Factor Models for High-Dimensional Time Series Forecasting: An Application to Revenue Management

Ruixue Li, Chaofeng Yuan, Nan Wang*

School of Mathematical Science, Heilongjiang University, Harbin 150080, China
**Correspondence: 2003137@hlju.edu.cn*

Abstract

Since the 2020s, the emergence of the COVID-19 pandemic has brought the development of transportation, tourism, and entertainment industries to a standstill. The idea of revenue management (RM) improved the profitability for different types of companies. Therefore, establishing a good RM model and accurate mathematical forecasting model is particularly important for struggling airlines. We herein propose a factor model based on high-dimensional time series that can efficiently use continuous time historical data and the related environmental historical data to predict the passenger load factor. Therefore, accurate and effective dimensional reduction and feature expression of high-dimensional matrix time series have profound practical significance for studying time series data. To verify the efficacy of the model and parameter estimation methods, we applied them to the booking rates of 11 flights over 365 days (year 2018). After experimental analysis and comparison tests with other methods studied in the paper has the best effect and the results of comparisons with different dimensions indicate that the error rate of the proposed method is less than 0.1.

Keywords: *time series; revenue management; forecasting; flight booking rates*

1. Introduction

Recently, high-dimensional time series models have achieved success in a wide range of applications in numerous fields. In finance, economics, medicine, and many other fields, as time data acquisition becomes more convenient, people often use matrix data to explore laws under time changes. The idea of revenue management (RM) has been popular since the 1980s, because RM improved the profitability for different types of companies [1,2]. The revenue management system (RMS), which comprises RM, pursues the maximization of revenue and builds a robust optimization algorithm through prediction value iteration [3,4].

For RM models, data including predetermined historical quantities can be used for each forecast point [5]. Therefore, we herein propose a factor model based on high-dimensional time series that can efficiently use continuous time historical data and the related environmental historical data to predict the passenger load factor.Traditional load factor forecasting tends to focus only on the departure day load factor results and performance evaluation, whereas our proposed model can predict the load factor of the day and the purchase rate of seats a few days prior.

2. Algorithm Design

2.1. The Factor Model

Let Y_t be a p × q observable matrix-valued time series, G_t' be a k3 × q observable covariate matrix-valued time series, and Ft be an unobservable matrix with $k_1 \times k_2$ dimensions. It is assumed that Y_t is generated by

$$Y_t = RF_tC' + \Gamma G_t' + E_t \tag{1}$$

where R and C are unknown $p \times k_1$ and $q \times k_2$ dimensional matrices of unknown parameters, and $E_t = \{e_{ij}^t\}$ is a p × q dimensional zero-mean white-noise sequence matrix.

2.2. Time Series Model

In Model (1), common fundamental factors of Ft drive the latent dynamics and co-movement of Y_t, while R and C reflect the importance of common factors and their interactions [6]. Let $H_t = [CF_t' \ G_t]' = \{h_{ij}^t\}$, $i = 1,\dots,(k_1 + k_3)$, $j = 1,\dots,q$. We further assume that the VAR(K) process given by

$$H_t = \Phi_1 \circ H_{t-1} + \cdots + \Phi_K \circ H_{t-K} + U_t \tag{2}$$

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 1247–1251, 2023.
https://doi.org/10.2991/978-94-6463-042-8_178

where {Φ_k}, k=1,...,K, are $(k_1 + k_3) \times q$ dimensional auto coefficient parameter matrices, and $U_t = \{u_{ij}^t\}$ is a $(k_1 + k_3) \times q$ dimensional zero-mean white-noise sequence matrix. The following assumptions pertain.

3. Experimental Results and Analysis

3.1. Comparison Between Different Methods

We compared our method with several existing methods for processing time series matrix data, including Wang et al.'s [6] method, Yu et al.'s [7] method, and linear regression (LR) (Table 1). Case 1 lets $k_1 = 1$, $k_2 = 1$, $k_3 = 2$, and H = 1 and the random seed is fixed; the process is iterated 200 times to explore the influence of p, q, and T on the model. In Table 1, "Our" method denotes the MSE of our estimated value and the true value, "Yu" method denotes the MSE of Yu et al.'s [7] estimated value and true value, and "Wang" method denotes the MSE of Wang et al.'s [6] estimated value and true value. Further, the columns corresponding to Γ, LR-Γ, R and C are all R^2 values between the estimated and true values. In this case, the latent factor matrix Ft is regarded as a number, and R and C are vectors.

Table 1. Comparison of the results of our method with those of existing methods.

p	q	T	MSE				R^2			
			Our	LR	Wang	Yu	Γ	LR-Γ	R	C
20	20	10	0.992	2.846	0.980		0.985	0.938	0.9961	0.9964
20	20	20	0.991	5.973	0.999	0.993	0.979	0.944	0.9994	0.9996
20	20	50	0.986	9.301	0.999	0.995	0.960	0.874	0.9999	0.9999
20	20	100	0.986	6.786	0.998	0.996	0.981	0.966	0.9999	0.9999
10	20	50	0.995	5.576	0.997	0.991	0.971	0.894	0.9998	0.9996
20	20	50	0.986	3.184	0.999	0.995	0.977	0.940	0.9995	0.9997
50	20	50	0.982	3.136	1.001	0.998	0.982	0.944	0.9994	0.9999
100	20	50	0.992	3.343	1.001	0.998	0.984	0.956	0.9994	0.9999
20	10	50	0.992	3.288	0.997	0.992	0.958	0.912	0.9982	0.9997
20	20	50	0.986	2.925	0.999	0.995	0.984	0.952	0.9993	0.9996
20	50	50	0.983	4.725	1.001	0.998	0.992	0.970	0.9999	0.9996
20	50	50	0.983	4.725	1.001	0.998	0.992	0.970	0.9999	0.9996

Table 1 shows that the MSE of the three methods showed different trends with increasing T. The MSE of our method shows a downward trend, and the proposed method achieves a better estimation effect. The MSE of Wang stabilizes at approximately 0.999 and the MSE of Yu is increasing. The accuracy of Yu's method slightly decreased. In general, Wang's method has the maximum MSE. When T = 10, Yu's method is superior to our method. When T > 10, our method is superior to Yu's method. Between the initial value LR−Γ and the estimated value Γ, the Γ value calculated by the algorithm has a better fitting effect, but the estimate of Γ does not change with an increase in T. The prediction effect of R and C improves with increasing T. With an increase in p and q, the MSE of Yu and Wang showed a rising trend, and our method showed a downward trend and then an upward trend. When T = 50, the effect of our model is estimated to be the best. The overall order of the estimated excellent effect is Yu < Wang < Our. When p increases, the estimation of C improves from 0.9996 to 0.9999, and the estimation accuracy of R decreases slightly from 0.9998 to 0.9994. When q increases, the estimation of R gradually improves from 0.9982 to 0.9999, whereas the estimation accuracy of C decreases slightly from 0.9997 to 0.9996.

3.2. Application to Flight Data

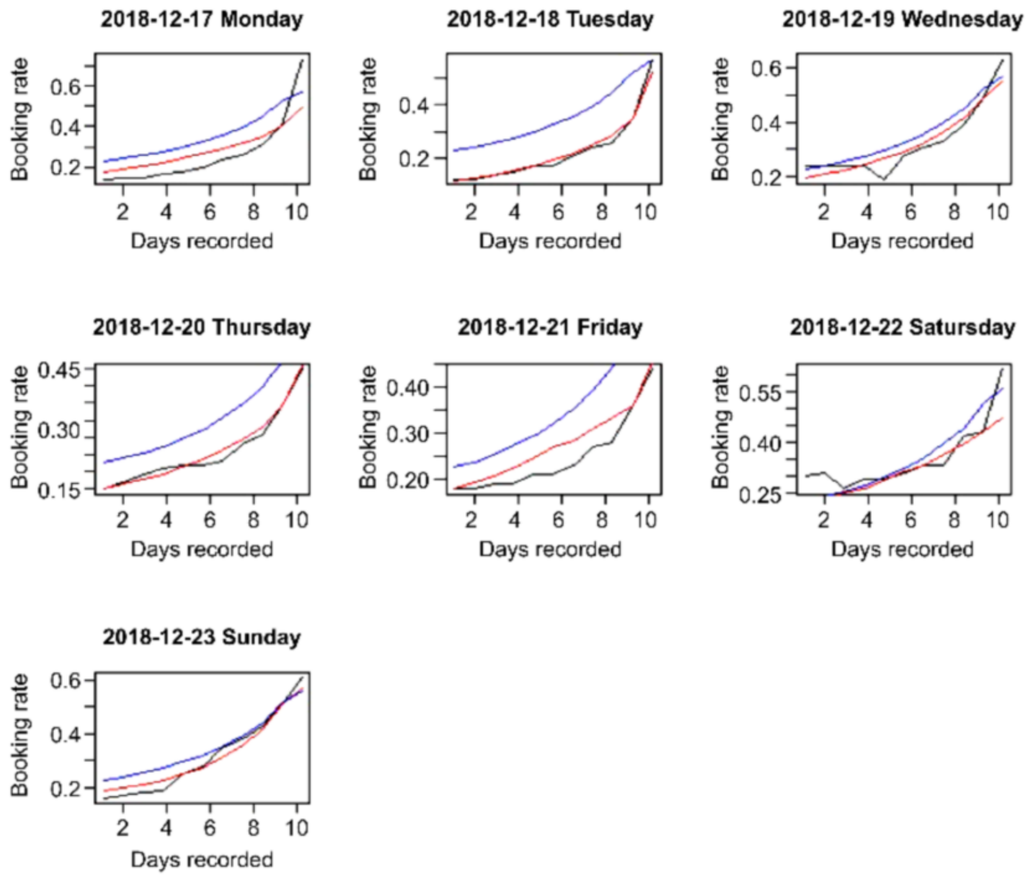
In this section, we show the practical application of the model to airline booking rate forecasting using real data from airlines. The booking rate data of the flight are stored in an SQL server database. The database contains the flight number, collection date, passenger load factor, model, number of seats sold, and other related information for each flight. The booking rate of 11 flights in 2018 was selected as the basic data, which were collected 365 days. Data were collected for the following flights: CA1684, CA1692, CA1698, CZ 684, CZ6213, CZ6225, CZ6318, CZ6482, CZ6657, CZ6658, and MU5614. Consider a certain flight as an example, such as CA1684. Because the booking rate information is collected four times per day in the database, the data collected at the latest time of the day is the booking rate for that day.

A single piece of data is continuous time series data. Matrix data collection started on January 1, 2018, and ended on December 31, 2018 (Table 2), and the predictor variables of 11 flights formed a matrix to construct a time series matrix. Here, "day0" denotes the booking rate on the day of departure, "dayn" denotes the booking rate n days before departure, and Y_t is the booking rate information of the 11 flights on day t. The data are presented in Table 2.

Table 2. Format of the original data Yt, taking January 1, 2018 as an example.

Time	CA1684	CA1692	CA1698	CZ684	CZ6213	CZ6225	CZ6318	CZ6482	CZ6657	CZ6658	MU5614
day0	0.94	0.52	0.47	0.64	0.90	0.92	0.84	0.64	0.89	0.78	0.74
day1	0.80	0.52	0.50	0.46	0.83	0.90	0.86	0.59	0.91	0.59	0.63
day2	0.59	0.43	0.52	0.44	0.69	0.83	0.75	0.54	0.66	0.46	0.59
day3	0.47	0.43	0.55	0.46	0.52	0.69	0.64	0.60	0.71	0.47	0.60
day4	0.38	0.38	0.47	0.39	0.43	0.64	0.60	0.47	0.68	0.37	0.55
day5	0.33	0.36	0.38	0.42	0.34	0.53	0.50	0.46	0.64	0.33	0.49
day6	0.24	0.38	0.34	0.44	0.35	0.44	0.43	0.49	0.61	0.50	0.42
day7	0.19	0.32	0.29	0.42	0.33	0.45	0.43	0.49	0.58	0.39	0.45
day8	0.16	0.30	0.26	0.42	0.32	0.38	0.41	0.41	0.46	0.32	0.46
day9	0.08	0.32	0.21	0.47	0.30	0.35	0.40	0.44	0.41	0.24	0.45
day10	0.08	0.33	0.20	0.43	0.28	0.29	0.39	0.43	0.40	0.26	0.40

Because our model and Yu's model are both based on high-dimensional time series, we compared the matrix fitting effects of the two models. The effect map of the one-week forecast for flight CA1684 is presented in Figure 1.


2018-12-17 Monday

2018-12-18 Tuesday

2018-12-19 Wednesday

2018-12-20 Thursday

2018-12-21 Friday

2018-12-22 Saturday

2018-12-23 Sunday

Figure 1. One-week booking rate forecast for flight CA1684.

The data in the seven plots in Figure 1 (the title of each figure part indicates the year-month-day and the corresponding day of the week) were extracted from the load factor of flight CA1684 predicted using two different methods. Each set of data contains the booking rate from 10 days before departure to the day of departure, and the fitting trend achieved the desired effect. The black line represents the original data, the red line represents our method, and the blue line represents Yu's method. Model of this report is highly sensitive to sudden changes in data and has a good fitting effect on changing data. Thus, the model is suitable for practical cases where data change rapidly.

To reflect the accuracy of the model's estimation at a single point, we compared two model with the ARIMA (Autoregressive Integrated Moving Average Model, one of the most common statistical models used for time series forecasting) composed of single points. In Figure 2, the horizontal axis represents the prediction day 1-15, and the vertical axis represents the percentage of the error rate on the nth day of the three prediction methods to the total error rate.

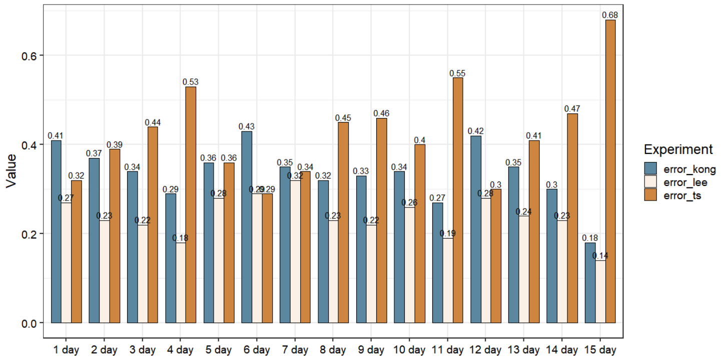


Figure 2. Comparison chart of data error rate percentages.

In Figure 2, the blue bars denote the percentage of error rate of Yu's method based on the overall error rate, white bars denote the percentage of error rate of our method, and brown bars denote the percentage of error rate of the time series ARIMA method. The error rates of our method are lower than those of the other two methods. In general, the error rate changes as follows: Our < Yu < ARIMA; among them, there are also cases in which the time series method is better than Yu's method. It includes our method, Yu's method, and the ARIMA time series method. Three methods corresponding to the error rates of different flights were extracted, and a box plot was generated, as shown in Figure 3.

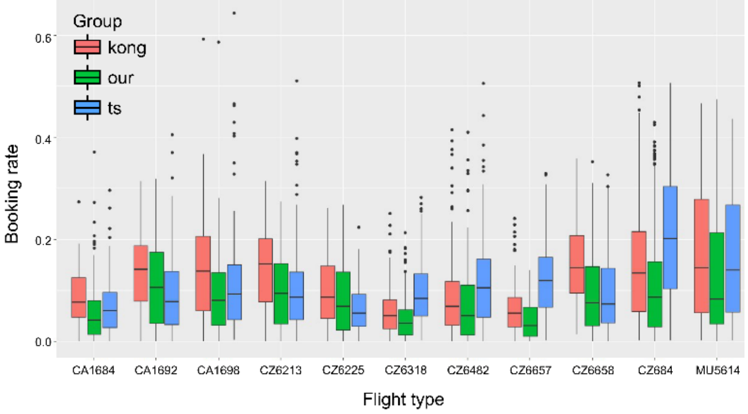


Figure 3. Error rates of flight forecast.

Figure 3 shows that the accuracy of the models is in the following order: our method > ARIMA time series and our method > Yu. ARIMA and Yu's methods have their own advantages and disadvantages; however, Yu's method directly uses a matrix to predict, which makes the time shorter. ARIMA must string the time points into a vector and then make the prediction, which takes more time. The error rate of our method was lower than that of the other two methods.

4. Conclusions

We proposed a method to explore and predict high-dimensional time series using a matrix dimensionality reduction model that fully utilizes known data. The proposed method uses a traditional dimensionality reduction matrix and projection estimation methods to reasonably split the original data. This offers a solution to the problem of dimensionality in high-dimensional data settings. We refer to the FAVAR model and generalize it to a matrix model to accelerate the speed of data processing. By selecting a small number of available relevant factors, we can fit a multivariate time series model to the overall trend and then make more detailed adjustments, allowing us to build the dependence structure of the latent factor model and dynamically estimate the data at each point. We conducted data simulations and proved that the model is stable under different parameters. Then, we used the model to empirically study the booking rate of 11 flights over 365 days. The results of the empirical research show that the model is robust. Compared with the existing forecasting approaches, the proposed method performs well in both simulations and empirical data analysis.

References

[1] Relihan III, W.J. The Yield-Management Approach to Hotel-Room Pricing. Cornell Hotel Restaur. Admin. Q. 1989, 30, 40–45. DOI:10.1177/001088048903000113.

[2] Choi, S.; Kimes, S.E. Electronic Distribution Channels' Effect on Hotel Revenue Management. Cornell Hotel Restaur. Admin. Q. 2002, 43, 23–31. DOI:10.1177/0010880402433002.

[3] Schwartz, Z. The Confusing Side of Yield Management: Myths, Errors, and Misconceptions. J. Hosp. Tourism Res. 1998, 22, 413–430. DOI:10.1177/109634809802200406.

[4] Phillips, R.L. Pricing and Revenue Optimization, 2nd ed.; Stanford University Press: Redwood City, US, 2021; pp. 20–66.

[5] Schwartz, Z.; Uysal, M.; Webb, T.; Altin, M. Hotel daily occupancy forecasting with competitive sets: a recursive algorithm. Int. J. Contemp. Hosp. Manag. 2016, 28, 267–285. DOI:10.1108/IJCHM-10-2014-0507.

[6] Wang, D.; Liu, X.; Chen, R. Factor Models for Matrix-Valued High-Dimensional Time Series. J. Econ. 2019, 208, 231–248. DOI:10.1016/j.jeconom.2018.09.013.

[7] Yu, L.; He, Y.; Yu, X.; Zhang, X. Projected estimation for Large-Dimensional Matrix Factor Models. Journal of Econometrics, 2021.