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, ...), 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).

3. Adaptation

In this section, we first establish our main result, the adaptation of the combined forecast under the weighting scheme (3) with respect to the squared risk, then present the corresponding result under the K–L risk (Yang, 2000a), and finally consider the adaptation with respect to other risk functions.

3.1. Adaptation under the squared risk

For any forecast $\widehat{f}_{d,t}$ of $f_d(G_t)$, we consider the squared loss

$$L = \sum_{d=0}^{D} \left(\widehat{f}_{d,t} - f_d(G_t)\right)^2.$$

The corresponding risk is given by

$$E_t(L) = E_t\left\{\sum_{d=0}^{D} \left(\widehat{f}_{d,t} - f_d(G_t)\right)^2\right\},$$

where the expectation E_t is taken with respect to the randomness of G_t. We consider the average forecasting risk up to the time n, denoted by R_n, as the performance measure of forecasts, i.e.,

$$R_n = \frac{1}{n}\sum_{t=1}^{n} E_n \sum_{d=0}^{D} \left(\widehat{f}_{d,t} - f_d(G_t)\right)^2.$$

For the jth and the combined forecasts, their average risks are denoted by

$$R_n(j) = \frac{1}{n}\sum_{t=1}^{n} E_n \sum_{d=0}^{D} \left(\widehat{f}_{d,t,j} - f_d(G_t)\right)^2$$

and

$$R_n(w) = \frac{1}{n}\sum_{t=1}^{n} E_n \sum_{d=0}^{D} \left(\widehat{f}_d(w_t) - f_d(G_t)\right)^2,$$

respectively, where $w = (w_1, \ldots, w_n)$.

Before stating the adaptation under the squared risk, we need the following assumption.

Condition (C0): There exists a constant $A > 0$ such that $\widehat{f}_{d,t,j} \geq A$ with probability 1 for all $j \in \{1, \ldots, J\}$, $d \in \{0, \ldots, D\}$ and $t \in \{1, \ldots, n\}$.

This condition is essentially the same as Condition 1 of Yuan and Ghosh (2008), where the lower bounds A_j's are needed to be greater than 0 (see the proof of Theorem 1 of Yuan and Ghosh, 2008). Note that Condition (C0) implies that all the individual forecasts are bounded away from 0, which is mild and easy to satisfy. For example, if $f_d(G_t) \geq \tau > 0$ (which means $f_d(G_t)$ to be bounded away from zero) and $|\widehat{f}_{d,t,j} - f_d(G_t)| \leq \tau_1 < \tau$ with probability 1 for all $j \in \{1, \ldots, J\}$, $d \in \{0, \ldots, D\}$ and $t \in \{1, \ldots, n\}$ (i.e., every forecast is not far away from the true value), where τ and τ_1 are two positive constants, then there exists a positive constant A such that $\widehat{f}_{d,t,j} \geq A$ holds true with probability 1.

The following theorem shows the adaptation of the proposed combination forecast under the squared risk.

Theorem 3.1. *Assume Condition* (C0) *is satisfied. Then the average risk of the combined forecast with the weight* (3) *satisfies*

$$R_n(w) \leq 4A^{-1} \inf_{1\leq j\leq J} \left(-An^{-1}\log\pi_j + R_n(j)\right). \tag{5}$$

Remark 1. Under the i.i.d. observations, a similar bound was obtained by Yang (2000a) for the density estimation. Yang (2004) also developed a similar bound for the combined forecast by the AFTER procedure.

Remark 2. From (5), up to the constant $4A^{-1}$ and the additive penalty $-An^{-1}\log\pi_j$, the combined forecast by the weighting scheme (3) achieves the performance of the optimal forecast among all J individual forecasts under the squared risk. In particular, under the equal priors $\pi_j = 1/J$, (5) reduces to

$$R_n(w) \leq 4A^{-1} \inf_{1\leq j\leq J} R_n(j) + 4n^{-1}\log J. \tag{6}$$

As commented by a referee, when A is small, the upper bound given by (6) may not be very tight. In this case, in view of the assumption (C0), we may need to reduce the number of categories, $D+1$, for the response Y, so that the probability for each category is not too small, and hence neither is the lower bound A in Condition (C0). On the other hand, if $\log J/\inf_{1\leq j\leq J}(nR_n(j)) \to 0$ as $n \to \infty$, then $nR_n(w)/\inf_{1\leq j\leq J}(nR_n(j)) = O(1)$, which means that the sum risk $nR_n(w)$ of the proposed method shares the same increasing rate as the optimal individual forecast. It may also be of interest to compare this property of risk bound with the asymptotically optimal property for the model selection or model averaging (cf., Li, 1987; Shao, 1997; Hansen, 2007). Although the latter is defined based on the risk for the best infeasible individual or combined estimator/forecast rather than the risk multiplied by a constant as in our risk bound, the asymptotic property makes sense only in the case of large sample size, and is established on the basis of linear estimators/forecasts. By contrast, our risk bound property holds true not only for the finite sample size but also without imposing any linearity restriction on the candidate estimators/forecasts.

Remark 3. Let $j^o = \arg\max_{1\leq j\leq J}\{\pi_j \prod_{t=1}^{n}(\prod_{d=0}^{D}(\widehat{f}_{d,t,j})^{I(Y_t=d)})\}$. By simple calculations, we can establish the lower bound of the average risk as follows:

$$R_n(w) \geq 4^{-1}AR_n(j^o) - n^{-1}A\log(J\pi_{j^o}). \tag{7}$$

See the Appendix for the proof of (7). Note that when using equal priors $\pi_1 = \cdots = \pi_J = 1/J$, the j^oth forecast has the biggest probability (among all J forecasts) of making totally correct choice from time 1 to time n. In this case, (7) simplifies to $R_n(w) \geq 4^{-1}AR_n(j^o)$. In practice, the j^oth forecast is infeasible for the prediction from time 1 to time n, because the selection of j^o depends on the forecast at time n, which is in fact unknown when making forecast (unless the forecasting is finished).

Remark 4. We point out that our combined forecast by the weighting scheme (3) also enjoys the adaptation under the K–L risk of Yang (2000a). Denote the average K–L risk of $\widehat{f}_d(w_t)$ by

$$D_n(w) = \frac{1}{n}\sum_{t=1}^{n} E_n \sum_{d=0}^{D} f_d(G_t)\log\frac{f_d(G_t)}{\widehat{f}_d(w_t)},$$

and likewise by $D_n(j)$ for the jth forecast. Proposition 1 of Yang (2000a) provides the following bound on the K–L risk:

$$D_n(w) \leq \inf_{1\leq j\leq J}\left(-n^{-1}\log\pi_j + D_n(j)\right). \tag{8}$$

This is similar to (5) except the constants $4A^{-1}$ and A. It implies that, up to an additive penalty $-n^{-1}\log\pi_j$, the combined forecast by the weighting scheme (3) achieves the performance of the optimal forecast among all J individual forecasts under the K–L risk.

Remark 5. The upper bound given by Theorem 3.1 is with respect to the optimal individual forecast. Sometimes, it is also of interest to target the best combined forecast based on all the J forecasts, for which, similar to Yang (2004), we can first average the J forecasts by some weighting scheme and then combine the averaged forecasts using the proposed weighting procedure.

Let us denote $\eta = (\eta_1, \ldots, \eta_J)'$, the weight vector belonging to $\Omega := \{\eta \in [0,1]^J : \sum_{j=1}^J \eta_j = 1\}$, and the combined forecast by using η is $\widehat{f}_d(G_t, \eta) = \sum_{j=1}^J \eta_j \widehat{f}_{d,t,j}$. Given $\epsilon > 0$, J and n, let N_ϵ be a suitable ϵ-net including $|N_\epsilon| < \infty$ points to discretize Ω under the ℓ_1^J distance; i.e., for each $\eta \in \Omega$, there exists an $\eta^* \in N_\epsilon$ such that $\sum_{j=1}^J |\eta_j - \eta_j^*| \le \epsilon$. Further, we define the combined forecast of all $\widehat{f}_d(G_t, \eta)$'s with $\eta \in N_\epsilon$ by using our proposed procedure with equal priors as $\widehat{f}_d^*(G_t, w_t^*) = \sum_{\eta \in N_\epsilon} w_{t,\eta}^* \widehat{f}_d(G_t, \eta)$, and $\widetilde{R}_n(w^*)$ as the corresponding risk, where $w_{t,\eta}^*$ is the weight for the forecast $\widehat{f}_d(G_t, \eta)$, $w_t^* = \{w_{t,\eta}^* : \eta \in N_\epsilon\}$, and $w^* = (w_1^*, \ldots, w_n^*)$. Then the upper bound depending on the best combined forecast can be derived as

$$\widetilde{R}_n(w^*) \le 8A^{-1} \inf_{\eta \in \Omega} R_n(\eta) + 8A^{-1}(D+1)\epsilon^2 + 4n^{-1} \log |N_\epsilon|, \quad (9)$$

where $R_n(\eta) = \frac{1}{n} \sum_{t=1}^n E_n \sum_{d=0}^D \left(\widehat{f}_d(G_t, \eta) - f_d(G_t)\right)^2$. See the Appendix for the proof of (9). Note that when $J < \sqrt{n}$, it follows from Theorem 1 of Schütt (1984) and the proof of Theorem 6 of Yang (2004) that there exists a discretization of Ω such that

$$8A^{-1}(D+1)\epsilon^2 + 4n^{-1} \log |N_\epsilon| \le A^* n^{-1} J \log(1 + nJ^{-1}), \quad (10)$$

where A^* is a positive constant independent of n. Therefore,

$$\widetilde{R}_n(w^*) \le 8A^{-1} \inf_{\eta \in \Omega} R_n(\eta) + A^* n^{-1} J \log(1 + nJ^{-1}). \quad (11)$$

Comparing (11) with (6), we see that the main difference between them is the term $\log J$ in (6) replaced by the term $J \log(1 + nJ^{-1})$ in (11). When $J \to \infty$, the latter tends to infinity at a faster rate than the former, which is a price paid for the reduction of risk from $\inf_{1 \le j \le J} R_n(j)$ to $\inf_{\eta \in \Omega} R_n(\eta)$. In practice, however, this weighting procedure targeting the best combined forecast is infeasible because the discretization of Ω is unknown.

3.2. Adaptation under other risk functions

In this subsection, we consider the adaptation under a general loss function $\psi(\rho)$ that is a nonnegative continuous function with $\psi(0) = 0$. It is straightforward to show that when $\rho \ge 0$, $\psi(\rho) = \psi(\sqrt{\rho^2})$, and otherwise, $\psi(\rho) = \psi(-\sqrt{\rho^2})$. We define $\phi(\delta) = \psi(\sqrt{\delta})$ and $\phi^*(\delta) = \psi(-\sqrt{\delta})$ for $\delta \ge 0$. To establish the adaptation of $\widehat{f}_d(w_t)$ under the loss function $\psi(\rho)$, we need two more conditions:

Condition (C1): There exists a constant $C > 0$ such that $\max_{0<\delta\le 1} |\phi'(\delta)| \le C$ and $\max_{0<\delta\le 1} |\phi^{*\prime}(\delta)| \le C$, where $\phi'(\cdot)$ and $\phi^{*\prime}(\cdot)$ are the first order derivatives of $\phi(\cdot)$ and $\phi^*(\cdot)$, respectively.

Condition (C2): There exists a constant $C^* > 0$ such that $\psi(\rho) \ge C^* \rho^2$ for $0 < \rho \le 1$.

Note that Conditions (C1) and (C2) correspond to the conditions (10) and (9) of Yang (2004), respectively, which are easily satisfied. Consider an example with widely used asymmetric LINEX loss function: $\psi(\rho) = \exp(a\rho) - a\rho - 1$ with a given constant a. In this case, $\phi(\delta) = \exp(a\delta^{1/2}) - a\delta^{1/2} - 1$ and $\phi^*(\delta) = \exp(-a\delta^{1/2}) + a\delta^{1/2} - 1$. So we have $\phi'(\delta) = a\delta^{-1/2}\{\exp(a\delta^{1/2}) - 1\}/2$, $\phi^{*\prime}(\delta) = -a\delta^{-1/2}\{\exp(-a\delta^{1/2}) - 1\}/2$, and $\lim_{\delta\to 0} \phi'(\delta) = \lim_{\delta\to 0} \phi^{*\prime}(\delta) = a^2/2$, and thus Condition (C1) holds. It can also be shown that there exists some a such that Condition (C2) holds (the proof is simple and available on request from the authors). Denote by $R_n^*(w)$ and $R_n^*(j)$ the risks of the combined and individual forecasts under the loss function $\psi(\rho)$, respectively. Note that $R_n^*(w)$ and $R_n^*(j)$ have the same expressions as $R_n(w)$ and $R_n(j)$ in Section 3.1, respectively, except that the squared loss function is replaced by $\psi(\rho)$. The following theorem shows the adaptation of the combined forecast under the loss function $\psi(\rho)$.

Theorem 3.2. *Assume Conditions* (C0)–(C2) *are satisfied. Then the average risk of the combined forecast by the weighting scheme* (3) *satisfies*

$$R_n^*(w) \le 4CA^{-1}C^{*-1} \inf_{1 \le j \le J} \left(-AC^* n^{-1} \log \pi_j + R_n^*(j)\right). \quad (12)$$

Remark 6. Note that the above adaptation property also depends on the weight (3) and thus supports the use of our weighting scheme. This is different from that in Section 2.6 of Yang (2004), where the adaptation under a convex loss function is based on a new weight that depends on a tuning parameter λ, the choice of which can have a dramatic effect on the weight for moderate sample sizes (Yang, 2004).

The use of general loss functions for forecasting discrete response variable can be found in some work such as Diebold and Rudebusch (1989) and Granger and Machina (2006). Consider a simple example: A risk averter predicts whether the price of a stock will go up (if up, buy the stock, otherwise, do not buy). Then she/he may put a larger loss for over-forecast of the probability of 'going up' than for under-forecast of this probability, and in this case, an asymmetric loss function appears desired.

Let us examine the case of $D = 1$. Note that $\widehat{f}_{0,t,j} + \widehat{f}_{1,t,j} = 1$ and $f_{0,j}(G_t) + f_{1,j}(G_t) = 1$. The sum of loss is therefore equal to $\sum_{d=0}^1 \psi(\widehat{f}_{d,t,j} - f_d(G_t)) = \psi(\widehat{f}_{0,t,j} - f_0(G_t)) + \psi(f_0(G_t) - \widehat{f}_{0,t,j})$, which is symmetric. To solve this problem, as in Diebold and Rudebusch (1989) and Granger and Machina (2006), we may change the loss function $\sum_{d=0}^1 \psi(\widehat{f}_{d,t,j} - f_d(G_t))$ to $\psi(\widehat{f}_{0,t,j} - f_0(G_t))$, and then an asymmetric loss function can be operated. For the squared loss, $\sum_{d=0}^1 (\widehat{f}_{d,t,j} - f_d(G_t))^2 = 2(\widehat{f}_{0,t,j} - f_0(G_t))^2$, so it is straightforward to show that Theorem 3.1 still holds when the loss function $\sum_{d=0}^1 (\widehat{f}_{d,t,j} - f_d(G_t))^2$ is replaced by $(\widehat{f}_{0,t,j} - f_0(G_t))^2$. Likewise, it can be shown that Theorem 3.2 holds as well when the loss function $\sum_{d=0}^1 \psi(\widehat{f}_{d,t,j} - f_d(G_t))$ is replaced by $\psi(\widehat{f}_{0,t,j} - f_0(G_t))$ (the proof is similar to that of Theorem 3.2 and available upon request from the authors).

4. The combined forecast with model screening

When the number of candidate forecasts J is very large, the computation cost of combining all these forecasts will be very substantial, so like the motivation of the ARMS algorithm of Yuan and Ghosh (2008), a model screening step is preferred in this case. The combining procedures with model screening step have also been supported by Yuan and Yang (2005), Claeskens et al. (2006), Zhang et al. (2012) and so on. Specifically, as commented by Hendry and Reade (2008), simple averaging without any screening could provide poor modeling and forecasting, while appropriate screening would lead to efficient averaging and forecasting.

Denote $\widehat{\Gamma}$ as a subset of $\{1, \ldots, J\}$ selected by a model screening method based on the initial set of observations G_0. In the simulation study and practical application of next two sections, we set G_0 to be the same set of observations as used in forecasting $f(G_1)$. In Yuan and Yang (2005) where observations are independent, random data splitting is adopted to determine the observation set for model screening. In the current paper, we are concerned with the time series context, and so we do not use the random data splitting but a set of initial observations for model screening.

Let J_1 be the number of elements in $\widehat{\Gamma}$. According to (3), without any prior information, the weight vector based on the set $\widehat{\Gamma}$ at time t is $\widetilde{w}_t = (\widetilde{w}_{t,1}, \ldots, \widetilde{w}_{t,J_1})'$, with

$$\widetilde{w}_{t,j} = \begin{cases} 1/J_1, & t = 1, \\ \dfrac{\prod_{l=1}^{t-1} \left(\prod_{d=0}^{D} \left(\widehat{f}_{d,l,j}\right)^{I(Y_l = d)} \right)}{\sum_{j' \in \widehat{\Gamma}} \left(\prod_{l=1}^{t-1} \left(\prod_{d=0}^{D} \left(\widehat{f}_{d,l,j'}\right)^{I(Y_l = d)} \right) \right)}, & t > 1. \end{cases} \quad (13)$$

The resulting combined forecast is $\widehat{f}_d(\widetilde{w}_t) = \sum_{j\in\widehat{\Gamma}} \widetilde{w}_{t,j}\widehat{f}_{d,t,j}$, and the associated squared risk is $R_n(\widetilde{w}) = \frac{1}{n}\sum_{t=1}^{n} E_n \sum_{d=0}^{D}(\widehat{f}_d(\widetilde{w}_t) - f_d(G_t))^2$, where $\widetilde{w} = (\widetilde{w}_1, \ldots, \widetilde{w}_n)$. The following theorem establishes the adaptation of $\widehat{f}_d(\widetilde{w}_t)$ under the squared risk.

Theorem 4.1. *Assume Condition* (C0) *is satisfied. Then the average risk of the combined forecast satisfies*

$$R_n(\widetilde{w}) \le 4n^{-1}E_{\widehat{\Gamma}} \log J_1 + \inf_{1\le j\le J}(4A^{-1}R_n(j) + 2\Pr\{j \notin \widehat{\Gamma}\}), \qquad (14)$$

where the expectation $E_{\widehat{\Gamma}}$ is taken with respect to the randomness of $\widehat{\Gamma}$.

Remark 7. Let $j^{\text{opt}} = \arg\min_{1\le j\le J} R_n(j)$. From (14), we have

$$\begin{aligned} R_n(\widetilde{w}) &\le 4n^{-1}E_{\widehat{\Gamma}} \log J_1 + 4A^{-1}R_n(j^{\text{opt}}) + 2\Pr\{j^{\text{opt}} \notin \widehat{\Gamma}\} \\ &= 4n^{-1}E_{\widehat{\Gamma}} \log J_1 + 4A^{-1}\inf_{1\le j\le J} R_n(j) + 2\Pr\{j^{\text{opt}} \notin \widehat{\Gamma}\}. \qquad (15) \end{aligned}$$

By comparison of (15) with (6) (the bound without any model screening step), there is a reduction from $4n^{-1}\log J$ to $4n^{-1}E_{\widehat{\Gamma}}\log J_1$, the benefit due to model screening, and an additional term $2\Pr\{j^{\text{opt}} \notin \widehat{\Gamma}\}$, the price paid for the model screening step. Therefore, whether the set selected by a model screening method contains j^{opt} (or the probability $\Pr(j^{\text{opt}} \notin \widehat{\Gamma})$) will affect the upper bound.

In recent literature, some methods of model screening have been suggested. Yuan and Yang (2005) proposed a 'top m' model screening procedure with the aid of AIC and BIC (i.e., the procedure excludes models that are not 'top m' AIC or BIC models). The resulting number of models may be smaller than $2m$. Claeskens et al. (2006) also suggested a screening method by using the information criteria. Different from the 'top m' method, they used a forward procedure, in each step of which the variable that yields the lowest value for the information criterion when added to the currently "best" model is added. From the perspective of computational burden, this method appears better than the 'top m', because it does not need to calculate the information criterion values for every model. However, this method takes only one model for each size, with the selected models nested, and even the model with the smallest AIC or BIC can be excluded. Therefore, in this paper, we choose the 'top m' model screening procedure. The initial set $\widetilde{G}_0$ is used to calculate the AIC and BIC values in the screening procedure.

5. A simulation study

In this section, we conduct a simulation study to compare our proposed adaptive forecast (termed as AF) with the forecasts selected by the two commonly used criteria, AIC and BIC, and weighted by the two methods, smoothed AIC (S-AIC) and smoothed BIC (S-BIC). Specifically, the S-AIC weights are given by

$$w_j^{\text{AIC}} = \exp\{-\text{AIC}_j/2\} \Big/ \sum_{j=1}^{J} \exp\{-\text{AIC}_j/2\},$$

where AIC_j is the AIC score corresponding to the jth forecast. The S-BIC weights are defined analogously. The performance of the 'top m' screening method discussed in Section 4 is also checked in this simulation study.

We generate our simulated data by the following autoregressive logistic model (de Vries et al., 1998)

$$\begin{cases} \Pr\{Y_t = 0|G_t\} = f_0(G_t) = \dfrac{1}{1 + \sum_{d=1}^{2}\exp\{B_{d,t}\}}, \\ \Pr\{Y_t = d|G_t\} = f_d(G_t) = \dfrac{\exp\{B_{d,t}\}}{1 + \sum_{d=1}^{2}\exp\{B_{d,t}\}}, \quad d = 1, 2, \end{cases}$$

where $t = 2, 3, \ldots, 2n$,

$$\begin{aligned} B_{d,t} &= \beta_{1,d} + \beta_{2,d}I(Y_{t-1} = 1) + \beta_{3,d}I(Y_{t-1} = 2) + \beta_{4,d}X_{t-1,1} \\ &\quad + \cdots + \beta_{7,d}X_{t-1,4}, \end{aligned}$$

$Y_1 = 0$, and the coefficient vectors

$$(\beta_{1,1}, \ldots, \beta_{7,1}) = (2, 0.1, 0.6, \kappa(1, 0.1, 0, 0.5))$$

and

$$(\beta_{1,2}, \ldots, \beta_{7,2}) = (1, 0.2, 0.4, \kappa(0.4, 0.6, 0.8, 0.2))$$

with a parameter κ controlling the distance from the full model (i.e., the model containing intercept, $I(Y_{t-1} = 1)$, $I(Y_{t-1} = 2)$, and $X_{t-1,1}, \ldots, X_{t-1,4}$ as regressors) to the restricted model (i.e., the model containing only intercept, $I(Y_{t-1} = 1)$ and $I(Y_{t-1} = 2)$). Here the four series $\{X_{t,i}\}$ with $i = 1, \ldots, 4$, are independently generated from $AR(1)$ with the autoregressive parameter 0.3. The four variables $X_{t,1}, \ldots, X_{t,4}$ are set to be auxiliary (i.e., they are possibly used in forecasting). Thus we have $2^4 = 16$ candidate models and so 16 forecasts. The parameter κ varies in $\{0, 0.2, 0.4, \ldots, 1.2\}$ so that AIC or BIC performs better than the other in about half cases of the set of κ, and $n = 50$. We set $m = 5$ in the screening step. The equal prior weights are adopted in this simulation and in the practical applications of Section 6 as well.

The forecasting begins with $t = n + 1$ and ends at $t = 2n$. To evaluate all methods, we compute the forecasting risk by

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{t=n+1}^{2n}\sum_{d=0}^{2}\left(\widehat{f}_{d,t}^{(k)} - f_d^{(k)}(G_t)\right)^2,$$

where $\widehat{f}_{d,t}^{(k)}$ denotes the combined forecast or selected individual forecast of $f_d^{(k)}(G_t)$ in the kth replication, and $K = 500$. For ease of comparison, we normalize the risk by dividing by the smallest risk of the individual forecasts. From Section 4, the additional term related to the upper bound of the risk of the AF with screening step is $\Pr(j^{\text{opt}} \notin \widehat{\Gamma})$, which can be approximated by

$$\Delta = \frac{1}{K}\sum_{k=1}^{K} I\left(j^{\text{opt}(k)} \notin \widehat{\Gamma}^{(k)}\right),$$

where $j^{\text{opt}(k)} = \arg\min_{1\le j\le J}\frac{1}{n}\sum_{t=n+1}^{2n}\sum_{d=0}^{2}(\widehat{f}_{d,t,j}^{(k)} - f_d^{(k)}(G_t))^2$, $\widehat{f}_{d,t,j}^{(k)}$ is the forecast of $f_d^{(k)}(G_t)$ by the jth candidate model of the kth replication, and $\widehat{\Gamma}^{(k)}$ is selected by the model screening step in the kth replication. Then, $1 - \Delta$ is calculated as the correctness ratio of the screening step. In addition, we calculate the selection correctness ratios of AIC and BIC by

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{t=n+1}^{2n} I\left(IC_t^{(k)} = j^{\text{opt}(k)}\right),$$

where $IC_t^{(k)}$ denotes the selected model by AIC or BIC at time t in the kth replication.

Risks and correctness ratios are shown by Fig. 1. The following conclusions are easily drawn from the figure.

(i) When κ is small (say <0.6), the BIC performs better than the AIC; however, the reverse phenomenon is observed when κ is large (say >0.8). This is expected because the BIC supports sparse models more than the AIC. It can also be noticed that when one model selection method has a larger correctness ratio, it achieves a lower risk than the other. This finding, in some sense, is similar to that of the "two simple models" simulation of Zou and Yang (2004), where it was found that the performance of model selection is related to the probability of choosing the best model.

(ii) The S-AIC and S-BIC always perform better than the AIC and BIC, respectively. This finding is similar to the result of Hansen

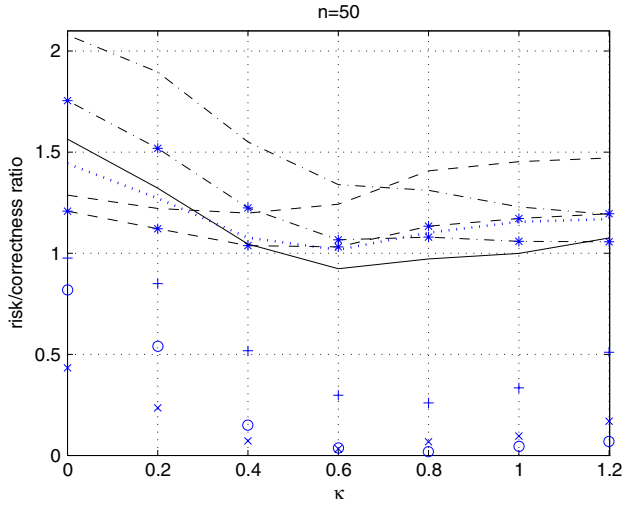

Fig. 1. Risk and correctness ratio of simulation under $n = 50$ (— AF, ⋯ AF with the 'top m' screening method, · — · — AIC, — — BIC, · — ∗ · — SAIC, — ∗ — SBIC, × correctness ratio by AIC, o correctness ratio by BIC, + correctness ratio by the 'top m' screening method).

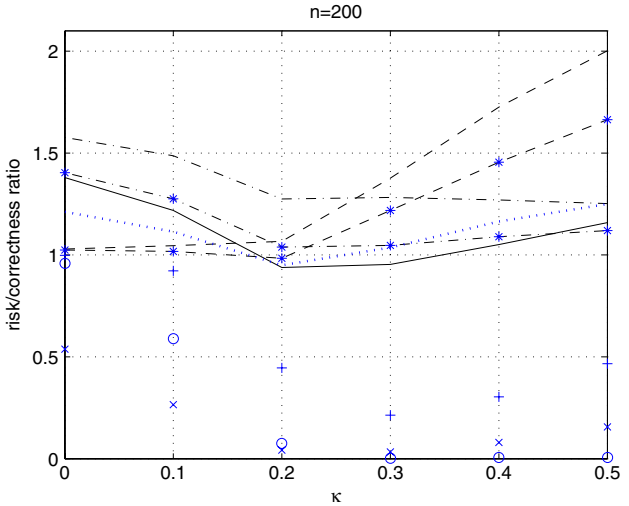

Fig. 2. Risk and correctness ratio of simulation under $n = 200$ (— AF, ⋯ AF with the 'top m' screening method, · — · — AIC, — — BIC, · — ∗ · — SAIC, — ∗ — SBIC, × correctness ratio by AIC, o correctness ratio by BIC, + correctness ratio by the 'top m' screening method).

Table 1
The average risk over all κ values in the simulations with different sample sizes.

n	AIC	BIC	S-AIC	S-BIC	AF	AF with model screening
50	1.51	1.33	1.25	1.13	1.13	1.18
200	1.36	1.37	1.16	1.23	1.12	1.12

(2007) regarding the comparison between the information criteria and their smoothed versions.

(iii) When κ is small (say <0.4), the S-BIC outperforms the S-AIC and the AF performs between them. When κ is large (say close to 1.2), the S-AIC performs better than the S-BIC and the AF still performs between them. However, when the S-AIC and S-BIC perform relatively closely, the AF leads to the lowest risk. More interestingly, when κ is in the range of values around 0.6, the AF achieves a risk lower than 1, implying that it performs better than using the optimal candidate model. Comparing the AF and S-AIC, the former performs better in most of range of the κ value space we used, which also holds when comparing the AF and S-BIC. The average risks over all κ values we considered are listed in the second row of Table 1. Clearly, the AF and S-BIC have close average risks which are smaller than those of other methods.

(iv) By and large, the AF with screening step produces a larger risk than the AF, although it can reduce the computational cost. However, when the correctness ratio of screening step is close to 1, it can reduce the risk (see, for example, the case with $\kappa = 0$). This performance demonstrates again that good screening would lead to good forecasts.

Further, we consider a simulation study with a larger sample size of $n = 200$, and the results are reported in Fig. 2. Broadly speaking, the conclusions drawn from Fig. 2 are similar to those in the previous simulation with $n = 50$. The main difference between them can be found in Table 1. Clearly, both the AIC and S-AIC have lower average risks than the BIC and S-BIC, respectively, while the AF has a lower average risk than the S-BIC, and the average risk of the AF with screening step is quite close to that of the AF itself.

6. Two real-world data examples

In this section, we evaluate the performance of the proposed adaptive forecasting method by two real-world data examples. The first one is concerned with forecasting the direction of Australian All Ordinary Index, which is relatively simple but of interest in investment with a binary response. The second example is to forecast the adjustment of Chinese deposit reserve ratio, where the adjustment can be regarded as a genuinely discrete response of three categories.

6.1. The direction of Australian all ordinary index

In this subsection, we apply the proposed adaptive method to forecasting the direction of the Australian All Ordinary Index, denoted as $Y_1, Y_2, \ldots$, where Y_t is the daily close price index in day t. The whole sample of the data set, depicted in Fig. 3, consists of the observations from 3rd January 2006 to 20th April 2011, covering the period of the recent global financial crisis. We focus on prediction of the direction of the index as an illustration of the advantage of the proposed adaptively combined forecasting for a binary discrete response. Note that forecasting sign of stock movements is often concerned with in investment. As commented by Leung et al. (2000), most trading practices adopted by financial analysts rely on accurate prediction of the price levels of financial instruments, but some recent studies have suggested that trading strategies guided by forecasts on the direction of the change in price level are more effective and may generate higher profits; see also Nyberg (2011) for a recent work on forecasting the direction of stock market.

For simplicity, we tentatively combine some simple and naive forecasts. Let us consider the daily return $r_t = \log(Y_t/Y_{t-1})$. First, we use a simple lag-1 autoregressive logistic model[2]:

$$\Pr\{r_t > 0 | r_{t-1}\} = \exp(\alpha + \beta r_{t-1}) / (1 + \exp(\alpha + \beta r_{t-1})),$$

under which the maximum likelihood estimators of the coefficients (α, β) and their standard errors (given in parentheses) are 0.20 (0.08) and −0.18 (0.11). This means that the higher daily return of one day tends to decrease the possibility of the index increasing in the next day, though the estimator of β is not very significant. We also perform a naive prediction method, that is,

if the index increases (decreases) in one day, then predict it increasing (decreasing) in the next day.

[2] The logit model has been adopted in forecasting the direction of stock return in the literature such as Ou and Penman (1989) and Hirshleifer and Shumway (2003).

Table 2
The hit-rates (in testing sample) in forecasting the direction of the Australian All Ordinary Index using the models built on different moving windows of the estimating sample.

Size of moving window	100	200	300	400	500	600	700	800	900	1000
Naive method	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52
Logistic autoregression	0.50	0.55	0.55	0.55	0.56	0.51	0.50	0.51	0.58	0.56
AF	0.55	0.57	0.56	0.56	0.58	0.58	0.58	0.57	0.59	0.57

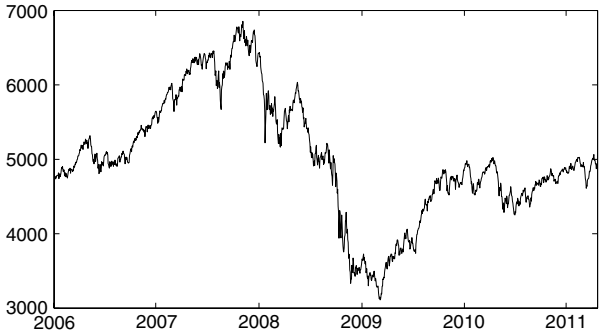

Fig. 3. The daily close price index of Australian all ordinary index.

Then, we examine our proposed adaptive method in combining the forecasts from these two methods. The last 300 observations in the whole sample are used as a testing sample. In addition, in many applications, it is often the case that a prediction may be sensitive to the estimating sample that is used to estimate the forecasting model. Too early observations may not be useful or even lead to worse result in prediction, so often one would use a moving window of sample for estimation in, for example, the autoregressive logistic model considered above. How to select an optimal moving window is often difficult in practice. We consider the sample sizes of the moving window varying from 100 to 1000 in this empirical study.

We report the hit-rates, in the testing sample, of the one-step-ahead direction predictions of the index based on the above methods with the models built on different moving windows of the estimating sample, which are displayed in Table 2. Obviously, the naive method is unrelated to the size of moving window for estimation, so it is stable in forecasting and the correct forecasting ratio keeps to be the understandably low constant rate of 0.52. However, the prediction performance by the autoregressive logistic model is very sensitive to the size of the moving window used for estimation, varying from the worst hit-rate of 0.50 (worse than the naive forecast), corresponding to the size of moving window equal to 100 or 700, to the optimal hit-rate of 0.58, corresponding to the size of moving window equal to 900. Now let us examine our proposed AF. It clearly follows from Table 2 that the AF uniformly obtains the highest hit-rate in each moving window case in comparison with the naive and logistic autoregression methods. The AF method is very stable in terms of the moving window for estimation; in particular when the size of moving window for estimation is greater than 500, the AF produces the correct forecasting ratio from 0.57 to 0.59, about 10% relative improvement compared with the naive method. In some cases, the prediction performance by the autoregressive logistic model is close to that by the AF, but it is sensitive to the change of the size of moving window.

6.2. The adjustment of Chinese deposit reserve ratio

In this subsection, we apply the proposed adaptive method to forecasting the adjustment of Chinese deposit reserve ratio. As a tool of monetary policy, the adjustment of deposit reserve ratio has a significant influence on economic system. The data set we used consists of the monthly observations from January 2006 to June 2012 from CEIC Databases at http://www.ceicdata.com/Chinese.htm. The adjustments Y_t's of these 78 months ($t = 1, \ldots, 78$) vary in a set $\{-1\%, -0.5\%, 0, 0.5\%, 1\%\}$, in which there are only one -1% and four 1%'s, so we reset the adjustment set as {*decrease*, *keep*, *increase*}, with the elements indexed by 0, 1, and 2. The numbers of the three categories of response are 6, 44, and 28, respectively, in the sample. Three variables collected for the forecasting are: V_1 (CPI), V_2 (increase rate of value-added of industry), and V_3 (increase rate of investment in the fixed assets). These kinds of variables on inflation, output, or investment are often used in policy forecasting (see, for example, Monokroussos, 2011). In order to reduce the effect of non-stability and seasonality, the variable V_1 has been operated with difference at lag 1, V_2 with difference at lag 12, and V_3 with both difference at lag 1 and difference at lag 12. Fig. 4 illustrates the time series of all the variables in this example.

Like Section 5, we are using a lag-1 logistic autoregressive regression model[3]:

$$\begin{cases} \Pr\left\{Y_t = d|Y_{t-1}, V_{1,t-1}, V_{2,t-1}, V_{3,t-1}\right\} \\ \quad = \dfrac{\exp(\widetilde{B}_{d,t})}{1 + \sum_{d=0}^{1} \exp(\widetilde{B}_{d,t})}, \quad d = 0, 1 \\ \Pr\left\{Y_t = 2|Y_{t-1}, V_{1,t-1}, V_{2,t-1}, V_{3,t-1}\right\} \\ \quad = \dfrac{1}{1 + \sum_{d=0}^{1} \exp(\widetilde{B}_{d,t})}, \end{cases}$$

where $\widetilde{B}_{d,t} = \gamma_{1,d} + \gamma_{2,d} I(Y_{t-1} = 2) + \gamma_{3,d} V_{1,t-1} + \gamma_{4,d} V_{2,t-1} + \gamma_{5,d} V_{3,t-1}$ and $t = 2, \ldots, 78$. As in Section 5, we set all three regressors, $V_{jt}, j = 1, 2, 3$, as uncertain variables and are unsure whether they should be in the model. Therefore we consider all the possible combinations of the three regressors and this leads to $2^3 = 8$ candidate forecast models. Based on the 8 forecasts, five selection and combination methods, including AIC, BIC, S-AIC, S-BIC, and AF, are examined. The 'top m' screening method is also performed with $m = 1, \ldots, 4$. Observations from July 2009 to June 2012 are used as a testing sample for forecasting evaluation. Also, as in Section 6.1, moving windows are adopted in the application. The sample size of moving window varies from 20 to 40.

The hit-rates of one-step-ahead predictions by these selection and combination methods over different moving window sizes are displayed in Table 3. It is clear that for each size of moving window, the AF method has higher hit-rate than the AIC, BIC, and associated averaging methods. For the AF with 'top m' screening step, we see that it has different performance when m changes. In most of cases, it produces lower hit-rate than the AF itself. In some cases (say the case with $m = 2$ and the size of moving window being 20), the AF with screening step has higher hit-rate than the AF itself. This indicates that good screening may produce improved forecast.

[3] Logistic regression or probit model has been widely used in the analysis or prediction of monetary policy in the literature. See, for example, Hu and Phillips (2004) and Chappell et al. (2007).

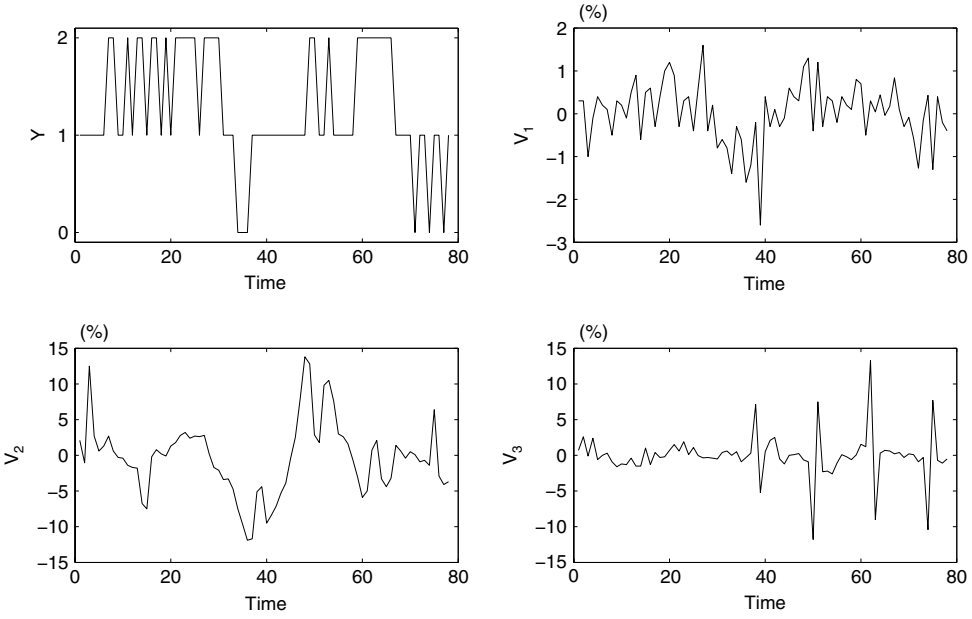

Fig. 4. The adjustment of Chinese deposit reserve ratio Y, CPI V_1, increase rate of value-added of industry V_2, and increase rate of investment in the fixed assets V_3.

Table 3
The hit-rates (in testing sample) in forecasting the adjustment of Chinese deposit reserve ratio using the models built on different moving windows of the estimating sample.

Size of moving window	20	25	30	35	40
AIC	0.53	0.58	0.56	0.56	0.67
BIC	0.56	0.58	0.56	0.58	0.67
S-AIC	0.56	0.58	0.53	0.56	0.67
S-BIC	0.56	0.61	0.56	0.58	0.64
AF	0.64	0.72	0.64	0.61	0.69
AF with "top 1" screening	0.58	0.58	0.56	0.50	0.67
AF with "top 2" screening	0.69	0.58	0.56	0.58	0.69
AF with "top 3" screening	0.69	0.56	0.56	0.58	0.69
AF with "top 4" screening	0.64	0.67	0.61	0.67	0.69

7. Concluding remarks

In this paper, we have considered the issue of adaptive forecast for discrete response time series. A new combining forecast method has been proposed with its adaptation under the squared and other risk functions established. Furthermore, we have extended the adaptation to the forecast combination with a model screening step. Both simulation study and two real-world data examples have displayed the promise of the proposed method.

Before ending the paper, we comment that as in Yuan and Yang (2005), we suggested the 'top m' screening step without any discussion of how to choose a proper m in practice. An appropriate choice of m should achieve a balance between the profit and price from the model screening step. It remains a challenging endeavor to develop an approach for the proper choice of m. In addition, like other publications on adaptive forecasting, the focus of the present paper is on the point forecast, which sometimes could not meet the need of practitioners. Developing an adaptive combination of interval forecasts would warrant our future research.

Acknowledgments

We are very grateful to the editor, Prof. Jianqing Fan, the associate editor, and the two anonymous referees for their very helpful comments and suggestions which substantially improved the original manuscript. Zhang's research partially supported by the two grants 71101141 and 70933003 from the National Natural Science Foundation of China and a special grant from the president of the Chinese Academy of Sciences, Lu's research by the Discovery Project grant and the Future Fellowships grant from the Australian Research Council, and Zou's research by a grant from the Hundred Talents Program of the Chinese Academy of Sciences and the two grants 70625004 and 11021161 from the National Natural Science Foundation of China, are also acknowledged.

Appendix

To simplify the expressions, we denote $u_{d,t,j} = f_d(G_t)/\widehat{f}_{d,t,j}$, $u_d(w_t) = f_d(G_t)/\widehat{f}_d(w_t)$, $v_{d,t,j} = \widehat{f}_{d,t,j} - f_d(G_t)$, and $v_d(w_t) = \widehat{f}_d(w_t) - f_d(G_t)$ for $d \in \{0, \ldots, D\}$, $t \in \{1, \ldots, n\}$, and $j \in \{1, \ldots, J\}$ in this appendix.

Proof of Theorem 3.1. Let $H_n = \prod_{t=1}^{n} \prod_{d=0}^{D} \{f_d(G_t)\}^{I(Y_t=d)}$ and $B_n = \sum_{j=1}^{J} \left\{\pi_j \prod_{t=1}^{n} \prod_{d=0}^{D} \left\{\widehat{f}_{d,t,j}\right\}^{I(Y_t=d)}\right\}$. For any $j^* \in \{1, \ldots, J\}$, we have

$$\log \frac{H_n}{B_n} \le \log \frac{\prod_{t=1}^{n} \prod_{d=0}^{D} \{f_d(G_t)\}^{I(Y_t=d)}}{\pi_{j^*} \prod_{t=1}^{n} \prod_{d=0}^{D} \left\{\widehat{f}_{d,t,j^*}\right\}^{I(Y_t=d)}}$$

$$= \log \frac{1}{\pi_{j^*}} + \sum_{t=1}^{n} \sum_{d=0}^{D} \left\{I(Y_t = d) \log u_{d,t,j^*}\right\}. \tag{A.1}$$

Further, let $h_t = \prod_{d=0}^{D} \{f_d(G_t)\}^{I(Y_t=d)}$ and $g_t = \sum_{j=1}^{J} [w_{t,j} \prod_{d=0}^{D} \{\widehat{f}_{d,t,j}\}^{I(Y_t=d)}]$. Then

$$B_n = \sum_{j=1}^{J} \left[\pi_j \prod_{d=0}^{D} \left\{\widehat{f}_{d,1,j}\right\}^{I(Y_1=d)}\right] \times \frac{\sum_{j=1}^{J} \left[\pi_j \prod_{d=0}^{D} \left\{\widehat{f}_{d,1,j}\right\}^{I(Y_1=d)} \prod_{d=0}^{D} \left\{\widehat{f}_{d,2,j}\right\}^{I(Y_2=d)}\right]}{\sum_{j=1}^{J} \left[\pi_j \prod_{d=0}^{D} \left\{\widehat{f}_{d,1,j}\right\}^{I(Y_1=d)}\right]}$$

$$\times \cdots \times \frac{\sum_{j=1}^{J}\left[\pi_j \prod_{t=1}^{n}\prod_{d=0}^{D}\left\{\widehat{f}_{d,t,j}\right\}^{I(Y_t=d)}\right]}{\sum_{j=1}^{J}\left[\pi_j \prod_{t=1}^{n-1}\prod_{d=0}^{D}\left\{\widehat{f}_{d,t,j}\right\}^{I(Y_t=d)}\right]}$$
$$= \sum_{j=1}^{J}\left[w_{1,j}\prod_{d=0}^{D}\left\{\widehat{f}_{d,1,j}\right\}^{I(Y_1=d)}\right] \times \sum_{j=1}^{J}\left[w_{2,j}\prod_{d=0}^{D}\left\{\widehat{f}_{d,2,j}\right\}^{I(Y_2=d)}\right]$$
$$\times \cdots \times \sum_{j=1}^{J}\left[w_{n,j}\prod_{d=0}^{D}\left\{\widehat{f}_{d,n,j}\right\}^{I(Y_n=d)}\right]$$
$$= g_1 \times g_2 \times \cdots \times g_n. \tag{A.2}$$

So,

$$\log\frac{H_n}{B_n} = \log\frac{h_1 h_2 \times \cdots \times h_n}{g_1 g_2 \times \cdots \times g_n} = \sum_{t=1}^{n}\log\frac{h_t}{g_t}. \tag{A.3}$$

From (A.1) and (A.3), we see that for any $j^* \in \{1, \ldots, J\}$,

$$\sum_{t=1}^{n}\log\frac{h_t}{g_t} \leqslant \log\frac{1}{\pi_{j^*}} + \sum_{t=1}^{n}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}. \tag{A.4}$$

Taking expectation on both sides of (A.4) with respect to the randomness of Y_n given G_n, we obtain

$$\sum_{t=1}^{n-1}\log\frac{h_t}{g_t} + E_{Y_n|G_n}\log\frac{h_n}{g_n}$$
$$\leq \log\frac{1}{\pi_{j^*}} + \sum_{t=1}^{n-1}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}$$
$$+ E_{Y_n|G_n}\sum_{d=0}^{D}\left\{I(Y_n=d)\log u_{d,n,j^*}\right\}$$
$$= \log\frac{1}{\pi_{j^*}} + \sum_{t=1}^{n-1}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}$$
$$+ \sum_{d=0}^{D}\left\{f_d(G_n)\log u_{d,n,j^*}\right\}, \tag{A.5}$$

where

$$E_{Y_n|G_n}\log\frac{h_n}{g_n} = \sum_{d=0}^{D}\left[\Pr\{Y_n=d|G_n\}\log\left\{f_d(G_n)\Big/\sum_{j=1}^{J}(w_{n,j}\widehat{f}_{d,n,j})\right\}\right]$$
$$= \sum_{d=0}^{D}\{f_d(G_n)\log u_d(w_n)\} = 2\sum_{d=0}^{D}\left\{-f_d(G_n)\log u_d^{-1/2}(w_n)\right\}$$
$$\geq 2\sum_{d=0}^{D}\left[f_d(G_n)\left\{1-u_d^{-1/2}(w_n)\right\}\right]$$
$$= 2\sum_{d=0}^{D}\left\{f_d(G_n) - \sqrt{f_d(G_n)\widehat{f}_d(w_n)}\right\}$$
$$= \sum_{d=0}^{D}\left\{f_d(G_n) + \widehat{f}_d(w_n) - 2\sqrt{f_d(G_n)\widehat{f}_d(w_n)}\right\}$$
$$= \frac{1}{4}\sum_{d=0}^{D}\left[4\left\{\sqrt{\widehat{f}_d(w_n)} - \sqrt{f_d(G_n)}\right\}^2 - v_d^2(w_n)\right]$$
$$+ \frac{1}{4}\sum_{d=0}^{D}v_d^2(w_n)$$
$$= \frac{1}{4}\sum_{d=0}^{D}\left(\left[\left\{1+\sqrt{\widehat{f}_d(w_n)}\right\}^2 - \left\{1+\sqrt{f_d(G_n)}\right\}^2\right]\right.$$
$$\times \left[\left\{1-\sqrt{f_d(G_n)}\right\}^2\right.$$
$$\left.\left.- \left\{1-\sqrt{\widehat{f}_d(w_n)}\right\}^2\right]\right) + \frac{1}{4}\sum_{d=0}^{D}v_d^2(w_n)$$
$$\geq \frac{1}{4}\sum_{d=0}^{D}v_d^2(w_n). \tag{A.6}$$

By inserting (A.6) into (A.5), it follows that

$$\frac{1}{4}\sum_{d=0}^{D}v_d^2(w_n) \leq \sum_{d=0}^{D}\left\{f_d(G_n)\log u_{d,n,j^*}\right\}$$
$$+ \sum_{t=1}^{n-1}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}$$
$$- \sum_{t=1}^{n-1}\log\frac{h_t}{g_t} + \log\frac{1}{\pi_{j^*}}. \tag{A.7}$$

Now, by taking expectation on both sides of formula (A.7) with respect to the randomness of G_n, we have

$$\frac{1}{4}E_n\sum_{d=0}^{D}v_d^2(w_n) \leq E_n\sum_{d=0}^{D}\left\{f_d(G_n)\log u_{d,n,j^*}\right\}$$
$$+ \sum_{t=1}^{n-1}E_n\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}$$
$$- \sum_{t=1}^{n-1}E_n\log\frac{h_t}{g_t} + \log\frac{1}{\pi_{j^*}}. \tag{A.8}$$

The first term on the right-hand side of (A.8) satisfies

$$E_n\sum_{d=0}^{D}\left\{f_d(G_n)\log u_{d,n,j^*}\right\} \leq E_n\sum_{d=0}^{D}\left\{f_d(G_n)(u_{d,n,j^*}-1)\right\}$$
$$= E_n\sum_{d=0}^{D}\left\{\frac{f_d^2(G_n) - 2f_d(G_n)\widehat{f}_{d,n,j^*} + \widehat{f}_{d,n,j^*}^2}{\widehat{f}_{d,n,j^*}}\right.$$
$$\left.+ \frac{f_d(G_n)\widehat{f}_{d,n,j^*} - \widehat{f}_{d,n,j^*}^2}{\widehat{f}_{d,n,j^*}}\right\}$$
$$= E_n\sum_{d=0}^{D}\left\{\frac{v_{d,n,j^*}^2}{\widehat{f}_{d,n,j^*}} + \left(f_d(G_n) - \widehat{f}_{d,n,j^*}\right)\right\}$$
$$= E_n\sum_{d=0}^{D}\frac{v_{d,n,j^*}^2}{\widehat{f}_{d,n,j^*}} + 1 - 1$$
$$\leq \frac{1}{A}E_n\sum_{d=0}^{D}v_{d,n,j^*}^2, \tag{A.9}$$

where the last step is from Condition (C0). For $1 \leq t \leq n-1$, the second term on the right-hand side of (A.8) satisfies

$$E_n\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\} = E_{Y_t,G_t}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}$$
$$= E_{G_t}E_{Y_t|G_t}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}$$

$$= E_{G_t}\sum_{d=0}^{D}\left\{f_d(G_t)\log u_{d,t,j^*}\right\} = E_t\sum_{d=0}^{D}\left\{f_d(G_t)\log u_{d,t,j^*}\right\}$$

$$\le \frac{1}{A}E_t\sum_{d=0}^{D}v_{d,t,j^*}^2 = \frac{1}{A}E_n\sum_{d=0}^{D}v_{d,t,j^*}^2, \tag{A.10}$$

where the inequality is from (A.9). As for the third term on the right-hand side of (A.8), using the same argument as (A.6), we see that for any $1 \le t \le n-1$,

$$\begin{aligned}E_n\log\frac{h_t}{g_t} &= E_{G_t}E_{Y_t|G_t}\log\frac{h_t}{g_t}\\ &= E_{G_t}\sum_{d=0}^{D}\left[\Pr\{Y_t=d|G_t\}\right.\\ &\quad\left.\times\log\left\{f_d(G_t)\Big/\sum_{j=1}^{J}\left(w_{t,j}\widehat{f}_{d,t,j}\right)\right\}\right]\\ &= E_t\sum_{d=0}^{D}\{f_d(G_t)\log u_d(w_t)\}\\ &\ge \frac{1}{4}E_n\sum_{d=0}^{D}v_d^2(w_t).\end{aligned} \tag{A.11}$$

Finally, combining (A.8)–(A.11), we have, for any $j^* \in \{1,\ldots,J\}$,

$$\sum_{t=1}^{n}\frac{1}{4}E_n\sum_{d=0}^{D}v_d^2(w_t) \le \log\frac{1}{\pi_{j^*}} + \sum_{t=1}^{n}\frac{1}{A}E_n\sum_{d=0}^{D}v_{d,t,j^*}^2, \tag{A.12}$$

by which we obtain the conclusion (5) and thus finish the proof. □

Proof of formula (7). The proof is very analogous to that of Theorem 3.1. First, by the definition of j^o in Remark 3 and the definitions of H_n and B_n in the proof of Theorem 3.1, we have

$$\begin{aligned}\log\frac{H_n}{B_n} &\ge \log\frac{\prod_{t=1}^{n}\prod_{d=0}^{D}\{f_d(G_t)\}^{I(Y_t=d)}}{J\pi_{j^o}\prod_{t=1}^{n}\prod_{d=0}^{D}\left\{\widehat{f}_{d,t,j^o}\right\}^{I(Y_t=d)}}\\ &= -\log(J\pi_{j^o}) + \sum_{t=1}^{n}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^o}\right\}.\end{aligned} \tag{A.13}$$

From (A.3) and (A.13), we see that

$$\begin{aligned}\sum_{t=1}^{n}\log\frac{h_t}{g_t} &\ge -\log(J\pi_{j^o})\\ &\quad + \sum_{t=1}^{n}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^o}\right\}.\end{aligned} \tag{A.14}$$

So by taking expectation with respect to the randomness of Y_n given G_n, it can be shown that

$$\begin{aligned}&\sum_{t=1}^{n-1}\log\frac{h_t}{g_t} + E_{Y_n|G_n}\log\frac{h_n}{g_n}\\ &\ge -\log(J\pi_{j^o}) + \sum_{t=1}^{n-1}\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^o}\right\}\\ &\quad + \sum_{d=0}^{D}\left\{f_d(G_n)\log u_{d,n,j^o}\right\}.\end{aligned} \tag{A.15}$$

From the second equality of (A.6) and using the proof similar to that of (A.9), we have

$$E_{Y_n|G_n}\log\frac{h_n}{g_n} = \sum_{d=0}^{D}\{f_d(G_n)\log u_d(w_n)\} \le \frac{1}{A}\sum_{d=0}^{D}v_d^2(w_n). \tag{A.16}$$

Using the derivation of (A.6), we obtain

$$E_n\sum_{d=0}^{D}\left\{f_d(G_n)\log u_{d,n,j^o}\right\} \ge \frac{1}{4}E_n\sum_{d=0}^{D}v_{d,n,j^o}^2. \tag{A.17}$$

Further, from (A.6) and (A.10), it follows that

$$\begin{aligned}E_n\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^o}\right\} &= E_t\sum_{d=0}^{D}\left\{f_d(G_t)\log u_{d,t,j^o}\right\}\\ &\ge \frac{1}{4}E_n\sum_{d=0}^{D}v_{d,t,j^o}^2,\end{aligned} \tag{A.18}$$

and from (A.9) and (A.11), we have

$$E_n\log\frac{h_t}{g_t} = E_t\sum_{d=0}^{D}\{f_d(G_t)\log u_d(w_t)\} \le \frac{1}{A}E_n\sum_{d=0}^{D}v_d^2(w_t). \tag{A.19}$$

Now, combining (A.15)–(A.19), it is seen that formula (7) is correct. □

Proof of formula (9). From the definition of N_ϵ, we see that for any $\widehat{f}_d(G_t,\eta)$ with $\eta\in\Omega$, there exists an $\eta^*\in N_\epsilon$ such that

$$|\widehat{f}_d(G_t,\eta) - \widehat{f}_d(G_t,\eta^*)| = \left|\sum_{j=1}^{J}(\eta_j-\eta_j^*)\widehat{f}_{d,t,j}\right| \le \epsilon. \tag{A.20}$$

By the same argument as in the proof of Theorem 3.1, it can be shown that

$$\widetilde{R}_n(w^*) \le 4A^{-1}\inf_{\eta\in N_\epsilon}R_n(\eta) + 4n^{-1}\log|N_\epsilon|. \tag{A.21}$$

Let $\widetilde{\eta} = \operatorname{argmin}_{\eta\in\Omega}R_n(\eta)$. From (A.20) and the triangle inequality, it follows that

$$\begin{aligned}\inf_{\eta\in N_\epsilon}R_n(\eta) &= \inf_{\eta\in N_\epsilon}\frac{1}{n}\sum_{t=1}^{n}E_n\sum_{d=0}^{D}\left(\widehat{f}_d(G_t,\eta)-f_d(G_t)\right)^2\\ &\le 2\inf_{\eta\in N_\epsilon}\frac{1}{n}\sum_{t=1}^{n}E_n\sum_{d=0}^{D}\left(\widehat{f}_d(G_t,\eta)-\widehat{f}_d(G_t,\widetilde{\eta})\right)^2\\ &\quad + 2\frac{1}{n}\sum_{t=1}^{n}E_n\sum_{d=0}^{D}\left(\widehat{f}_d(G_t,\widetilde{\eta})-f_d(G_t)\right)^2\\ &\le 2(D+1)\epsilon^2 + 2\inf_{\eta\in\Omega}R_n(\eta).\end{aligned} \tag{A.22}$$

Combining (A.21) and (A.22), we obtain formula (9). □

Proof of Theorem 3.2. From Condition (C1), we see that for any $1\le t\le n$,

$$\begin{aligned}&\sum_{d=0}^{D}\left\{\widehat{f}_d(w_t)-f_d(G_t)\right\}^2 = \sum_{d=0}^{D}v_d^2(w_t)\left\{I\left(\widehat{f}_d(w_t)\ge f_d(G_t)\right)\right.\\ &\quad\left.+I\left(\widehat{f}_d(w_t)<f_d(G_t)\right)\right\}\\ &\ge \sum_{d=0}^{D}\left\{C^{-1}\max_{0<\delta\le1}|\phi'(\delta)|v_d^2(w_t)I\left(\widehat{f}_d(w_t)\ge f_d(G_t)\right)\right.\\ &\quad\left.+C^{-1}\max_{0<\delta\le1}|\phi^{*\prime}(\delta)|v_d^2(w_t)I\left(\widehat{f}_d(w_t)<f_d(G_t)\right)\right\}\end{aligned}$$

$$\geq C^{-1}\sum_{d=0}^{D}\left\{\left[\phi\left(v_d^2(w_t)\right)-\phi(0)\right]I\left(\widehat{f}_d(w_t)\geq f_d(G_t)\right)\right.$$
$$\left.+\left[\phi^*\left(v_d^2(w_t)\right)-\phi^*(0)\right]I\left(\widehat{f}_d(w_t)<f_d(G_t)\right)\right\}$$
$$=C^{-1}\sum_{d=0}^{D}\left\{\psi\left(\sqrt{v_d^2(w_t)}\right)I\left(\widehat{f}_d(w_t)\geq f_d(G_t)\right)\right.$$
$$\left.+\psi\left(-\sqrt{v_d^2(w_t)}\right)I\left(\widehat{f}_d(w_t)<f_d(G_t)\right)\right\}$$
$$=C^{-1}\sum_{d=0}^{D}\psi\left(v_d(w_t)\right). \tag{A.23}$$

On the other hand, by Condition (C2), it can be seen that for any $1\leq t\leq n$,

$$\sum_{d=0}^{D}v_{d,t,j^*}^2\leq C^{*-1}\sum_{d=0}^{D}\psi\left(v_{d,t,j^*}\right). \tag{A.24}$$

Now, combining (A.6) and (A.23), we have

$$E_{Y_n|G_n}\log\frac{h_n}{g_n}\geq 4^{-1}C^{-1}\sum_{d=0}^{D}\psi\left(v_d(w_n)\right), \tag{A.25}$$

and combining (A.11) and (A.23), we obtain that

$$E_n\log\frac{h_t}{g_t}\geq 4^{-1}C^{-1}E_n\sum_{d=0}^{D}\psi\left(v_d(w_t)\right). \tag{A.26}$$

Similarly, from (A.9), (A.10) and (A.24), it follows that

$$E_n\sum_{d=0}^{D}\left\{f_d(G_n)\log u_{d,n,j^*}\right\}\leq A^{-1}C^{*-1}E_n\sum_{d=0}^{D}\psi\left(v_{d,n,j^*}\right) \tag{A.27}$$

and

$$E_n\sum_{d=0}^{D}\left\{I(Y_t=d)\log u_{d,t,j^*}\right\}\leq A^{-1}C^{*-1}E_n\sum_{d=0}^{D}\psi\left(v_{d,t,j^*}\right). \tag{A.28}$$

Plugging (A.25)–(A.28) into (A.5), formula (12) is obtained. □

Proof of Theorem 4.1. Fix a forecasting procedure $j^*\in\{1,\ldots,J\}$. Let $\widetilde{g}_t=\sum_{j\in\widehat{\Gamma}}\left[\widetilde{w}_{t,j}\prod_{d=0}^{D}\left\{\widehat{f}_{d,t,j}\right\}^{I(Y_t=d)}\right]$, $u_d(\widetilde{w}_t)=f_d(G_t)/\widehat{f}_d(\widetilde{w}_t)$, and $v_d(\widetilde{w}_t)=\widehat{f}_d(\widetilde{w}_t)-f_d(G_t)$. Assume that $j^*\in\widehat{\Gamma}$. First note that formulas (A.5) and (A.6) still hold when g_t, g_n, $1/\pi_j^*$, w_n, $w_{n,j}$, and $\sum_{j=1}^{J}$ are replaced by $\widetilde{g}_t$, $\widetilde{g}_n$, J_1, $\widetilde{w}_n$, $\widetilde{w}_{n,j}$, and $\sum_{j\in\widehat{\Gamma}}$, respectively. From Condition (C0), formula (A.9) holds with probability 1 when E_n's are removed. From Condition (C0) and the fact that the screening step is based on the initial set $\widetilde{G}_0$, it is also straightforward to show that when E_n, E_{Y_t,G_t}, E_{G_t}, E_t, and $w_{t,j}$ are replaced by $E_{G_n|\widehat{\Gamma}}$, $E_{\{Y_t,G_t\}|\widehat{\Gamma}}$, $E_{G_t|\widehat{\Gamma}}$, $E_{G_t|\widehat{\Gamma}}$, and $\widetilde{w}_{t,j}$, respectively, formulas (A.10) and (A.11) remain true. So from the proof of Theorem 3.1, it follows that when $j^*\in\widehat{\Gamma}$,

$$\sum_{t=1}^{n}\frac{1}{4}E_{G_n|\widehat{\Gamma}}\sum_{d=0}^{D}(\widehat{f}_d(\widetilde{w}_t)-f_d(G_n))^2$$
$$\leq E_{G_n|\widehat{\Gamma}}\log J_1+\sum_{t=1}^{n}\frac{1}{A}E_{G_n|\widehat{\Gamma}}\sum_{d=0}^{D}(\widehat{f}_{d,t,j^*}-f_d(G_t))^2$$
$$=\log J_1+\sum_{t=1}^{n}\frac{1}{A}E_{G_n|\widehat{\Gamma}}\sum_{d=0}^{D}(\widehat{f}_{d,t,j^*}-f_d(G_t))^2. \tag{A.29}$$

In addition,

$$\sum_{d=0}^{D}\left\{\widehat{f}_d(\widetilde{w}_t)-f_d(G_t)\right\}^2$$
$$\leq\sum_{d=0}^{D}\left\{\left[\widehat{f}_d(\widetilde{w}_t)\right]^2+[f_d(G_t)]^2\right\}\leq 2. \tag{A.30}$$

Now, from (A.29) and (A.30), we see that for any $j^*\in\{1,\ldots,J\}$,

$$\sum_{t=1}^{n}\frac{1}{4}E_{G_n|\widehat{\Gamma}}\sum_{d=0}^{D}(\widehat{f}_d(\widetilde{w}_t)-f_d(G_n))^2$$
$$=I(j^*\in\widehat{\Gamma})\sum_{t=1}^{n}\frac{1}{4}E_{G_n|\widehat{\Gamma}}\sum_{d=0}^{D}(\widehat{f}_d(\widetilde{w}_t)-f_d(G_n))^2$$
$$+I(j^*\notin\widehat{\Gamma})\sum_{t=1}^{n}\frac{1}{4}E_{G_n|\widehat{\Gamma}}\sum_{d=0}^{D}(\widehat{f}_d(\widetilde{w}_t)-f_d(G_n))^2$$
$$\leq\log J_1+\sum_{t=1}^{n}\frac{1}{A}E_{G_n|\widehat{\Gamma}}\sum_{d=0}^{D}(\widehat{f}_{d,t,j^*}-f_d(G_t))^2$$
$$+n/2I(j^*\notin\widehat{\Gamma}), \tag{A.31}$$

which implies formula (14). □

References

Aguirregabiria, V., Mira, P., 2010. Dynamic discrete choice structural models: a survey. Journal of Econometrics 156, 38–67.

Akhtar, S., Scarf, P., 2012. Forecasting test cricket match outcomes in play. International Journal of Forecasting 28, 632–643.

Altavilla, C., De Grauwe, P., 2010. Forecasting and combining competing models of exchange rate determination. Applied Economics 42, 3455–3480.

Barron, A., Birgé, L., Massart, P., 1999. Risk bounds for model selection by penalization. Probability Theory and Related Fields 113, 301–413.

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Operations Research Quarterly 20, 451–468.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.

Chappell, H.W., McGregor, R.R., Vermilyea, T.A., 2007. The role of the bias in crafting consensus: FOMC decision making in the Greenspan era. International Journal of Central Banking 3, 39–60.

Claeskens, G., Croux, C., Venkerckhoven, J., 2006. Variable selection for logit regression using a prediction-focused information criterion. Biometrics 62, 972–979.

de Vries, S.O., Fidler, V., Kuipers, W.D., Hunink, M.G., 1998. Fitting multistate transition models with autoregressive logistic regression: supervised exercise in intermittent claudication. Medical Decision Making 18, 52–60.

Diebold, F., Rudebusch, G., 1989. Scoring the leading indicators. Journal of Business 62, 369–391.

Dueker, M., 1999. Measuring monetary policy inertia in target fed funds rate changes. Federal Reserve Bank of St. Louis Review 81, 3–9.

Eckstein, Z., Wolpin, K., 1989. The specification and estimation of dynamic stochastic discrete choice models: a survey. Journal of Human Resources 24, 562–598.

Ghosh, D., Yuan, Z., 2009. An improved model averaging scheme for logistic regression. Journal of Multivariate Analysis 100, 1670–1681.

Granger, C.W.J., Machina, M.J., 2006. Forecasting and decision theory. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting, Vol. 1. Elsevier, Amsterdam, pp. 82–98.

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 19, 197–204.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hansen, B.E., 2008. Least squares forecast averaging. Journal of Econometrics 146, 342–350.

Hansen, B.E., Racine, J.S., 2012. Jacknife model averaging. Journal of Econometrics 167, 38–46.

Hendry, D., Reade, J., 2008. Forecasting using model averaging. Working Paper. Nuffield College, University of Oxford.

Hirshleifer, D., Shumway, T., 2003. Good day sunshine: stock returns and the weather. Journal of Finance 3, 1009–1032.

Hu, L., Phillips, P.C.B., 2004. Dynamics of the federal funds target rate: a nonstationary discrete choice approach. Journal of Applied Econometrics 19, 851–867.

Kedem, B., Fokianos, K., 2002. Regression Models for Time Series Analysis. In: Wiley Series in Probability and Statistics.

Leung, G., Barron, A.R., 2006. Infromation theory and mixing least-squares regressions. IEEE Transactions on Information Theory 52, 3396–3410.
Leung, M.T., Daouk, H., Chen, A.S., 2000. Forecasting stock indices: a comparison of classification and level estimation models. International Journal of Forecasting 16, 173–190.
Li, K.-C., 1987. Asymptotic optimality for C_L, cross validation and generalized cross-validations: discrete index set. Annals of Statistics 15, 958–975.
Liang, H., Zou, G., Wan, A.T. K., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. Journal of the American Statistical Association 106, 1053–1066.
Monokroussos, G., 2011. Dynamic limited dependent variable modeling and US monetary policy. Journal of Money, Credit and Banking 43, 519–534.
Müller, G., Czado, C., 2005. An autoregressive ordered probit model with application to high frequency financial data. Journal of Computational and Graphical Statistics 14, 320–338.
Müller, G., Czado, C., 2009. Stochastic volatility models for ordinal valued time series with application to finance. Statistical Modelling 9, 69–95.
Nyberg, H., 2011. Forecasting the direction of the US stock market with dynamic binary probit models. International Journal of Forecasting 27, 561–578.
Ou, J.A., Penman, S.H., 1989. Financial statement analysis and the prediction of stock returns. Journal of Accounting and Economics 11, 295–329.
Rapach, D., Strauss, J., 2008. Forecasting US employment growth using forecast combining methods. Journal of Forecasting 27, 75–93.
Schütt, C., 1984. Entropy numbers of diagonal operators between symmtric Banach spaces. Journal of Approximation Theory 40, 121–128.
Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica 7, 221–264 (with discussion).
Wan, A.T. K., Zhang, X., Zou, G., 2010. Least squares model averaging by Mallows criterion. Journal of Econometrics 156, 277–283.
Yang, Y., 2000a. Mixing straregies for density estimation. Annals of Statistics 28, 75–87.
Yang, Y., 2000b. Adaptive estimation in pattern recognition by combinging different procedures. Statistica Sinica 10, 1069–1089.
Yang, Y., 2001a. Adaptive regression by mixing. Journal of the American Statistical Association 96, 574–586.
Yang, Y., 2001b. Minimax rate adaptive estimation over continuous hyper-earameters. IEEE Transactions on Information Theory 47, 2081–2085.
Yang, Y., 2003. Regression with multiple candidate models: selecting or mixing? Statistica Sinica 13, 783–809.
Yang, Y., 2004. Combining forecasting procedures: some theoretical results. Econometric Theory 20, 176–222.
Yuan, Z., Ghosh, D., 2008. Combining multiple biomarker models in logistic regression. Biometrics 64, 431–439.
Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how? Journal of the American Statistical Association 100, 1202–1214.
Zhang, X., Wan, A.T. K., Zhou, Z., 2012. Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. Journal of Business and Economic Statistics 30, 132–142.
Zou, H., Yang, Y., 2004. Combining time series models for forecasting. International Journal of Forecasting 20, 69–84.