ASYMPTOTIC EXPANSION OF RISK OF FORECASTING OF AUTOREGRESSIVE TIME SERIES WITH MISSING DATA

A.S. Huryn
Belarusian State University
Minsk, BELARUS
e-mail: Huryn@bsu.by

Abstract

The problem of forecasting of autoregressive time series with missing data is considered in the paper and the asymptotic expansion of the risk of forecasting under estimated parameters is constructed.

1 Introduction

The autoregressive model is widely used in practice for the statistical analysis and forecasting of time series in various applications: economy, communication, meteorology, geology, environmental protection, astronomy and many others [3, 5, 6, 9]. The time series in this applications are often observed with missing data [8]. There are two main sources of missing data. Firstly, some data can not be registered in certain time moments, for example because they do not exist or there are measurement problems (holidays, failures of measurement device) [2]. Secondly, some data are ignored by experts because of unreliability or low level of trust (noises, outliers, heteroscedasticity) [7].

This paper is devoted to the construction of statistical forecasts of autoregressive time series with missing data and to the evaluation of the risks of forecasting.

2 Mathematical model

Let the observed time series $Y_t \in \mathbb{R}^d$ be defined in time moments $t \in \mathbb{Z}$ and satisfy the vector autoregressive model VAR(1) of the dimension $d \in \mathbb{N}$ on the probabilistic space (Ω, F, P) [1]:

$$Y_t = BY_{t-1} + U_t, t \in \mathbb{Z}, \tag{1}$$

where $B \in \mathbb{R}^{d\times d}$ is a matrix parameter of autoregression, all eigenvalues of the matrix B are inside the unit circle, $U_t \in \mathbb{R}^d, t \in \mathbb{Z}$, is a vector innovation process defined on (Ω, F, P), the random vectors $\{U_t : t \in \mathbb{Z}\}$ are independent in the aggregate and have normal distribution: $\mathcal{L}\{U_t\} = \mathcal{N}(0_d, \Sigma), t \in \mathbb{Z}, |\Sigma| \neq 0$.

There are missing values in the data. In order to describe the structure of missing values let define the vector missing pattern $O_t \in \mathbb{R}^d, t \in \mathbb{Z}$, by the following way:

$$O_{ti} = \begin{cases} 1, & \text{if } Y_{ti} \text{ is observed} \\ 0, & \text{if } Y_{ti} \text{ is not observed} \end{cases}, t \in \mathbb{Z}, i \in \{1, \ldots, d\}.$$

Let define the minimal and the maximal time moments with the observed components:

$$t_- = \mathbf{Min}\left\{t \in \mathbb{Z} : \sum_{i=1}^{d} O_{ti} > 0\right\}, t_+ = \mathbf{Max}\left\{t \in \mathbb{Z} : \sum_{i=1}^{d} O_{ti} > 0\right\},$$

without loss of generality $t_- = 1$, $t_+ = T \in \mathbb{N}$. If $O_t = 1_d, t \in \{1, \ldots, T\}$, then there are no missing values and the classical observation model takes place [1].

The vector time series, defined by (1), is stationary and has zero mathematical expectation $\mathbf{E}\{Y_t\} = 0_d, t \in \mathbb{Z}$, the covariance matrix $\mathbf{Cov}\{Y_t, Y_t\} = G_0 \in \mathbb{R}^{d\times d}, t \in \mathbb{Z}$, and the covariance function $\mathbf{Cov}\{Y_{t+\tau}, Y_t\} = G_\tau \in \mathbb{R}^{d\times d}, t, \tau \in \mathbb{Z}$, where

$$G_0 = \sum_{i=0}^{+\infty} B^i \Sigma (B')^i,$$

$$G_\tau = B^\tau G_0 I_{\tau \geq 0} + G_0 (B')^{-\tau} I_{\tau < 0}, \tau \in \mathbb{Z}. \quad (2)$$

The problem is to construct the statistical forecasts of the future values $Y_{T+\tau} \in \mathbb{R}^d$ with depth $\tau \in \mathbb{N}$ of the vector autoregressive time series VAR(1) (1) given the observed vector time series $Y_1, \ldots, Y_T$ with missing pattern $O_1, \ldots, O_T$ and to evaluate the risk of forecasting.

3 Statistical estimators of parameters

Let denote that the following equations take place for the model (1):

$$G_1 = BG_0, G_0 = BG_1' + \Sigma.$$

Assume that the missing pattern $O_t, t \in \{1, \ldots, T\}$ satisfies the assumption:

$$\sum_{t=1}^{T-k} O_{t+k,i} O_{tj} > 0, k \in \{0, 1\}, i, j \in \{1, \ldots, d\},$$

construct the sample covariances $\hat{G}_k \in \mathbb{R}^{d\times d}, k \in \{0, 1\}$:

$$(\hat{G}_k)_{ij} = \frac{\sum_{t=1}^{T-k} Y_{t+k,i} Y_{tj} O_{t+k,i} O_{tj}}{\sum_{t=1}^{T-k} O_{t+k,i} O_{tj}}, k \in \{0, 1\}, i, j \in \{1, ..., d\}, \quad (3)$$

and, if $|\hat{G}_0| \neq 0$, construct the estimators of parameters of the vector autoregressive model with missing data using the sample covariances (3):

$$\hat{B} = \hat{G}_1 \hat{G}_0^{-1}, \hat{\Sigma} = \hat{G}_0 - \hat{G}_1 \hat{G}_0^{-1} \hat{G}_1' \in \mathbb{R}^{d\times d}. \quad (4)$$

The asymptotic properties (consistency and asymptotic normality) of estimators (4) can be found in [4].

4 Statistical forecasting

Let denote that the optimal in the maximum likelihood sense one step predictor of the vector autoregressive model (1) under totally observed last vector ($O_T = 1_d$) is [4]

$$\hat{Y}_{T+1} = BY_T.$$

If the model parameters B, Σ are unknown let construct the so called "plug-in" forecasting statistic using the given by (4) estimator $\hat{B}$:

$$\hat{Y}_{T+1} = \hat{B}Y_T. \tag{5}$$

The following theorem gives the asymptotic expansion of the matrix risk

$$R = \mathbf{E}\left\{\left(\hat{Y}_{T+\tau} - Y_{T+\tau}\right)\left(\hat{Y}_{T+\tau} - Y_{T+\tau}\right)'\right\} \in \mathbb{R}^{d\times d} \tag{6}$$

of the forecasting statistic (5).

Theorem 1. *Let the model (1) takes place, the last vector is observed without missing values ($O_T = 1_d$), the one step ($\tau = 1$) forecasting statistic (5) is used under estimated by (4) parameters, the missing pattern satisfies the assumption:*

$$\frac{\sum_{t=1}^{T-k} O_{t+k,i}O_{tj}}{T-k} \xrightarrow[T\to+\infty]{} \vartheta_{k,i,j} \in (0,1],$$

$$\frac{\sum_{t,t'=1}^{T-1} O_{t+k,i}O_{tj}O_{t'+k',i'}O_{t'j'}\delta_{t-t',\tau}}{T-|\tau|-1} \xrightarrow[T\to+\infty]{} \tilde{\vartheta}_{\tau,k,k',i,j,i',j'} \in [0,1].$$

and the following assumption takes place:

$$\forall n \in \mathbb{N}, \exists C \in (0,+\infty), \exists T_0 \in \mathbb{N},$$

$$\forall T \geq T_0, \forall \theta \in (0,1) \qquad \mathbf{E}\left\{\frac{1}{\left|I_d + \theta G_0^{-1}\left(\hat{G}_0 - G_0\right)\right|^{2n}}\right\} \leq C. \tag{7}$$

Then the asymptotic expansion of the matrix risk (6) takes place under $T \to +\infty$:

$$R = \Sigma + \frac{1}{T}A + o\left(\frac{1}{T}\right)1_{d\times d}, \tag{8}$$

where

$$A = \sum_{i,j,i',j'=1}^{d}\sum_{\tau=-\infty}^{+\infty} BI(i,j)\left(BI(i',j')G_0^{-1}\right)' \times \left((G_\tau)_{ii'}(G_\tau)_{jj'} + (G_\tau)_{ij'}(G_\tau)_{ji'}\right)C_{\tau,0,0,i,j,i',j'}$$

$$- \sum_{i,j,i',j'=1}^{d} \sum_{\tau=-\infty}^{+\infty} \Big(BI(i,j) \left(I(i',j') G_0^{-1} \right)' + I(i',j') \left(BI(i,j) G_0^{-1} \right)' \Big)$$
$$\times \left((G_{\tau-1})_{ii'} (G_\tau)_{jj'} + (G_\tau)_{ij'} (G_{\tau-1})_{ji'} \right) C_{\tau,0,1,i,j,i',j'}$$

$$+ \sum_{i,j,i',j'=1}^{d} \sum_{\tau=-\infty}^{+\infty} I(i,j) \left(I(i',j') G_0^{-1} \right)'$$
$$\times \left((G_\tau)_{ii'} (G_\tau)_{jj'} + (G_{\tau+1})_{ij'} (G_{\tau-1})_{ji'} \right) C_{\tau,1,1,i,j,i',j'} \in \mathbb{R}^{d \times d},$$

$$C_{\tau,k,k',i,j,i',j'} = \frac{\tilde{\vartheta}_{\tau,k,k',i,j,i',j'}}{\vartheta_{k,i,j} \vartheta_{k',i',j'}}, \tau \in \mathbb{Z}, k, k' \in \{0,1\}, i, j, i', j' \in \{1, ..., d\},$$

and $I(i,j) \in \mathbb{R}^{d \times d}$, $(I(i,j))_{kl} = \begin{cases} 1, & \text{if } (k,l) = (i,j) \\ 0, & \text{if } (k,l) \neq (i,j) \end{cases}, i, j, k, l \in \{1, \ldots, d\}.$

References

[1] Anderson T. (1976). *The Statistical Forecasting of Time Series*. Mir, Moscow. (in Russian)

[2] Brubaker S.R., Wilson G.T. (1976). Interpolating Time Series with Application to the Estimation of Holiday Effects on Electricity Demand. *Applied Statistics*. Vol. **25**, pp. 107–116.

[3] Hsiao C. (1979). Autoregressive Modelling of Canadian Money and Income Data. *Journal of the American Statistical Association*. Vol. **74**, pp. 553–560.

[4] Kharin Yu.S., Huryn A.S. (2005). "Plug-in" Statistical Forecasting of Vector Autoregressive Time Series with Missing Values. *Austrian Journal of Statistics*. Vol. **34**, n. **2**, pp. 163–174.

[5] Landers T.E., Lacoss R.T. Some Geophysical Applications of Autoregressive Spectral Estimates. sl IEEE Transactions on Geoscience Electronics. Vol. **15**, pp. 26–32.

[6] Resnick S. (1997). Heavy Tail Modeling and Teletraffic Data. *Annals of Statistics*. Vol. **25**, pp. 1805–1869.

[7] Rousseeuw P.J., Leroy A.M. (1987). *Robust Regression and Outlier Detection*. Chapman and Hall, London.

[8] Schafer J.L. (1977). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

[9] Tong H. (1977). Some Comments on the Canadian Lynx Data. *Journal of the Royal Statistical Society. Series A*. Vol. **140**, pp. 432–436.