Journal of Econometrics 176 (2013) 80–91

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Adaptively combined forecasting for discrete response time series

Xinyu Zhang[a,*], Zudi Lu[b], Guohua Zou[a]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Mathematical Sciences, The University of Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 6 February 2012
Received in revised form
5 April 2013
Accepted 22 April 2013
Available online 30 April 2013

JEL classification:
C25
C53

Keywords:
Adaptation
Discrete response
Forecast combination
Model screening
Time series

ABSTRACT

Adaptive combining is generally a desirable approach for forecasting, which, however, has rarely been explored for discrete response time series. In this paper, we propose an adaptively combined forecasting method for such discrete response data. We demonstrate in theory that the proposed forecast is of the desired adaptation with respect to the widely used squared risk and other significant risk functions under mild conditions. Furthermore, we study the issue of adaptation for the proposed forecasting method in the presence of model screening that is often useful in applications. Our simulation study and two real-world data examples show promise for the proposed approach.

1. Introduction

The time series data with discrete response exist widely in many research fields, including economics, finance, health, and even sports. For example, Dueker (1999) and Monokroussos (2011) used this kind of time series to describe the US monetary policy; Müller and Czado (2005, 2009) utilized it in the study of financial analysis; Kedem and Fokianos (2002) applied it to modeling the sleep state of a newborn infant and Akhtar and Scarf (2012) adopted it for the prediction of match outcomes in test cricket. For the surveys on early and recent developments of time series models with discrete responses, the reader is referred to Eckstein and Wolpin (1989) and Aguirregabiria and Mira (2010), respectively. Obviously, an accurate forecast for discrete response time series data is significantly desirable. This paper will be devoted to developing an effective procedure for combining forecasts in the context of time series with discrete responses.

As is well known, different estimation or forecasting procedures may generally perform well in different cases. However, in practice, it is often very hard to choose out the best procedure, even though a large number of model selection approaches exist in the literature. Furthermore, model selection is often unstable in the sense that small change in data may lead to a significant difference in the chosen models, and thus cause an unnecessarily high variability in the final estimation/prediction. Therefore, a combination of candidate procedures is highly desirable. In addition, a combined forecast avoids ignoring useful information from the relationship between the response and the covariates and also provides a kind of insurance against selecting a very poor candidate model. We refer to Bates and Granger (1969), Zou and Yang (2004) and Leung and Barron (2006), among others, for further discussions.

Various combination methods have been suggested for forecasting in the literature. In the classical forecasting combination (cf., Bates and Granger, 1969; Granger and Ramanathan, 1984), combining weights are typically selected based on the estimated variances of individual forecast errors. The resulting combined forecast by this kind of procedures, however, lacks of theoretical supports. Combining procedures based on the scores of information criteria such as AIC and BIC (Buckland et al., 1997) are also commonly used in practice, but they need the maximum likelihood values of all candidate models fitted. Recently, asymptotically optimal combining approaches have attracted a lot of attention and various procedures have been proposed. Examples include Mallows model averaging (MMA) by Hansen (2007, 2008) and Wan et al. (2010), optimal mean squared error averaging by Liang et al. (2011), and Jackknife model averaging (JMA) by Hansen and Racine (2012). But all the work on asymptotically optimal combining procedures consider to average the linear estimators. Differently, this

* Corresponding author.
E-mail address: xinyu@amss.ac.cn (X. Zhang).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.04.019

paper will develop an adaptive combination procedure,[1] which is applicable in a more general framework because it neither restricts the form of the estimators/forecasts averaged nor requires the likelihood values fitted. Also, as pointed out by a referee, any forecasting procedure (e.g., MMA) can be included in the candidate set of the adaptive combination procedure so that the final risk can adaptively achieve minimax rates for multiple scenarios.

In recent literature, adaptive forecasting studies are focused on continuous random variables. Yang (2004) proposed an adaptive algorithm, called aggregated forecast through exponential reweighting (AFTER). Zou and Yang (2004) used it to combine time series models (e.g., ARIMA) and indicated the advantage of AFTER over some commonly used model selection approaches. Since then, the AFTER algorithm has been applied to a variety of forecasting issues, such as the US employment growth (Rapach and Strauss, 2008) and the exchange rate (Altavilla and De Grauwe, 2010). To the best of our knowledge, however, the adaptive property has seldom been investigated for discrete response time series, which will be studied in this paper. We will not only establish the adaptive property of our proposed combination procedure under usual squared risk, but also demonstrate that the proposed combined forecast procedure enjoys the adaptation under other significant general risk functions, including, for example, the asymmetric LINEX loss function. In addition, we will consider the adaptation of the proposed method in the presence of model screening that is often useful in applications. The advantages of the proposed approaches will be illustrated by both simulation study and real-world data examples.

The remainder of this paper is structured as follows. Section 2 begins with the setup of the problem and combined forecast. Section 3 contains theorems on the adaptation of the proposed combined forecast based on the squared risk and other important risk functions. Section 4 further presents its adaptation by adding a model screening step to the combining procedure. Sections 5 and 6 report results from the simulation study and real-world data analysis, respectively. Section 7 concludes. The technical proofs are relegated to an Appendix.

2. Problem setup and combined forecast

Suppose that we are interested in forecasting a discrete response variable Y at some time t $(=1, 2, \ldots)$, taking on $D+1$ categories. We denote by $\widetilde{G}_0$ the initial information set available at time $t=0$ and by $\{Y_1, Y_2, \ldots, Y_{t-1}\}$ the observations of Y at time $\{1, 2, \ldots, t-1\}$. At each time t, X_t denotes the covariates possibly related to Y_t. For $t > 0$, let $Z^{t-1} = \left\{\widetilde{G}_0, (X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1})\right\}$ be all the historical information available up to time $t-1$, and set $G_t = (Z^{t-1}, X_t)$. Suppose that, for each time t, the conditional probability of $Y_t = d$ given G_t is modeled by

$$\Pr\{Y_t = d | G_t\} = f_d(G_t), \quad d = 0, 1, \ldots, D, \tag{1}$$

where D means $D+1$ categories for the values of Y_t, and $f_d(\cdot)$'s are unknown probability functions, satisfying $\sum_{d=0}^{D} f_d(G_t) = 1$. Here D can be larger than 1 and so the response Y does not need to be binary. The $D+1$ values are also allowed to be ordered. Note that Y_{t-1} is in G_t, so similar to the AFTER, the algorithm we will propose can be used to forecast time series with autoregressive structure.

In this problem of forecasting, the key concern is to forecast $f_d(G_t)$, $d = 0, 1, \ldots, D$. Suppose we have J candidate forecasting procedures, based on which we aim to construct a combined forecast with adaptive property. For $j = 1, \ldots, J$, we denote $\widehat{f}_{d,j}(G_t)$ as the forecast of $f_d(G_t)$ by the jth candidate procedure. For simplicity, we write $\widehat{f}_{d,t,j} = \widehat{f}_{d,j}(G_t)$. Unlike the combining procedures such as the MMA and JMA mentioned in the Introduction, no restriction is imposed on the J forecasts in this paper. These forecasts are flexible, and can be constructed from different classes of methods and/or under different assumptions. The combined forecast that we propose is in the form

$$\widehat{f}_d(G_t, w_t) = \sum_{j=1}^{J} w_{t,j} \widehat{f}_{d,t,j} \tag{2}$$

with the weight vector $w_t = (w_{t,1}, \ldots, w_{t,J})'$ and its jth element given by

$$w_{t,j} = \begin{cases} \pi_j, & t = 1, \\ \dfrac{\pi_j \prod_{l=1}^{t-1} \left(\prod_{d=0}^{D} \left(\widehat{f}_{d,l,j}\right)^{I(Y_l = d)} \right)}{\sum_{j'=1}^{J} \left(\pi_{j'} \prod_{l=1}^{t-1} \left(\prod_{d=0}^{D} \left(\widehat{f}_{d,l,j'}\right)^{I(Y_l = d)} \right) \right)}, & t > 1, \end{cases} \tag{3}$$

where $\pi_j > 0$ is the prior weight given to the jth forecast, satisfying $\sum_{j=1}^{J} \pi_j = 1$, and $I(\cdot)$ denotes the indicator function as usual. For simplicity, we write $\widehat{f}_d(w_t) = \widehat{f}_d(G_t, w_t)$.

Note that the weight in (3) depends on the prior information, the past forecasts and the corresponding actual realizations. In particular, such weights are dynamic, i.e., they are updated with a new observation. Ignoring the prior information, we see that the bigger the value of $\prod_{l=1}^{t-1} \left(\prod_{d=0}^{D} \left(\widehat{f}_{d,l,j}\right)^{I(Y_l = d)} \right)$, the larger the weight $w_{t,j}$. This value is the probability of making totally correct choice before time t, and thus can be thought of as a measure of the past forecasting accuracy from time 1 to time $t-1$. Therefore, the proposed weights are related to both the past forecasting accuracy and the prior information. Obviously, if the prior weights are equal, then the proposed approach sets a bigger weight to the procedure with higher forecasting accuracy in the past time, which accords to our intuition. Also note that by (3), we have

$$w_{t,j} = \frac{w_{t-1,j} \prod_{d=0}^{D} \left(\widehat{f}_{d,t-1,j}\right)^{I(Y_{t-1} = d)}}{\sum_{j'=1}^{J} \left(w_{t-1,j'} \prod_{d=0}^{D} \left(\widehat{f}_{d,t-1,j'}\right)^{I(Y_{t-1} = d)} \right)}. \tag{4}$$

So similar to the AFTER procedure, the weight of form (3) has a Bayesian interpretation as well: If we view the weight $w_{t-1,j}$ as the prior probability put on the jth forecast before observing Y_{t-1}, then $w_{t,j}$ is the posterior probability of the jth forecast after Y_{t-1} is obtained.

In the case of combining binary predictions, Yuan and Ghosh (2008) and Ghosh and Yuan (2009) developed procedures for adaptive regression by mixing with model screening (ARMS) and improved ARMS by extending the works of Yang (2001a, 2003) and Yuan and Yang (2005). The form of our weighting scheme in (3) is similar to those in these papers, but the latter depends on data splitting and is not suitable for forecasting time series. Our weight form (3) is also similar to the weight implied by the mixing strategy for density estimation in Yang (2000a), where the adaptation of the mixing strategy is shown under Kullback–Leibler (K–L) risk. We will present this result in Remark 4 as a support of using (3) for combining procedures.

[1] The adaptation in the current paper is on minimax-rate adaptation. Referring to Yang (2001b), we briefly describe the minimax-rate adaptive property here. Let $g \in \mathcal{G}_\theta$ be the vector of interest, where θ is the hyper-parameter belonging to Θ, and $\{\widehat{g}_{j,n}, n \geq 1\}$ be a sequence of estimators by the estimation procedure j, $j \in \{1, \ldots, J\}$. The minimax risk in $\mathcal{G}_\theta$ at n is defined as $R(\mathcal{G}_\theta, n) = \inf_{1 \leq j \leq J} \sup_{g \in \mathcal{G}_\theta} E\{l(g, \widehat{g}_j)\}$, where $l(\cdot,\cdot)$ denotes some distance. If the estimation procedure j^* satisfies $\limsup_{n\to\infty} \left[R^{-1}(\mathcal{G}_\theta, n) \sup_{g \in \mathcal{G}_\theta} E\{l(g, \widehat{g}_{j^*})\}\right] < \infty$ for every $\theta \in \Theta$, then we say that the estimation procedure j^* is minimax-rate adaptive over $\{\mathcal{G}_\theta : \theta \in \Theta\}$. Similar definitions can be found in Barron et al. (1999) and Yang (2000b).