ARIMA Forecasting Chinese Macroeconomic Variables Based on Factor and Principal Component Backdating

Wei Wang and Yan Liu*

School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, P.R. China

Abstract: In this paper the backdating methods based on factors and principal components are applied for the first time to emulate the historical macroeconomic variables in China. The numerical results show that these procedures are useful to backdate some missing or not available historical data. ARIMA forecasting experiments based on backdated historical data are conducted and compared with forecasting procedures using directly factors and principal components. Our results suggest that some key variables like GDP can indeed be forecasted more precisely with the principal components backdated data.

Keywords: Backdating, Factor model, Principal components, ARIMA forecasting, GDP of China.

1. INTRODUCTION

With the implementation of five-year plan in China, studies on the behavior of macroeconomic variables have been paid more and more attention by economists and policy makers. Tracking and predicting the tendencies of some important macroeconomic variables are especially what researchers interested in. To guarantee the accuracy and effectiveness of these analyses, enough data of time series are usually required. A fundamental statistical approach to handle longitudinal data is time series analysis method. A popular model is autoregressive integrated moving average (ARIMA) model, which was first proposed by Box and Jenkins (1968) [1] and had been further studied in a series of works of Box and Pierce (1970) [2], Box *et al*. (1974) [3] and Bartholomew (1976) [4]. Saboia (1977) [5] used this model to forecast female-birth data for Norway during the years 1976-2000 with time series of the years 1919-1974. Ledolter (1979) [6] first investigated the sensitivity of ARIMA models to study non-normality of the distribution of the shock price data. Very recently, Garg *et al*. (2015) [7] analyzed traffic noise levels by modeling the data of day-night average sound level with AMIRA process. Time-series methods have been frequently applied to the studies of complex and dynamic economic phenomena and for an overview readers are referred to the book of Tsay (2010) [8]. Recently Corte *et al*. (2010) [9] and Bianco *et al*. (2012) [10] showed that fundamentals-based econometric models obtain statistically significant improvements upon random-walk model when they are modeling certain macroeconomic variables such as short-term exchange rate. Xiao *et al*. (2014) [11] studied financial market volatility by establishing a multiscale ensemble forecasting model which combines ARIMA with feed forward neural network (FNN). ARIMA model is used to generate a linear forecast, and FNN is developed as a tool for nonlinear pattern recognition to correct the estimation error in ARIMA forecast. These studies motivate us to construct effective models of time series to forecasting some important macroeconomic variables of China.

It should be mentioned that in practice some historical data may be insufficient because of varieties of reasons such as missing or incomplete statistics. To solve this problem, Angelini *et al*. (2006) [12] first put forward a factor-backdating method to construct historical data. Factors from some macroeconomic variables relative to interest were extracted and then a linear model between interest and the factors was established to backdate interest data before Germany was unified in 1991. Factor-backdating method has great ability to handle the situations in which time series data are missing in some of the cross-section units and this advantage makes the procedure more effective and useful. Based on this method, Angelini *et al*. (2011) [13] and Anderson *et al*. (2011) [14] analyzed separately historical financial data for the Euro area. Brüggemann and Zeng (2015) [15] further investigated the effectiveness of backdating method in forecasting a number of macroeconomic Euro-area variables by contrasting predicting data obtained by this method with that by autoregressive (AR) models and logistic smooth transition auto regression (LSTAR) models. To the best of our knowledge, there is no literature about analysis of backdating method to study the macroeconomic variables of China.

This paper aims to analyze the behavior of macroeconomic variables of China by using backdating

*Address correspondence to this author at the School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, P.R. China;
Tel: (86)27-68752957; Fax: (86)27-68752256; E-mail: yanliu@whu.edu.cn

ISSN: 1814-8085 / E-ISSN: 1927-5129/17

methods combined with classical simulation models. Take GDP, a key variable, as an example to expand in-depth. Methods based on factors and principal components are used separately to simulate the historical GDP data. Short-term GDP data are also forecasted by embedding backdating method in traditional ARIMA time series models. All the estimated data are compared with the real ones. Throughout the paper, the set of indicators includes Electricity production (EP), Railway freight traffic (RFT), Index of raw materials supply(RMI), Retail sales of consumer goods (CGR), as discussed in Fernald, Hsu, and Spiegel (2015) [16]. In addition to most frequently used variable GDP, Financial institutions deposits (FID). Financial institutions loans (FIL) and National financial expenditure (NFE) are also analyzed. All the procedures are achieved by R statistical software.

The paper is constructed as follows. Section 2 introduces related backdating models based on factors and principal components and all the obtained backdated GDP data before 1999 are compared to the real ones. ARIMA forecasting experiments based on factor and principal component backdated data are conducted in Section 3. Predicting GDP data after 2010 are obtained and compared to the real ones. Section 4 concludes the major results of the paper.

2. BACKDATING METHODS AND EVALUATION

2.1. Factor-Based Backdating

Exploratory factor analysis (EFA) is used frequently to gain insights into latent structure underlying obtained data. In this section the factor-based procedures to backdate GDP historical data for China are introduced and a detailed description is given as follows.

As in Stock and Watson (2002a, b) [17, 18], we suppose that $X_s^t\ (s=1,2,\ldots,p)$ is an original variable, $X_t=\left(X_1^t,X_2^t,\ldots,X_p^t\right)'$ is a *p*-dimensional time series represented by an unobserved common factor F_t and an idiosynacratic component e_t, where $F_t=\left(F_1^t,F_2^t,\ldots,F_k^t\right)'$ is a *k*-dimensional vector, $e_t=\left(\varepsilon_1^t,\varepsilon_2^t,\ldots,\varepsilon_p^t\right)'$. Then the vector of time series are written as

$$X_t=\mu+\Lambda F_t+e_t,\ t=1,2,...,N,$$

where X_t is a $p\times 1$ vector, Λ is a $p\times k$ matrix, F_t is the $k\times 1\left(k\le p\right)$ vector of common factors and e_t, is a $p\times 1$ vector of idiosyncratic components independent of F_t. Here Λ is called factor loading matrix. Denote $X=\left(X_1,X_2,\ldots,X_N\right)'$, $F=\left(F_1,F_2,\ldots,F_N\right)'$, and $e=\left(e_1,e_2,\ldots,e_N\right)'$ when the number of samples is N, and assume that

$$E(X)=\mu,\ D(X)=\Sigma,$$

$$\mathrm{E}(F)=\mathbf{0},\ D(F)=\begin{pmatrix}1&&&\\&1&&\\&&\ddots&\\&&&1\end{pmatrix}=\mathrm{I}_k,$$

$$\mathrm{E}(e)=\mathbf{0}\ ,\ D(e)=\begin{pmatrix}\sigma_1^2&&&\\&\sigma_2^2&&\\&&\ddots&\\&&&\sigma_N^2\end{pmatrix},\ COV(F,e)=\mathbf{0}.$$

The common factors are first extracted from the time series data before the backdate procedure. In the factor model approach, we estimate factor model using maximum likelihood estimation (MLE). Suppose that there are *k* common factors from *p* original variables, and then we get

$$F_j^t=\alpha_{j1}^tX_1^t+\alpha_{j2}^tX_2^t+\cdots+\alpha_{jp}^tX_p^t,\ j=1,2,\cdots,k,\ t=1,2,\cdots N,$$

where the parameters $\alpha_{j1}^t,\alpha_{j2}^t,\cdots,\alpha_{jp}^t$, $j=1,2,\cdots,k$, $t=1,2,\cdots N$, are estimated by Bartlett factor score method due to the condition that *k* ≤ *p*.

In our application, we split the entire sample period (from 1995 to 2015) into a backdating period (from T_0 to T_1), a real period (from T_1 + 1 to T_2), and a forecasting evaluation period (from T_2 + 1 to T_3). In the following, we set T_0 to 1995, T_1 to 1998,T_2 to 2010 and T_3 to 2015 which means that we will backcast the GDP data from 1995 to 1998 by backdating methods. In the first step, we make the initial indicators stationary by two order difference operation and then extract factors from data of 4 variables including EP, RFT, RMI and CGR before 2011. In the second step, we relate the factor time series to the macroeconomic series of GDP from T_1+1 to T_2 by a regression model, denoted by y_t, as the dependent variables and the estimated factor time series $\hat{F}_j^t$ as explanatory variables. That is, we use the model

$$y_t=\beta_0+\beta_1\hat{F}_1^t+\cdots+\beta_k\hat{F}_k^t+\varepsilon_t\ ,\ t\in\left[T_1+1,T_2\right] \tag{2.1}$$

and estimate the parameters $\beta_0,\beta_1,\cdots,\beta_k$ by ordinary least square (OLS). Here the macroeconomic series of GDP are also stationary after using two order difference operation. In the third step of our procedure, we backdate the GDP time series for period before 1999 by

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1\hat{F}_1^t + \cdots + \hat{\beta}_k\hat{F}_k^t \ , \ t \in \left[T_0, T_1\right] \tag{2.2}$$

The backdated data results of $\hat{y}_t$ and the related comparisons are provided in Section 2.3.

2.2. Principal Component-Based Backdating

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables which are called principal components. The number of principal components is less than or equal to the number of original variables. PCA was first introduced by Pearson (1901) [19] and has been mostly used as a tool in exploratory data analysis and for making predictive models.

In this subsection, prior to the backdate procedure we first use PCA to extract *k* principal components from the data set of four indicators including EP, RFT, RMI, CGR. Suppose that at time *t* the total number of economic indicators is *p* and we denote these indicators as *p* random variables $X_1^t, X_2^t, \ldots, X_p^t$.

Firstly, we extract principal components, $G_1^t, G_1^t, \cdots, G_p^t$, by the model

$$\begin{cases} G_1^t = u_{11}X_1^t + u_{21}X_2^t + \cdots + u_{p1}X_p^t \\ G_2^t = u_{12}X_1^t + u_{22}X_2^t + \cdots + u_{p2}X_p^t \\ \qquad\cdots \\ G_p^t = u_{1p}X_1^t + u_{2p}X_2^t + \cdots + u_{pp}X_p^t \end{cases}$$

where variables G_i^t and coefficients u_{ji}, *i,j* = 1,2,⋯, *p*, satisfying that:

1. For each principal component the sum of the square of coefficients is equal to 1:

$$\sum_{j=1}^{p} u_{ji}^2 = 1 \ , \ i = 1,2,\cdots,p \ ,$$

2. The principal components are independent of each other:

$$COV(G_i^t, G_j^t) = 0 \ \ i \neq j \ , \ i,j = 1,2,\cdots,p$$

3. The variance of the principal components satisfying

$$D(G_1) \geq D(G_2) \geq \cdots \geq D(G_p)$$

Denote the orthonormal sample matrix as X_t, and $COV(X_t)$ as Σ. Then we derive a orthogonal matrix U such that

$$U^{'}\Sigma U = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_p$ are *p* characteristic roots of Σ, and those related *p* characteristic vectors compose the principal components $G_1^t, G_1^t, \cdots, G_p^t$. Then, we can extract *k* (*k* ≤ *p*) principal components $G_1^t, G_1^t, \cdots, G_k^t$ by cumulative contribution $C_k = \sum_{i=1}^{k}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i$ satisfying that $C_k \geq 85\%$, that is, these *k* principal components account for 85% contribution to the total sample variance. Therefore, we obtain time series set of G_j^t, *j* = 1,2,⋯ ,*k, t* = 1,2,⋯*N.*

Relating the principal components time series to the macroeconomic series of GDP from T_1+1 to T_2 by a regression using series of GDP of China, denoted by y_t, as the dependent variable and the estimated principal components time series $\tilde{G}_j^t$ as explanatory variables, we get the model

$$y_t = \beta_0 + \beta_1\tilde{G}_1^t + \cdots + \beta_k\tilde{G}_k^t + \varepsilon_t \ , t \in \left[T_1 + 1, T_2\right], \tag{2.3}$$

and estimate the parameters $\beta_0,\beta_1,\cdots,\beta_k$ by ordinary least square (OLS). Finally, we backdate GDP time series $\tilde{y}_t$ by

$$\tilde{y}_t = \hat{\beta}_0 + \hat{\beta}_1\tilde{G}_1^t + \cdots + \hat{\beta}_k\tilde{G}_k^t \ , \ t \in \left[T_0, T_1\right] \tag{2.4}$$

The polynomial and exponential fitting methods for principal components time series $\tilde{G}_j^t$ are also investigated in the following of this subsection.

In the polynomial fitting model, we rewrite estimated principal components time series $\tilde{G}_j^t$ as $\tilde{G}_j^{t(1)}$ and represent the model as

$$\tilde{G}_j^{t^{(1)}} = \gamma_{j0}^{(1)} + \gamma_{j1}^{(1)}(t-t_0) + \cdots + \gamma_{jn}^{(1)}(t-t_0)^n,\ j=1,2,\cdots,k,\quad t\in[T_0,T_2] \tag{2.5}$$

where $\gamma_{j0}^{(1)},\gamma_{j1}^{(1)},\cdots,\gamma_{jn}^{(1)}$, $j=1,2,\cdots,k$, are parameters to be estimated. Here t_0 is set to1994 and n is set to 2. Then, we run a regression like equation (2.3) in which time series of GDP are dependent variables and estimated principal components $\tilde{G}_j^{t^{(1)}}$ are explanatory variables. Finally, we backdate following equation (2.4) and then obtain the backdated GDP which are written as $\tilde{y}_t^{(1)}$.

In parallel, we denote the estimated principal components time series $\tilde{G}_j^t$ as $\tilde{G}_j^{t^{(2)}}$ and introduce the exponential regression model as below

$$\tilde{G}_j^{t^{(2)}} = \gamma_{j0}^{(2)} + \gamma_{j1}^{(2)}\exp\{\gamma_{j2}^{(2)}(t-t_0)\},\ j=1,2,\cdots,k,\ t\in[T_0,T_2] \tag{2.6}$$

where $\gamma_{j0}^{(2)}$, $\gamma_{j1}^{(2)}$, $\gamma_{j2}^{(2)}$, $j=1,2,\cdots,k$ are parameters to be estimated. Then we rerun backdating procedures of (2.3) and (2.4) using $\tilde{G}_j^{t^{(2)}}$ as explanatory variables. The backdated data of GDP are denoted accordingly as $\tilde{y}_t^{(2)}$.

The backdated data results of $\tilde{y}_t, \tilde{y}_t^{(1)}, \tilde{y}_t^{(2)}$ and the related comparisons are provided in Section 2.3.

2.3. Backdating Comparisons

In this subsection, the numerical results of $\hat{y}_t, \tilde{y}_t, \tilde{y}_t^{(1)}$ and $\tilde{y}_t^{(2)}$ obtained in Section 2.1 and Section 2.2 and the corresponding comparisons with the real ones are provided. Recall that we split the entire sample period (from 1995 to 2015) into a backdating period (from 1995 to 1998), a real period (from 1999 to 2010), and a forecasting evaluation period (from 2011 to 2015). We will derive the GDP data from 1995 to 1998 by backdating methods introduced in Section 2.1 and Section 2.2 and then compare them to the real ones.

In EFA model we first extract 1 factor and then get the backdating GDP $\hat{y}_t$ by (2.2). In PCA-based backdating we also extract 1 principal component from the set of data of the four indicators. And then we compute backdated GDP $\tilde{y}_t$ by PCA model, $\tilde{y}_t^{(1)}$ by polynomial fitting and $\tilde{y}_t^{(2)}$ by exponential fitting. All of the backdated data are given in Table **1**.

In addition, following above procedures backdated variables of FID, FIL, NFE can also be computed. Here we just provide RMSE in Table **2**.

Tables **1** and **2** suggest that backdating methods works well to emulate historical macroeconomic variables for China. Generally speaking, backdating based on classical PCA provides most precise data and this method is followed closely by that based on

Table 1: Backdating GDP from 1995 to 1998

Year	Real value	EFA	Principal	based	model
		$\hat{y}_t$	$\tilde{y}_t$	$\tilde{y}_t^{(1)}$	$\tilde{y}_t^{(2)}$
1995	60794	80424	58192	53251	33177
1996	71177	91529	67204	53977	41883
1997	78973	74449	70461	57790	51752
1998	84402	87638	62810	64690	62940

Note: The unit of the data in Table **1** is billion yuan.

Table 2: RMSE of Backdating Variables from 1995 to 1998

	EFA	Principal	based	model
	$\hat{y}_t$	$\tilde{y}_t$	$\tilde{y}_t^{(1)}$	$\tilde{y}_t^{(2)}$
GDP	0.2184	0.1432	0.2238	0.3739
FID	0.2314	0.2076	0.3785	0.7138
FIL	0.1544	0.1671	0.2389	0.4361
NFE	0.4084	0.2146	0.4120	0.7801

factors. Meanwhile, traditional curve fitting approach to choose principal components leads to higher error relative to real data.

3. FORECASTING METHODS AND EVALUATION

The effectiveness of forecasting based on factors and PCA is investigated in this section. Prior to these two forecasting procedures, we first use traditional ARIMA time series model to get prediction because of the rising tendency of GDP, which is described in detail in Section 3.1. Forecasting methods based on factors and PCA are introduced separately in Section 3.2 and Section 3.3. The data results and corresponding comparisons are given in Section 3.4.

3.1. Autoregressive Integrated Moving Average (ARIMA)

It is well known that ARIMA time series model has the form

$$y_{t+h} = f(Z_t;\theta_{ht}) + \varepsilon_{t+h},$$

where Z_t is the vector of explanatory variables, θ_{ht} is a vector of possibly time-varying parameters and ε_t is the error of this model. We focus on forecasting GDP in China h periods ahead, denoted as y_{t+h}, where h represents the forecasting horizon. The h-step ahead forecast is given by

$$\hat{y}_{t+h} = f(Z_t;\hat{\theta}_{ht}),$$

where the unknown parameter vector θ_{ht} is estimated by $\hat{\theta}_{ht}$, and the h-step forecast error is

$$e_{t+h} = y_{t+h} - \hat{y}_{t+h},$$

where the forecasting horizon is chosen as h = 1,2,3,4 and 5.

The widely used ARIMA (p,d,q) model is defined as

$$\begin{cases} \Phi(B)\nabla^d y_t = \Theta(B)\varepsilon_t \\ \mathrm{E}(\varepsilon_t) = 0,\ D(\varepsilon_t) = \sigma_a^2 \\ E(\varepsilon_t\varepsilon_s) = 0,\ s \neq t \\ E(y_t\varepsilon_t) = 0,\ \forall s < t \end{cases}$$

where B is the backward shift operator satisfying that $By_t = y_{t-1}$, $B^j y_t = y_{t-j}$, $\nabla^d = (1-B)^d$, d is the number of times for difference operation, and $\{\varepsilon_t, t = 1,2,\cdots\}$ is a sequence of independently distributed random variables with mean zero and variance σ_a^2 (white noise). ARIMA(p,d,q) model is the product of an autoregressive part AR(p)

$$\Phi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$

an integrating part

$$I(d) = \nabla^{-d}$$

and a moving average MA(q) part

$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta^q B^q .$$

An example of an ARIMA model is the ARIMA(1,0,0) model, first order autoregressive model, which is used recently to forecast the EMU aggregate of some variables of interest h periods ahead by Brüggemann and Zeng (2015) [15].

In this subsection we study the problem of minimum variance prediction for stationary time series. In the first step, stationary property of the time series y'_t, t=1,2,..., is checked and a stationary sequence of GDP data, y'_t, y'_{t-1}, ..., is obtained by using difference operators if it is not stationary. It turns out that in our case GDP time series become stationary after difference operations twice. In the second step, a white noise test is applied to the stationary sequence and ARMA model is used to simulate the model if it is not a white noise. In the third step, we predict y'_{t+h} using y'_t, y'_{t-1}, y'_{t-2} .... In view of the fact that

$$(y'_{t+h} \mid y'_t, y'_{t-1}, y'_{t-2}, \cdots) \sim N(\mathrm{E}(y'_{t+h}), D(y'_{t+h}))$$

we can choose the conditional expectation of y'_{t+h} as its prediction value, that is

$$\hat{y}'_{t+h} = \mathrm{E}(y'_{t+h} \mid y'_t, y'_{t-1}, y'_{t-2}, \cdots) .$$

Since stationary series $\{y'_t,\ t = 1,2,\cdots\}$ can be written as

$$y'_{t+h} = \varphi_1 y'_{t+h-1} + \varphi_2 y'_{t+h-2} + \cdots + \varphi_p y'_{t+h-p} + \varepsilon_{t+h} - \theta_1\varepsilon_{t+h-1} \\ -\theta_2\varepsilon_{t+h-2} - \cdots - \theta_q\varepsilon_{t+h-q}$$

we obtain that when $h \leq \max\{p,q\}$,

$$\hat{y}'_{t+h} = \varphi_1 \hat{y}'_{t+h-1} + \varphi_2 \hat{y}'_{t+h-2} + \cdots + \varphi_{h-1}\hat{y}'_{t+1} + \varphi_h y'_t + \cdots + \varphi_p y'_{t+h-p} \\ -\theta_1\varepsilon_t - \theta_2\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t+h-q}$$

and when $h > q$,

$$\hat{y}'_{t+h} = \varphi_1 \hat{y}'_{t+h-1} + \varphi_2 \hat{y}'_{t+h-2} + \cdots + \varphi_p \hat{y}'_{t+h-p} .$$

Table 3: Forecasting Results for GDP of ARIMA Models

	Model	2011	2012	2013	2014	2015
	Real value	471563.7	519322.0	588019.0	636463.0	676708.0
$\hat{y}'_t$	ARIMA (0,2,1)	452516.8	503520.8	554524.7	605528.7	656532.7
	ARIMA (1,2,1)	447063.3	497420.1	546242.9	595555.3	644711.5
$\hat{y}_t$	ARIMA (0,2,1)	450026.5	498540.1	547053.8	595567.4	644081.1
	ARIMA (1,2,1)	437255.7	490099.8	531183.4	580354.7	623964.1
$\tilde{y}_t$	ARIMA (0,2,1)	449827.6	498142.4	546457.2	594771.9	643086.7
	ARIMA (1,2,1)	444731.4	493429.2	540400.8	587916.2	635260.3
$\tilde{y}_t^{(1)}$	ARIMA (0,2,1)	451637.2	501761.6	551886.0	602010.4	652134.8
	ARIMA (1,2,1)	447880.0	497800.7	546834.8	596090.1	645290.2
$\tilde{y}_t^{(2)}$	ARIMA (0,2,1)	451105.4	500698.1	550290.7	599883.3	649476.0
	ARIMA (1,2,1)	446893.6	496441.2	544848.8	593568.3	642202.5

Note: The unit of the data in Table **3** is billion yuan.

In this ARIMA (*p,d,q*) model, we first use the real historical GDP data from 1995 to 2010 to predict data from 2011 to 2015 and the results are denoted by $\hat{y}'_t$. And then we use the backdating data from 1995 to 1998 obtained in Section 2 and real data from 1999 to 2010 to forecast the future data.

The forecasted data results based on ARIMA, factor backdated data, PCA backdated data, PCA with polynomial fitting backdated data and PCA with exponential fitting backdated data are denoted respectively by $\hat{y}'_t$, $\hat{y}_t$ $\tilde{y}_t$, $\tilde{y}_t^{(1)}$, $\tilde{y}_t^{(2)}$ and the related comparisons are provided in Table **3**.

3.2. Factor-Based Forecasting

In this subsection, we discuss the forecasting method based on factors. As in the procedure of backdating, we first extract *k* factors from four original variables covering the entire period from T_0 to T_3. And then we relate the factor time series to the macroeconomic series of GDP from T_0 to T_2, that is, we construct a regression model like (2.1) using series of GDP as the dependent variable and the estimated factor time series as explanatory variables for $t \in [T_0, T_2]$. Finally we forecast GDP data following equation (2.2) for $t \in [T_2 + 1, T_3]$.

The forecasted data results of $\hat{y}_t$, and the related comparisons are provided in Table **4**.

3.3. Principal Component-Based Forecasting

In this subsection, we discuss the forecasting method based on PCA. Firstly, we extract *k* principal components denoted by G_j^t from the set of four indicators from 1995 to 2015. Secondly, we run a regression between GDP and principal components time series $\tilde{G}_j^t$. The model has the same form as (2.3) for $t \in [T_0, T_2]$. Finally, we forecast GDP time series for period after 2010 following equation (2.4) for $t \in [T_2 + 1, T_3]$.

We also try to fit G_j^t by polynomial and exponential functions and denote the emulated values as $\tilde{G}_j^{t(1)}$ and $\tilde{G}_j^{t(2)}$ which have the same form as (2.5) and (2.6) respectively for $t \in [T_0, T_3]$. And linear regression models are established as in Section 3.2 in which y_t is dependent variable and $\tilde{G}_j^t, \tilde{G}_j^{t(1)}$ and $\tilde{G}_j^{t(2)}$ are the independent variables.

The forecasted data results of $\tilde{y}_t$, $\tilde{y}_t^{(1)}$, $\tilde{y}_t^{(2)}$ and the related comparisons are provided in Table **4**.

3.4. Forecasting Comparisons

This subsection investigates the efficiencies of forecasting methods discussed in Section 3.1-3.3. Recall that in ARIMA models, we first use real value of GDP from 1995 to 2010 to make a prediction. Then we use backdated GDP data from 1995 to 1998 obtained respectively by factor-model and by three principal-component-models introduced in Section 2 and real GDP data from 1999 to 2010 to forecast GDP after 2010. The numerical results are provided in Table **3**.

Table 4: Forecasting Results of GDP from 2011 to 2015

Year	Real value	Factor	Principal	based	model
		$\hat{y}_t$	$\tilde{y}_t$	$\tilde{y}_t^{(1)}$	$\tilde{y}_t^{(2)}$
2011	471564	341947	441845	432364	452847
2012	519322	376238	464537	481879	519874
2013	588019	353520	498408	534518	596718
2014	636463	359493	507046	590282	684814
2015	676708	331012	500211	649170	785811

Table 5: RMSE of Forecasting Results of ARIMA Models with Different Historical Data

Historical data	GDP	FIL	FID	NFE
$\hat{y}_t'$	0.04225^{0}	0.02886^{1}	0.03441^{0}	0.06034^{1}
$\hat{y}_t$	0.05488^{0}	0.02828^{1}	0.03216^{1}	0.05719^{1}
$\tilde{y}_t$	0.05576^{0}	0.02788^{1}	0.03932^{0}	0.07073^{1}
$\tilde{y}_t^{(1)}$	0.04681^{0}	0.02792^{1}	0.03213^{0}	0.04113^{1}
$\tilde{y}_t^{(2)}$	0.04941^{0}	0.02799^{1}	0.03334^{0}	0.04337^{1}

Note: We take q=1,d=2 and denote the value of p at the top-right corner of the data.

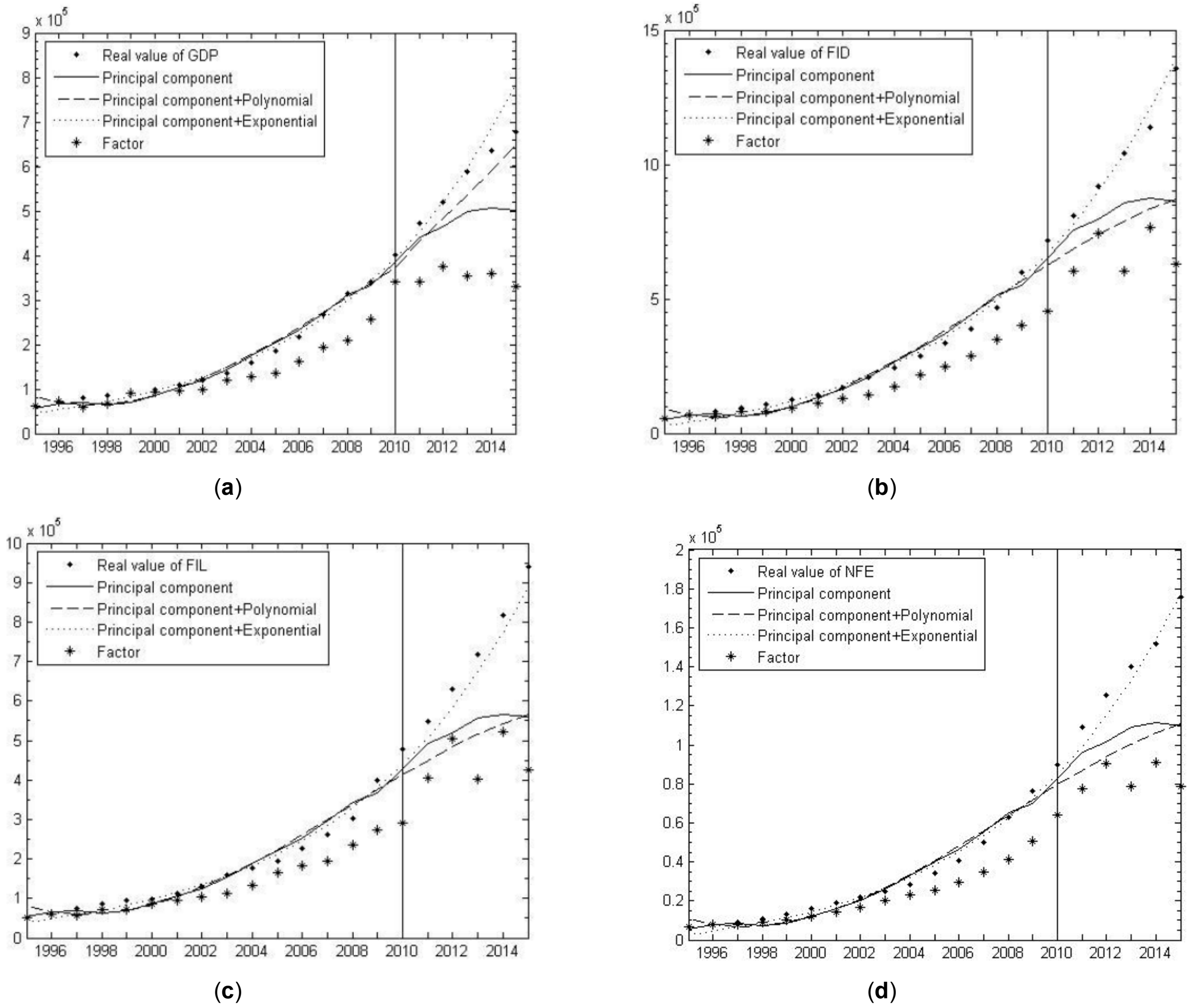
Table **3** shows that ARIMA models with backdated GDP data (from 1995 to 2010) based on PCA perform better than that based on EFA. Meanwhile FIL，FID and NFE are also forecasted in the same way and corresponding RMSE are shown in Table **5**. We conclude that historical data based on PCA provide more precise forecasting than that based on EFA and forecasting data perform especially well when principal components time series are described with polynomial and exponential functions. It is interesting that forecasting data of FIL and NFE with historical data based on PCA combined with polynomial and exponential function fitting even predict more precisely than that based on real historical data.

The numerical results for forecasted GDP based on factors and PCA are given in Table **4** and corresponding RMSE for forecasting GDP, FIL, FID and NFE are computed in Table **6**. To see more clearly, we also draw prediction curves yielded by forecasting methods discussed in Section 3.2 and Section 3.3, see Figure **1**. It follows that forecasting methods based on factor and PCA are practicable to emulate future macroeconomic data. Compared to the real data from 2011 to 2015, the data forecasted by PCA provides more efficiency than that obtained by factors. In particularly principal components described with exponential functions provide the optimal forecasting data.

It follows from Tables **3** and **4** that real data are usually higher than forecasted data. The practice has showed that implementation of five-year plan is impactful to promote economic growth in China. Comparisons in Tables **5** and **6** suggest that

Table 6: RMSE of 5-Step Forecasting Variables from 2011 to 2015

Var.	EFA	Principal	based	model
	$\hat{y}_t$	$\tilde{y}_t$	$\tilde{y}_t^{(1)}$	$\tilde{y}_t^{(2)}$
GDP	0.39009	0.17187	0.07391	0.08193
FID	0.36726	0.21926	0.25326	0.03486
FIL	0.38260	0.26452	0.29610	0.06344
NFE	0.40552	0.24658	0.28683	0.05856



(**a**)

(**b**)

(**c**)

(**d**)

Figure 1: Forecasting curves for four variables based on factors and PCA models.

forecasting data based on ARIMA model and PCA with principal components described with exponential functions perform better than that obtained by other methods.

4. CONCLUSION

In this paper the backdating procedures based on factors and PCA are introduced for the first time to emulate historical China macroeconomic time series data. Our numerical results for GDP, FID, FIL, NFE data illustrate that they are effective methods to handle the situation where some historical data are missing or not available in the desired quality. ARIMA forecasting experiments based on the factor-backdated and PCA-backdated data are conducted and compared with forecasting procedures based directly on factors and PCA. Our results suggest that some key variables like real GDP can indeed be forecasted more precisely with PCA backdated data. Overall, our results indicate that for some important macroeconomic variables the backdating procedure based on factors or PCA is a valuable method to construct time series data for China.

ACKNOWLEDGEMENTS

The authors would like to thank anonymous referees for their thoughtful comments on an earlier version of this paper. This work was supported by the National Natural Science Foundation of China (No. 11001209 and No. 11371284).

REFERENCES

[1] Box GEP, Jenkins GM. Some Recent Advances in Forecasting and Control Part I. J R Stat Soc Ser C Appl Stat 1968; 17(2): 91-109.

[2] Box GEP, Pierce DA. Distribution of residual autocorrelations in auto regressive integrated moving average time series models. Journal of the American Statistical Association 1970; 65(332): 1509-1526.

[3] Box GEP, Jenkins GM, MacGregor JF. Some recent advances in forecasting and control Part II. J R Stat Soc Ser C Appl Stat 1974; 23(2): 158-179.

[4] Bartholomew D, Box GEP, Jenkins GM. Time Series Analysis: Forecasting and Control. Holden-day Series in Time Series Analysis, Revised Ed, San Francisco: Holden-Day, Incorporated 1976; 199-201.

[5] Saboia JLM. Autoregressive integrated moving average (ARIMA) models for birth fore- casting. Journal of the American Statistical Association 1977; 72(358): 264-270. https://doi.org/10.1080/01621459.1977.10480989

[6] Ledolter J. Inference robustness of ARIMA models under non-normality—special application to stock price data. Metrika 1979; 26(1): 43-56. https://doi.org/10.1007/BF01893469

[7] Garg N, Soni K, Saxena TK, Maji S. Applications of Auto Regressive Integrated Moving Average (ARIMA) approach in time-series prediction of traffic noise pollution. Noise Control Eng J 2015; 63(2): 182-194. https://doi.org/10.3397/1/376317

[8] Tsay, Ruey S. Analysis of financial time series. 3rd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ 2010. https://doi.org/10.1002/9780470644560

[9] Corte PD, Sestieri G. The predictive information content of external imbalances for exchange rate returns: How much is it worth? Review of Economics & Statistics 2011; 94(1): 100-115. https://doi.org/10.1162/REST_a_00157

[10] Bianco MD, Camacho M, Quiros GP. Short-run forecasting of the euro-dollar exchange rate with economic fundamentals. J Int Money Finance 2012; 31(2): 377-396. https://doi.org/10.1016/j.jimonfin.2011.11.018

[11] Xiao Y, Xiao J, Liu J, Wang SY. A multiscale modeling approach incorporating ARIMA and ANNS for financial market volatility forecasting. Journal of System Science & Complexity 2014; 27(1): 225-236. https://doi.org/10.1007/s11424-014-3305-4

[12] Angelini E, Henry J, Marcellino M. Interpolation and backdating with a large in-formation set. J Econ Dyn Control 2006; 30(12): 2693-2724. https://doi.org/10.1016/j.jedc.2005.07.010

[13] Angelini E, Marcellino M. Econometric analyses with backdated data–unified Germany and the Euro area. Econ Model 2011; 28(3): 1405-1414. https://doi.org/10.1016/j.econmod.2011.02.002

[14] Anderson HM, Dungey M, Osborn DR, Vahid F. Financial integration and the costruction of historical financial data for the Euro area. Econ Model 2011; 28(4): 1498-1509. https://doi.org/10.1016/j.econmod.2011.02.027

[15] Brüggemann R, Zeng J. Forecasting Euro-Area macroeconomic variables using a factor model approach for backdating. Oxf Bull Econ Stat 2015; 77(1): 22-39. https://doi.org/10.1111/obes.12053

[16] Fernald J, Hsu E, Spiegel MM. Is China Fudging its Figures? Evidence from Trading Partner Data. Federal Reserve Bank of San Francisco Working Paper 2015-12. http://www.frbsf.org/economic-research/publications/working-papers/wp2015-12.pdf.

[17] Stock JH, Watson MW. Macroeconomic forecasting using diffusion indexes. J Bus Econ Stat 2002a; 20(2): 147-162. https://doi.org/10.1198/073500102317351921

[18] Stock JH, Watson MW. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 2002b; 97(460): 1167-1179. https://doi.org/10.1198/016214502388618960

[19] Pearson K. On lines and planes of closest fit to systems of points in space. The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science 1901; 559-572.

Received on 17-06-2016 Accepted on 22-02-2017 Published on 12-04-2017

https://doi.org/10.6000/1927-5129.2017.13.16