qin, Y. Liu, and S. Tang, "Partial least squares, steepest descent, and conjugate gradient for regularized predictive modeling," *AIChE Journal*, vol. 69, no. 4, p. e17992, 2023.

[7] W. K. Härdle, L. Simar, W. K. Härdle, and L. Simar, "Canonical correlation analysis," *Applied multivariate statistical analysis*, pp. 443–454, 2015.

[8] D. Peña and V. J. Yohai, "Generalized dynamic principal components," *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 1121–1131, 2016.

[9] D. Peña, E. Smucler, and V. J. Yohai, "Forecasting multiple time series with one-sided dynamic principal components," *Journal of the American Statistical Association*, 2019.

[10] S. Richthofer and L. Wiskott, "Predictable feature analysis," in *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*. IEEE, 2015, pp. 190–196.

[11] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002.

[12] C. Shang, B. Huang, F. Yang, and D. Huang, "Probabilistic slow feature analysis-based representation learning from massive process data for soft sensor modeling," *AIChE Journal*, vol. 61, no. 12, pp. 4126–4139, 2015.

[13] X. Yuan, J. Rao, Y. Wang, L. Ye, and K. Wang, "Virtual sensor modeling for nonlinear dynamic processes based on local weighted PSFA," *IEEE Sensors Journal*, vol. 22, no. 21, pp. 20 655–20 664, 2022.

[14] Y. Dong and S. J. Qin, "A novel dynamic PCA algorithm for dynamic data modeling and process monitoring," *Journal of Process Control*, vol. 67, pp. 1–11, 2018.

[15] Dong, Yining and Qin, S Joe, "Dynamic-inner canonical correlation and causality analysis for high dimensional time series data," *IFAC-PapersOnLine*, vol. 51, no. 18, pp. 476–481, 2018.

[16] Y. Dong, S. J. Qin, and S. P. Boyd, "Extracting a low-dimensional predictable time series," *Optimization and Engineering*, pp. 1–26, 2021.

[17] S. J. Qin, "Latent vector autoregressive modeling and feature analysis of high dimensional and noisy data from dynamic systems," *AIChE Journal*, vol. 68, no. 6, p. e17703, 2022.

[18] Qin, S Joe, "Latent vector autoregressive modeling for reduced dimensional dynamic feature extraction and prediction," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 3689–3694.

[19] Y. Mo, J. Yu, and S. J. Qin, "Probabilistic reduced-dimensional vector autoregressive modeling for dynamics prediction and reconstruction with oblique projections," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 7623–7628.

[20] J. Yu and S. J. Qin, "Latent state space modeling of high-dimensional time series with a canonical correlation objective," *IEEE Control Systems Letters*, vol. 6, pp. 3469–3474, 2022.

[21] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for R," *Journal of statistical software*, vol. 27, pp. 1–22, 2008.

[22] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015, pp. 1855–1870.