Forthcoming in *Econometric Reviews*

On-line Learning and Forecast Combination in Unbalanced Panels☆

Kajal Lahiri[a], Huaming Peng[a], Yongchen Zhao[a]

[a]*Department of Economics, University at Albany, SUNY*

Abstract

This paper evaluates the performance of a few newly proposed on-line forecast combination algorithms, and compares them with some of the existing ones including the simple average and that of Bates and Granger (1969). We derive asymptotic results for the new algorithms that justify certain established approaches to forecast combination including trimming, clustering, weighting and shrinkage. We also show that when implemented on unbalanced panels, different combination algorithms implicitly impute missing data differently, so that the performance of the resulting combined forecasts are not comparable. After explicitly imputing the missing observations in the U.S. Survey of Professional Forecasters (SPF) over 1968 IV-2013 I, we find that the equally weighted average continues to be hard to beat, but the new algorithms can potentially deliver superior performance at shorter horizons, especially during periods of volatility clustering and structural breaks.

Keywords: On-line learning, Recursive algorithms, Unbalanced panel, SPF forecasts

JEL Classification: C22; C53; C14.

☆An earlier version of the paper was presented at the New York Camp Econometrics VI (Lake Placid, April 2011) and the 17th International Panel Data Conference (Montreal, July 2011). We thank Cheng Hsiao, Tom Wansbeek, three anonymous referees and the handling editor for many helpful comments.

September 2, 2014

On-line Learning and Forecast Combination in Unbalanced Panels☆

1. Introduction

Since the seminal work of Bates and Granger (1969), the potential benefit of combining multiple forecasts instead of simply choosing the single best has long been recognized. The basic idea is that under certain conditions, combined forecasts can be more accurate than individual ones in a panel. Moreover, combining forecasts can be a useful hedge against structural breaks and model instability, see Timmermann (2006) for a survey. However, despite the development of many new forecast combination methods during the past forty years, empirical studies still find that a simple equally-weighted average (henceforth SA) perform admirably well compared to more elaborate procedures. This "forecast combination puzzle", as dubbed originally by Stock and Watson (2004), is related to several issues. One is the "curse of dimensionality", due to the fact that the performance-based combination methods require the estimation of a large number of weight parameters. A number of well-developed Monte Carlo studies and analytical results have illustrated this problem.[1] There are other equally important factors too: A forecaster's past record may not be a good indicator of her future performance due to structural breaks, outliers, new information, uncertainty shocks, and other ex-ante unobservables. These factors can make the relative performance rankings and combination weights unstable, unpredictable, and potentially misleading.[2]

Parallel to the aforementioned literature, another promising approach to forecast combination has developed in recent years using aggregation algorithms and on-line learning. In this approach, time varying combination weights are naturally built into on-line recursive algorithms that do not require the knowledge of the full covariance matrix of the forecast errors. Yang (2004) distinguishes between two broad categories of combination methods.

[1] See Capistrán and Timmermann (2009), Issler and Lima (2009), and Smith and Wallis (2009).

[2] Aiolfi and Timmermann (2006) document such crossings in the context of model based forecasts.

In the first category, called *combining for adaptation*, combination methods are designed to produce combined forecasts that are as good as the best in the panel. In the second category, called *combining for improvement*, methods are designed to produce forecasts that are better than each individual series of forecasts. The Bates and Granger (BG) procedure falls in the second category. Yang (2004) suggests a new automated combination method, the Aggregated Forecast Through Exponential Reweighting (henceforth AFTER), which is a variant of the aggregating algorithm proposed by Vovk (1990), and belongs to the first category. AFTER has been found to be useful in many recent applications.[3]

In addition to AFTER, we consider another on-line recursive machine learning algorithm featuring shrinkage (henceforth MLS) due to Sancetta (2010). Unlike BG, this on-line algorithm tends to select a few forecasters with good performance, thus requiring a much smaller number of estimated parameters. Wei and Yang (2012) note that with alternative forecasts being similar and stable, BG tends to be unnecessarily conservative, and AFTER, in these situations, can perform better. On the other hand, when the best forecaster changes over time and the performance is unstable, gradient-based methods such as MLS can be advantageous. Therefore, the aforementioned methods can be complimentary, depending on the dynamic forecasting environment confronting decision makers in real time.

Wei and Yang (2012) extend the AFTER algorithm designed for squared loss (s-AFTER) to allow for absolute error loss (L_1-AFTER) and Huber loss (h-AFTER), with a special emphasis on reducing the influence of outliers, which are more likely to occur in the presence of structural breaks. In real life situations, however, since the future is unlikely to be predictable, one cannot determine *a priori* which combination strategy to adopt.

This paper has three main objectives. First, we make a modest attempt to derive the asymptotic limits of AFTER-type algorithms, especially the s-AFTER, under various conditions. These algorithms take the number of observations as given and try to evaluate the

[3]See Rapach and Strauss (2005), Leung and Barron (2006), Rapach and Strauss (2007), Fan et al. (2008), Inoue and Kilian (2008), Sanchez (2008), andAltavilla and Grauwe (2010).

performance of competing forecasts via non-asymptotic risk bounds (see Massart (2007)). Hopefully, the large sample limits or approximations we derive in this paper will provide new insights into the working mechanism of AFTER-type algorithms, and explain the rationale for their superior performance under certain conditions.

Second, we evaluate these newly developed on-line algorithms and compare them to a number of existing combination algorithms using the U.S. Survey of Professional Forecasters (SPF) from 1968:IV to 2013:I. Taking advantage of the survey's long time span, we hope to understand the relative strengths and weaknesses of these combination algorithms under the dynamic and complex forecasting scenario of real life.

Finally, we examine the implications of missing data on the comparison of alternative combination algorithms. Schmidt (1977) initiated much of the early work on the estimation of panel data models in unbalanced panels. In the forecasting context, Capistrán and Timmermann (2009) are the first to examine the implication of incomplete panels due to entry, exit and re-entry of experts. We argue that if the missing observations are not treated explicitly and uniformly, different combination algorithms would implicitly yield different imputed values. This means, when applied to unbalanced panels, different combination methods are effectively implemented on different data sets. In our exercises, we address the issue of incomplete panels by explicit imputations, improvising on a procedure suggested by Genre et al. (2013).

The organization of the paper is as follows. Section 2 presents theoretical results on the asymptotic forms of the s-AFTER algorithm. Section 3 discusses the SPF data and data-related issues, including the incomparability between combination methods when applied to unbalanced panels. We also report results from a small-scale simulation study. Section 4 contains a comparison of the alternative combination algorithms. The behavior of the weights implied by the on-line algorithms and BG are examined in Section 5. Section 6 concludes. The mathematical proofs are relegated to the Appendix.

2. Asymptotic Limits of AFTER-type algorithms

As Massart (2007) has pointed out, given the available data, the promise of these on-line algorithms is that the combined forecast would perform almost as well as if the set of best forecasters were known. The interesting question is then how these algorithms realize their potential. One major approach to address this issue is via large sample asymptotics, though more often than not these algorithms are designed to predict short-run evolution of certain phenomena in real time. Thus, in this section, we study the impact of idiosyncratic forecast errors on the s-AFTER combination weights under various asymptotic convergence rates for the idiosyncratic errors.

Let y_t be the real-valued continuous random variable of interest at time t, and X_t be the column vector of random variables representing the information observed by all forecasters prior to the occurrence of y_t. Define $Z_t = (X_t^{'}, y_{t-1})'$ and let $\mathcal{F}_{t-1} = \sigma(\cdots, Z_{t-1}, Z_t)$ be the σ-field generated by current and past Z_t so that $\{y_t, \mathcal{F}_t\}$ is an adapted stochastic sequence. Similarly, we define $\mathcal{F}_{t-1}^j = \sigma(\cdots, Z_{t-1}^j, Z_{jt}^j)$, with $Z_t^j = (X_t^{j'}, y_{t-1})'$ and X_t^j characterizing the information available to the jth forecaster only before the realization of y_t. It follows from the theory of rational expectation that y_t can be decomposed into two components as $y_t = m_t + \varepsilon_t$, where $m_t = E(y_t|\mathcal{F}_{t-1})$ is the conditional mean of y_t given the information contained in $\mathcal{F}_{t-1}$, and ε_t is a random error due to aggregate shocks unforeseeable to forecasters at the time expectations are formed.[4] Clearly, ε_t is a martingale difference sequence relative to the filtration $\{\mathcal{F}_t : t > 0\}$ and m_t is orthogonal to ε_t. Now let $\hat{y}_{jt}$ denote the forecast of y_t made by the jth forecaster conditional on $\mathcal{F}_{t-1}^j$. Noted that $\hat{y}_{jt}$ need not be model-based, with an unknown specification. The forecast combination problem is how to assign weights to these n forecasters at time t after observing $(y_\tau, \hat{y}_{j\tau})$ and the associated forecast error $e_{j\tau}$ $(= y_\tau - \hat{y}_{j\tau})$ for $\tau = 1, \cdots, t-1$ and $j = 1, \cdots, n$.

Since the individual forecast error can be rewritten as $e_{j\tau} = \varepsilon_\tau + \varepsilon_{j\tau}$, where $\varepsilon_{j\tau} = m_\tau - \hat{y}_{j\tau}$, it has a factor interpretation with ε_τ being the common factor and $\varepsilon_{j\tau}$ the idiosyncratic

[4] Note that m_t exists provided $E|y_t| < \infty$, which is satisfied under our assumptions made in this section.

error. The latter could be due to differential interpretation of the public signal by individual forecasters, different loss functions, errors in judgment, and a number of other factors, see Lahiri and Sheng (2010).

In an attempt to combine a group of forecasts based on recent performance, Yang (2004) proposes the s-AFTER algorithm based on minimizing squared error loss in real time. More specifically, when forecast errors are normal and variances are unknown, the weights of s-AFTER are estimated as

$$\hat{\omega}_{jt}^{s-AFTER} = \frac{\prod_{\tau=t_o+1}^{t-1} \hat{\sigma}_{je\tau}^{-1} \exp(-\frac{1}{2} \sum_{\tau=t_o+1}^{t-1} \frac{e_{j\tau}^2}{\hat{\sigma}_{je\tau}^2})}{\sum_{j=1}^{n} \prod_{\tau=t_o+1}^{t-1} \hat{\sigma}_{je\tau}^{-1} \exp(-\frac{1}{2} \sum_{\tau=t_o+1}^{t-1} \frac{e_{j\tau}^2}{\hat{\sigma}_{je\tau}^2})}, \tag{1}$$

where $\hat{\sigma}_{je\tau}^2 = \frac{1}{(\tau-1)} \sum_{s=1}^{\tau-1} e_{js}^2$, and the first t_o observations are used as a training sample to obtain an initial estimate for $\hat{\sigma}_{je(t_o+1)}^2$ at time $t_o + 1$. In this scheme, the lower is the value of $\hat{\sigma}_{je\tau}^2$, the higher the weight. Also, the latest squared forecast error of a forecaster is evaluated relative to its estimated value $\hat{\sigma}_{je\tau}^2$ based on past performance. A large squared forecast error today relative to its expected value is interpreted as a sign of potential deterioration of the forecaster's performance. In addition, the impact of the relative squared forecast error is accumulative, though the contribution of that forecast to the combination is exponentially reduced, see also Zou and Yang (2004).

Wei and Yang (2012) extend the AFTER algorithm by adopting absolute error loss (L_1 loss) and Huber loss, and propose two new algorithms for forecast combination to address the issue of outliers (L_1-AFTER and h-AFTER). Since the large sample approximations of L_1-AFTER and h-AFTER mimic those of s-AFTER, here in this section we focus exclusively on the s-AFTER algorithm.

As additional data are gathered, more precise estimates of $\hat{\sigma}_{ejt}^2$ are readily obtainable. In addition, the availability of an increasingly large information set allows forecasters to comprehend the dynamic properties of the actual process of y_t better. Thus, it may be argued that, under a stable forecasting environment, the idiosyncratic errors made by rational forecasters would gradually disappear. On the other hand, because of diversity in efforts,

expertise, and experience across forecasters, it is reasonable to expect that the speed at which the idiosyncratic errors approach zero may vary among forecasters. Therefore, it is of particular interest to investigate how the varying convergence rates for the idiosyncratic errors across forecasters impact the s-AFTER combination weights in large sample.

To develop the limit theory, besides the fact that ε_τ is a martingale difference sequence relative to the filtration $\{\mathcal{F}_t : t > 0\}$, the following conditions are also needed.

Assumption A (Common errors): $E(\varepsilon_\tau^2|\mathcal{F}_{\tau-1}) = \sigma_{\varepsilon\tau}^2$, $\sigma_{\varepsilon\tau}^2 - \sigma_\varepsilon^2 = O_p(\frac{1}{\tau^\alpha})$ with $\alpha > \frac{1}{2}$, $0 < \sigma_\varepsilon^2 < \infty$, $E|\varepsilon_\tau|^{4+\delta} < M < \infty$ for some $\delta > 0$ and $\frac{1}{t}\sum_{\tau=1}^t E[(\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2)^2|\mathcal{F}_{\tau-1}] \to_p \omega$ for some $\omega > 0$ as $t \to \infty$.

Assumption B (Idiosyncratic errors)

(i): $\frac{1}{t-1}\sum_{\tau=1}^{t-1} E(\frac{\varepsilon_{j\tau}}{\sigma_{\varepsilon\tau}})^2 = O(t^{-\beta_j})$ with $\beta_j > 0$ for all j, $s_j = \lim_{t\to\infty} t^{-1+\beta_j}\sum_{\tau=1}^{t-1} E\Big(\frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2}\Big)$, $d_{jk} = \lim_{t\to\infty} t^{-1+\beta_j}\sum_{\tau=1}^{t}\Big[E\Big(\frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2}\Big) - E\Big(\frac{\varepsilon_{k\tau}^2}{\sigma_{\varepsilon\tau}^2}\Big)\Big]$.

(ii): $\frac{1}{t-1}\Big\{\sum_{\tau=1}^{t-1} var((\frac{\varepsilon_{j\tau}}{\sigma_{\varepsilon\tau}})^2) + \sum_{\tau\neq s} cov((\frac{\varepsilon_{j\tau}}{\sigma_{\varepsilon\tau}})^2, (\frac{\varepsilon_{js}}{\sigma_{\varepsilon s}})^2)\Big\} = O(t^{-\gamma_j})$ with $\gamma_j > 0$ for all j.

Assumption C: ε_τ is independent of $\varepsilon_{j\tau}$ for all j and τ.

Assumption A imposes moment conditions for ε_τ and a weak law of large numbers for the square of martingale difference sequence $\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2$. ε_t is allowed to be conditionally heteroscedastic in finite time, but the degree of heteroscedasticity is regulated by a specified rate of convergence, so that limit theory can be obtained without resorting to conditions such as stationarity and ergodicity. The uniformly boundedness of $(4+\delta)^{th}$ moment condition is slightly stronger than its counterpart given in Proposition 3 of Yang (2004), while the high level assumption of weak law of large numbers in here regarding asymptotic variance is conventional in martingale framework (see, for example, McLeish (1975); Phillips and Solo (1992)).[5] These two conditions are essential for establishing the conditional Lindeberg condition and hence the central limit theorem for the martingale difference sequence $\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2$. Assumption B(i) strengthens the consistency condition in Yang (2004) (viz., the relative idiosyncratic forecast errors for the jth forecaster vanishes over time) by explicitly

[5] Yang (2004) assumes uniform boundedness of the fourth moments in Proposition 3.

specifying the convergence rates so that the asymptotic limits of the AFTER-type weights can be obtained. Each individual forecaster is allowed to have its own potentially distinct rate, which enables us to examine the potential of the AFTER algorithm to fulfill its objective: adapt to the best few forecasters for combination. The restriction $\beta_j > 0$ for all j is for analytical convenience and it can be readily relaxed at the expense of some technical complications. Note that s_j gauges the long run relative squared idiosyncratic errors scaled by t^{β_j} while $d_{jk} = s_j - s_k$ calibrates the difference between s_j and s_k when $\beta_j = \beta_k$, and hence d_{jk} is a relative measure of the long run squared idiosyncratic errors between forecasters j and k. Assumption B(ii), when joined with B(i), controls the speed at which the average relative squared error tends to the long run relative squared error. The independence of errors due to common shocks and those associated with idiosyncratic factors in Assumption C is standard in factor analysis literature. Finally, Assumptions A, B(i), B(ii) and C together imply that forecast errors are asymptotically stationary and asymptotically homogenous across forecasters, though the speed of convergence may differ across forecasters.

The following lemma summarizes the asymptotic properties of the estimate $\hat{\sigma}^2_{ej\tau}$ and is sufficient for our present needs.

Lemma 1. *Under Assumptions A, B(i), B(ii) and C,*
(i) if $\beta_j > \frac{1}{2}$, *then*

$$\frac{1}{(t-1)^{1/2}} \sum_{\tau=1}^{t-1} (e^2_{j\tau} - \sigma^2_{\varepsilon\tau}) \to_d N(0, \omega) \quad as \quad t \to \infty;$$

(ii) if $0 < \beta_j \leq \frac{1}{2}$, *then*

$$\frac{1}{(t-1)^{1/2}} \sum_{\tau=1}^{t-1} (e^2_{j\tau} - \sigma^2_{\varepsilon\tau} - E\varepsilon^2_{j\tau}) \to_d N(0, \omega) \quad as \quad t \to \infty;$$

Lemma 1 establishes a $\sqrt{t}$ convergence rate for the estimate $\hat{\sigma}^2_{jet}$ $(= \frac{1}{t-1} \sum_{\tau=1}^{t-1} e^2_{j\tau})$. In particular, the asymptotic bias term or the so called long run relative squared idiosyncratic errors $\left(\frac{1}{t-1} \sum_{\tau=1}^{t-1} E\varepsilon^2_{j\tau}\right)$ will play no role in the limit distribution as evident in Lemma 1(i)

provided the idiosyncratic errors approach zero at a rate greater than $\sqrt{t}$. By contrast, Lemma 1(ii) shows that $\hat{y}_{jt}$ is asymptotically approaching the conditional mean m_t or the long run efficient forecast with the asymptotic bias $\frac{1}{t-1}\sum_{\tau=1}^{t-1} E\varepsilon_{j\tau}^2$ tending to zero for the jth forecaster provided $0 < \beta_j \leq \frac{1}{2}$. The result follows from the fact that the aggregate shock, which is common to all individual forecasters, eventually dominates the idiosyncratic error. Even though the latter vanishes, it does not vanish fast enough to eliminate the asymptotic bias in the limit distribution for the jth forecaster. The $\sqrt{t}$ rate of convergence for $\hat{\sigma}_{jet}^2$ assists in establishing the asymptotic limits or approximation for $\hat{\omega}_{jt}^{s-AFTER}$.

To derive the large sample limits or approximations for $\hat{\omega}_{jt}^{s-AFTER}$, let $\beta_{(i)}$ be the ith largest value among $\{\beta_j, j = 1, 2, \cdots, n\}$. Also define $B_m = \{\beta_j : \beta_j = \beta_{(m)}\}$ and $B_m^c = \{\beta_{(m+1)}, \cdots, \beta_{(n)}\}$ for $\beta_{(1)} = \cdots = \beta_{(m)} > \beta_{(m+1)} \geq \cdots \geq \beta_{(m)}$.

Proposition 1. *Suppose that Assumptions A, B(i), B(ii) and C hold with* $0 < \beta_j < \frac{1}{2}$ *for all* j*, then as* $t \to \infty$, $t_o \to \infty$ *and* $\frac{t_o}{t} \to 0$,

(i)

$$\hat{\omega}_{jt}^{s-AFTER} \to_p 0,$$

provided $\beta_j < \beta_k$ *for some* $k \neq j$ *or* $\beta_j = \beta_k$ *with* $d_{jk} > 0$ *for some* $k \neq j$*;*

(ii)

$$\hat{\omega}_{jt}^{s-AFTER} \to_p 1,$$

provided $\beta_j > \beta_k$ *for all* $k \neq j$ *or* $\beta_j = \beta_k$ *with* $d_{jk} < 0$ *for all* $k \neq j$*; and*

(iii)

$$\hat{\omega}_{jt}^{s-AFTER} = \begin{cases} \dfrac{\prod_{\tau=t_o+1}^{t-1} \hat{\sigma}_{je\tau}^{-1} \exp(-\frac{1}{2}\sum_{\tau=t_o+1}^{t-1} \frac{e_{j\tau}^2}{\hat{\sigma}_{je\tau}^2})}{\sum_{\beta_k \in B_m} \prod_{\tau=t_o+1}^{t-1} \hat{\sigma}_{ke\tau}^{-1} \exp(-\frac{1}{2}\sum_{\tau=t_o+1}^{t-1} \frac{e_{k\tau}^2}{\hat{\sigma}_{ke\tau}^2})} & \text{if } \beta_j \in B_m \text{ and } d_{jk} = 0 \\ & \text{for all } \beta_k \in B_m \text{ with } k \neq j \\ 0 & \text{if } \beta_j \in B_m^c \end{cases}$$

approximately for large t_o *and* $t \gg t_o$.[6]

[6] t$\gg t_o$ means t is much greater then t_o.

This proposition summarizes the asymptotic results on the conditions which determine the relative importance of the individual forecasters. More specifically, under Assumption A, B(i), B(ii) and C, if the jth forecaster performs worse than at least one other forecaster in terms of slower speed of squared relative errors approaching zero (namely, $\beta_j < \beta_k$ for some k or $d_{jk} > 0$ with $\beta_j = \beta_k$), then this forecaster will be punished with zero weight in combination by the s-AFTER algorithm. By contrast, if the jth forecaster outperforms all other forecasters in terms of long run relative squared errors, s-AFTER algorithm will assign full weight to the jth forecaster, and nothing to others. Finally, when a group of m forecasters perform almost identically in the sense that $d_{jk} = 0$ for some j and all $k \neq j$ in the group, and this cluster of forecasters perform better than the rest, s-AFTER algorithm will pick these m forecasters and assign them nontrivial weights according to their actual relative squared errors as shown in Proposition 1(iii) . Intuitively, these m forecasters are so close to each other that they are hard to distinguish based purely on long run forecasting performance. Under such a scenario, the approximation errors are not neglectable and play an important role in determining the combination weights among the m forecasters in an attempt to guard against instability of relative performance ordering. Proposition 1(iii) considers cases where idiosyncratic errors tend to zero slowly.

It is also of significance to see how s-AFTER responds to a fast convergence rate for the cluster of good forecasters. The result is contained in Proposition 2.

Proposition 2. *Suppose that Assumptions A, B(i), B(ii) and C hold with* $\beta_{(m)} > 1$ *and* $\beta_{(m+1)} \leq 1$, *then,*

$$\hat{\omega}_{jt}^{s-AFTER} \rightarrow_p \begin{cases} \frac{1}{m}, & \beta_j \in B_m, \\ 0, & \beta_j \in B_m^c. \end{cases}$$

as $t \rightarrow \infty$, $t_o \rightarrow \infty$ *and* $\frac{t_o}{t} \rightarrow 0$.

Proposition 2 states that s-AFTER algorithm works like a simple average scheme among the group of m good forecasters while penalizing the rest with null weight under this particular set of convergence rate conditions for idiosyncratic errors. Intuitively, the group

of good forecasters are almost homogenous even in a moderately large finite sample due to their fast rates of convergence. As a result, "the best long run forecaster" is conceptually undefined, and hence the s-AFTER algorithm will assign equal weight ($\frac{1}{m}$) to all the forecasters in this group as time passes by, justifying SA forecast combination rather than selecting the single best, see Chen and Yang (2007).

We show that in the presence of slowly declining variability of errors over time for all forecasters, s-AFTER algorithm identifies a cluster of a few good forecasters whose forecasts approach long run efficiency faster than others, and assigns nontrivial weights to these forecasters. We also characterize situations where a simple average of the chosen few is the optimal solution. In this sense, the AFTER-type algorithm rationalizes many of the approaches advocated in the literature, including trimming, clustering a homogenous group of forecasters, optimal weighting, and shrinkage as suggested in Aiolfi and Timmermann (2006).

In Propositions 1 and 2, the forecast errors for all forecasters are eventually dominated by the common shock - the unforecastable component of the target series.[7] Of course, this is an extreme scenario, and in reality, forecasters often have to respond to a constantly changing environment due to structural breaks and other instabilities. Thus the idiosyncratic errors may not vanish or even diminish. Under these situations idiosyncratic errors may be at least as important as the errors associated with the aggregate shocks in determining the combination weights. As a result, forecasters will be heterogeneous even asymptotically, and the results are summarized in Proposition 3.

Proposition 3. *Suppose Assumption A, C hold, and* $\{\varepsilon_{j\tau}, \mathcal{F}_{\tau}^{j}\}$ *is an adapted stochastic sequence such that* $\varepsilon_{j\tau}$ *is strictly stationary and ergodic with* $E\varepsilon_{j\tau}^2 = \sigma_{j\varepsilon}^2$, $E\varepsilon_{j\tau}^4 < \infty$. *In addition,* $E[E(\varepsilon_{jo}^2 - \sigma_{j\varepsilon}^2|\mathcal{F}_{-t}^{j})]^2 \to 0$ *as* $t \to \infty$, $\sum_{s=0}^{\infty}\left(E(\varepsilon_{jo}^2|\mathcal{F}_{-s}^{j}) - E(\varepsilon_{jo}^2|\mathcal{F}_{-s-1}^{j})\right)^{1/2} < \infty$, *and*

[7] The results in Proposition 1 also apply to the case where only some of the forecasts approach long run efficiency.

var $\left(\frac{1}{t^{1/2}}\sum_{\tau=1}^{t}(\varepsilon_{j\tau}^2-\sigma_{j\varepsilon}^2)\right)>0$ *for all t. Then for large* t_o *and* $t>>t_o$

$$\left(\frac{\hat{\omega}_{kt}^{s-AFTER}}{\hat{\omega}_{jt}^{s-AFTER}}\right)^{\frac{1}{t-t_o}}=\left(\frac{\omega_{kt}^{BG}}{\omega_{jt}^{BG}}\right)^{\frac{1}{2}}$$

approximately provided $\sigma_{\varepsilon j}^2>>\sigma_{\varepsilon\tau}^2$ *for all* j *and all* τ, *where* $\omega_{jt}^{BG}=\frac{\sigma_{jet}^{-2}}{\sum_{j=1}^{n}\sigma_{jet}^{-2}}$, $\sigma_{jet}^2=\sigma_{\varepsilon t}^2+\sigma_{j\varepsilon}^2$.

In proving Proposition 3, we replace Assumption B in Propositions 1 and 2 with some primary conditions that allow the estimated idiosyncratic variances to approach their non-zero population counterparts at a rate of $\sqrt{t}$. Interestingly, it follows from Proposition 3 that $\hat{\omega}_{jt}^{s-AFTER}$ can be approximately written as

$$\hat{\omega}_{jt}^{s-AFTER}=\frac{\left(\omega_{jt}^{BG}\right)^{\frac{t-t_o}{2}}}{\sum_{k=1}^{n}\left(\omega_{kt}^{BG}\right)^{\frac{t-t_o}{2}}}.$$

This means that s-AFTER weights are roughly a normalized power transform of BG weights. Given that the power is positive because $t>t_o$, the s-AFTER algorithm would magnify the weights produced by BG when the forecaster can be distinguished from the rest of the forecasters. Conversely, it tends to discount the weight assigned to a forecaster by BG method if the weight is below some data-dependent threshold level. This explains why s-AFTER is more sensitive to relative forecast errors and, unlike BG, would eventually pick the best forecaster if its past and most recent performance are well above the rest of the forecasters.

We now proceed to evaluate the on-line forecast combination algorithms and compare them with a number of existing procedures using real life forecast data.

3. U.S. SPF and the Missing Data Issues

3.1. SPF Data and Variables

The data we use to evaluate the performance of the combination algorithms is the U.S. Survey of Professional Forecasters (SPF). It is a high quality, long standing, and widely used quarterly survey on macroeconomic forecasts in the United States. The survey was

initially conducted by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER). Starting from 1990, the survey was taken over by the Federal Reserve Bank of Philadelphia. This change in administration led to a unique missing data pattern in the survey, thus a challenge for empirical work.

Based on current quarter GDP price deflator (PGDP), Figure 1 shows the amount and pattern of missing data in SPF over its entire life. A black square in the figure represents a data point; and a blank spot represents a missing forecast. Strikingly, the amount of missing data far exceeds the amount of available data. Taking current-quarter PGDP forecasts as an example, a fully balanced panel with 430 forecasters from 1968:IV to 2013:I without missing data would have 76,540 data points. However, we have only 6,793 data points, meaning that 91% of the data is missing! As for the pattern of missing data, Figure 1 shows that before 1990, there were a large number of forecasters whose forecasts started from the initial years around 1970. Then, about half of these forecasters stopped forecasting midway while the rest kept forecasting until around 1990. Only six forecasters who joined the survey in its early days remain in the sample after 2000. On the other hand, starting from 1990, many new forecasters joined the survey in batches every few years, and about half of them kept forecasting. The amount and pattern of missing data for other variables and horizons are similar.

3.2. Incomparability of Alternative Combined Forecasts in Unbalanced Panels

Like SPF, the majority of panel data sets available for empirical researchers are unbalanced in nature. This subsection builds on Capistrán and Timmermann (2009) and explores the question: what happens if empirical researchers blindly apply various combination methods without properly controlling for the unbalanced structure of the data at hand.

For simplicity, suppose that an analyst observes n forecasters, the earliest time at which forecast data are available is $t = 1$ while the latest time at which forecast data are available is $t = T$. Due to entry and exit of forecasters from time to time, the data is unbalanced, in the sense that $\hat{y}_{jt}$ are missing for some forecasters at some time periods. Let $S_t^A =$

$\left\{j : \hat{y}_{jt} \text{ is observed at time } t \text{ and } j = 1, \cdots, n\right\}$ be the set of forecasters whose forecasts are available at time t, and define $S_t^{NA} = \left\{j : \hat{y}_{jt} \text{ not available at time } t \text{ and } j = 1, \cdots, n\right\}$ as the set of forecasters whose forecasts are not available at time t. Let the number of observed forecasts at time t be n_t, leaving $n - n_t$ elements in S_t^{NA}.

If simple averaging method is applied only to observed data, then effectively, the weights of SA in a unbalanced panel with S_t^A and S_t^{NA} $(t = 1, \cdots, T)$ are governed by

$$\hat{\omega}_{jt}^{SA} = \begin{cases} 0, & \hat{y}_{jt} \text{ is not available} \\ \frac{1}{n_t}, & \text{otherwise} \end{cases}$$

from which we get the combined forecast $\hat{y}_t^{SA} = \sum_{j=1}^n \hat{\omega}_{jt}^{SA} \hat{y}_{jt} = \frac{1}{n_t} \sum_{j=1}^n \hat{y}_{jt}$. Let $\hat{y}_{jt}^*$ be the computed input for the unobserved $\hat{y}_{jt}$, and define $\hat{y}_t^{SA} = \frac{1}{n}\left(\sum_{j \in S_t^A} \hat{y}_{jt} + \sum_{j \in S_t^{NA}} \hat{y}_{jt}^*\right)$. Then since $\sum_{j \in S_t^A} \hat{y}_{jt} = n_t \hat{y}_t^{SA}$, we can obtain $\frac{1}{n - n_t} \sum_{j \in S_t^{NA}} \hat{y}_{jt}^* = \hat{y}_t^{SA}$. This implies that blindly applying simple average method without properly controlling for the unbalanced structure gives rise to imputed values $\hat{y}_{jt}^*\left(j \in S_t^{NA}\right)$ such that the average of these imputed values equals the average of the observed data. For example, if only one data point is not observed, then SA simply imputes it as the average of the observed data from other forecasters without recognizing the idiosyncrasies in the missing forecast.

Next, suppose the Bate and Granger's (BG) method is applied to an unbalanced panel. The weights of BG method are defined by

$$\hat{\omega}_{jt}^{BG} = \begin{cases} 0, & \hat{y}_{jt} \text{ is not available} \\ \hat{\omega}_{jt}^{BGA} = \frac{\hat{\sigma}_{je(t-1)}^{-2}}{\sum_{j \in S_t^A} \hat{\sigma}_{je(t-1)}^{-2}}, & \text{otherwise} \end{cases},$$

from which we have combined forecast $\hat{y}_{jt}^{BG} = \sum_{j=1}^n \hat{\omega}_{jt}^{BG} \hat{y}_{jt} = \sum_{j \in S_t^A} \left(\frac{\hat{\sigma}_{je(t-1)}^{-2}}{\sum_{j \in S_{t-1}^A} \hat{\sigma}_{je(t-1)}^{-2}}\right) \hat{y}_{jt}$. If we assume $\hat{\sigma}_{je(t-1)}^{-2}$ are available for all j at time $t-1$ and $\hat{y}_{jt}^*$ are being filled into the missing spots at time t, then $\hat{y}_{jt}^{BG} = \sum_{j \in S_t^A} \left(\frac{\hat{\sigma}_{je(t-1)}^{-2}}{\sum_{j=1}^n \hat{\sigma}_{je(t-1)}^{-2}}\right) \hat{y}_{jt} + \sum_{j \in S_t^{NA}} \left(\frac{\hat{\sigma}_{je(t-1)}^{-2}}{\sum_{j=1}^n \hat{\sigma}_{je(t-1)}^{-2}}\right) \hat{y}_{jt}^*$. As $\hat{y}_{jt}^{BG}$ $\sum_{j \in S_{t-1}^A} \hat{\sigma}_{je(t-1)}^{-2} = \sum_{j \in S_t^A} \hat{\sigma}_{je(t-1)}^{-2} \hat{y}_{jt}$, we obtain $\sum_{j \in S_{t-1}^{NA}} \left(\frac{\hat{\sigma}_{je(t-1)}^{-2}}{\sum_{j \in S_{t-1}^{NA}} \hat{\sigma}_{je(t-1)}^{-2}}\right) \hat{y}_{jt}^* = \hat{y}_{jt}^{BG}$. This means that inadvertently applying the BG method to unbalanced panels produces imputed values $\hat{y}_{jt}^*\left(j \in S_t^{NA}\right)$ such that the BG weighted average of imputed values equal the BG weighted

average of the observed data at time t, i.e., the combined forecast. Note that for imputed values, the weights are based on either prior knowledge or estimates obtained from past available data. In the simplest case when only one data point is not observed, BG implicitly impute the value as the BG weighted average of the observed data.

Therefore, combined forecasts produced by different combination methods may not be comparable when the methods are applied to an unbalanced panel, since different methods implicitly use different (conceptually balanced) data sets. Consequently, existing results comparing the performance of various combination methods may be misleading if the data used in the comparisons contain missing values. Special attention to this issue is needed in interpreting these empirical results.

Facing the missing data problem, different researchers have handled the situation differently. Issler and Lima (2009) trim a large incomplete panel down to a smaller ($N = 18$, $T = 41$) balanced panel. While a simple and practical approach, this type of trimming could also be interpreted as discarding some values of the original unbalanced panel and replacing them by the averages of the remaining observations. As discussed in the following sections of the paper, this could potentially reduce the heterogeneity in individual forecasters and thus limiting the scope of performance-based combination methods. Capistrán and Timmermann (2009) trim the U.S. Survey of Professional Forecasters (SPF) data, which contains a great amount of missing observations (see our Figure 1), by requiring forecasters to have a minimum of 10 common contiguous observations. They compare the performance of 12 combination procedures relative to SA based on the trimmed unbalanced panel. Poncela et al. (2011) use one-period-ahead forecasts from SPF and restrict the analysis to those individuals who have been in the panel for a minimum of 7 years and never missed more than 4 consecutive forecasts. The resulting unbalanced panel is balanced in two ways: One, the missing one-quarter-ahead forecasts are replaced by the two-quarter-ahead forecasts from the same forecaster. If the two-quarter-ahead forecasts are also missing, then three-quarter-ahead forecasts are used, and so on. The other way is to replace the missing value from a forecaster by that same forecaster's historical mean forecast. Since all the

combination methods are implemented on the same imputation-balanced panel, the results are not affected by the issue raised here. However, imputation schemes like these minimize the commonality of individual forecasts and emphasize the idiosyncratic component in the forecasts.

3.3. A Monte Carlo Study

We have discussed how one could potentially obtain misleading results regarding the relative performance of competing combination algorithms when the comparison is made on the basis of unbalanced panels. To demonstrate this point and to provide a brief comparison of the performance of the combination algorithms, we conduct a set of small-scale simulations using a set up similar to that of Capistrán and Timmermann (2009). Using notations in the previous section, the actual values (y_t) and forecasts ($\hat{y}_{jt}$) are generated based on a two-factor model with factors X_{1t} and X_{2t}:

$$y_t = X_{1t} + X_{2t} + u_{yt} \tag{2}$$

$$\hat{y}_{jt} = \theta_{j1}X_{1t} + \theta_{j2}X_{2t} + u_{jt} \tag{3}$$

where $u_{yt} \sim \mathrm{N}(0, \sigma_y^2)$ and $u_{jt} \sim \mathrm{N}(0, \sigma_j^2)$. Factor dynamics are introduced by letting $X_{rt} = 0.9X_{r(t-1)} + w_{rt}$, $r = 1, 2$, with standard normal innovations. Forecasts are made in real time: one-step out-of-sample, using no information from the future. All errors are independent.

We report the results from two experiments.[8] In the first experiment, we set the performance of forecasters to be heterogeneous cross-sectionally but constant over time: We set $\theta_{j1} = \theta_{j2} = 1$, $\forall j$, but we draw the inverse of the variance σ_j^{-2} from Gamma(2,2) distribution, introducing heterogeneity in performance. In the second experiment, we consider a more realistic scenario where the performance of forecasters are not only heterogeneous cross-sectionally but also varying over time. In addition to the specification of the first experiment, in each period t for each forecaster j, with probability 0.2, σ_j^{-2} is re-drawn from the Gamma(2,2) distribution, introducing variations over time.

[8] Results from additional experiments with varying sample sizes, number of forecasters, and specifications of heterogeneity in forecasters' performance are available from the authors.

In both experiments, apart from using the balanced panel generated according to the above specifications, we mimic the missing data pattern of the U.S. SPF data and construct unbalanced panels on the basis of the balanced panel. For each forecaster, conditioned on having non-missing forecast in the current period, with probability 0.21, we set her next-period forecast to missing; with probability 0.34, we set a forecast to non-missing, conditioned on a missing forecast in the previous period. Both probabilities are estimated using current-quarter GDP price deflator forecasts in SPF, as used in the empirical application in the following sections. Finally, we impute the missing values by setting them as the simple average of available forecasts for the same period, creating balanced panels from the unbalanced panel. Note that due to imputations, these balanced panels are different from the balanced panel based on which the unbalanced panel is created. In real life, only the unbalanced panel and the panels balanced by imputations are available.

In both experiments, simple average (SA), Bates and Granger (BG), and the three AFTER algorithms (s-AFTER, h-AFTER, and L_1-AFTER) are implemented on each of the three data sets: the balanced panel, the unbalanced panel, and the balanced panel with missing values imputed using the average of non-missing values. We consider several sample size and number of forecasters pairs: We first choose large n relative to T where $n = 10, T = 50$, moderate n and t where $n = 5, T = 50$, and large t relative to n where $n = 5, T = 100$. Then, we consider progressively large panels with $n = 20, T = 100$, $n = 20, T = 200$, and $n = 40, T = 300$.

Based on 500 replications, Table 1 reports the percentage deviations of the mean squared forecast errors (MSE) of BG method and the three AFTER algorithms relative to the MSE of the SA method, i.e., a value of 3% means that the MSE of the combined forecasts is 3% higher than the MSE of the SA method, on average over all the replications. Table 1 also reports the average MSE of the SA method as a reference. These simulation results clearly demonstrate the effect of missing values on the performance of the combination algorithms. For example, when forecasters' performance are stable over time, with $n = 40, T = 300$, MSE of s-AFTER method is 1% higher than that of SA when both are implemented on a

balanced panel, 0.5% higher when implemented on the imputation-balanced panel, but is 4% higher when implemented on the unbalanced panel. When only the unbalanced panel is used, one may conclude that the performance of s-AFTER is inferior to that of SA, which is clearly not the case if the imputation-based panel is used in the comparison. The rest of the results illustrate this point as well. We again emphasize the necessity to control for the effect of missing observations when attempting to compare the performance of combination methods, as different methods implicitly impute missing values differently.

Note that in Table 1, the MSE of SA method is always the same regardless of whether the unbalanced panel or the imputation-balanced panel is used, because the imputation method used here is the same as the imputation implied by SA. Another point worth noting is that, even though the AFTER algorithms do not produce combined forecasts that are superior to the SA benchmark, they do display performance that is close to the best forecaster in the panel. This behavior is exactly as expected since data are generated from time-series models like what used here, and that the AFTER algorithms belong to the category of combination for adaptation, see Wei and Yang (2012) and Sancetta (2010) for similar simulation results are discussions.

3.4. Missing Data Imputations for SPF

Based on the pattern of missing data unique to the SPF as discussed before, we choose to construct two subsamples of SPF instead of using the entire data as a whole, for the obvious reason that combined forecasts for the earlier years would be based on a different group of forecasters than that for the recent years even if the data is used in its entirety. The first subsample includes the initial years from 1968:IV to 1990:IV. The second subsample goes from 2000:I to 2013:I. We can thus utilize the part of the sample where data points are highly concentrated, and avoid the part that contains too many missing data points. From the 39 regularly surveyed variables in SPF, we select the growth rate of real GDP (RGDP), seasonally-adjusted annual rate of change for GDP price deflator (PGDP), the CPI inflation rate (CPI), and the seasonally-adjusted quarterly average of monthly unemployment rate (UNEMP) as our target variables. For each variable, we examine the forecasts made for the

current quarter and the following 3 quarters.

We limit our attention to frequent forecasters – ones with sufficient number of observed forecasts – to further reduce the amount of missing data. In both subsamples, the amount of missing data is kept as low as possible while maintaining a reasonable number of forecasters.[9] Specifically, we require forecasters to have at least 45 forecasts in subsample 1 or at least 36 forecasts in subsample 2. As a result, depending on variables and subsamples, around 15 forecasters remain. On average, the amount of missing data is about 40% for subsample 1 and 15% for subsample 2.

To accurately measure the performance of different combination methods in the incomplete panel, we impute the missing data explicitly. Two imputation methods are considered. The first method gives imputed values as $\hat{y}^*_{it} = (1/n_t) \sum_{j \in S^A_t} \hat{y}_{jt}$ if $i \notin S^A_t$, where n_t is the total number of elements in S^A_t. This method replaces missing forecasts with the simple average of non-missing forecasts period-by-period. This is the implied imputation when simple average is used for combination. The downside of this method is that it reduces the level of forecast dispersion thus homogenizing forecasting performance, which limits the combination algorithms' ability to distinguish good performers from poor ones. Especially for the performance-based methods, it would be more reasonable if imputed values reflect, at least partially, the past performance of the forecaster.

Such concerns lead us to the second imputation method, based on Genre et al. (2013), where imputed values are given by $\hat{y}_{jt} - \tilde{\bar{y}}_t = \hat{\beta}_i[\sum_{s=1}^{4}(\hat{y}_{jt-s} - \tilde{\bar{y}}_{t-s})]$, with $\tilde{\bar{y}}_t = (1/n_t) \sum_{j \in S^A_t} \hat{y}_{jt}$ is the mean of available forecasts at time t. Intuitively, a missing forecast is replaced by the adjusted mean forecast for that period. The adjustment, which is forecaster-specific, is made according to the recent average deviations of the forecasts made by that forecaster from the mean forecasts. This "hot deck" imputation method is superior, in principle, to the first method, because the imputed values are forecaster-specific and incorporate both the common component of all forecasters and the idiosyncrasy of that forecaster, see Andridge

[9] We observe no clear relationship between performance and participation. Capistrán and Timmermann (2009) and Genre et al. (2013) report similar findings.

and Little (2010). In particular, if a forecaster tends to produce forecasts that are far from the average, her imputed forecasts would reflect that characteristic.

In practice, both imputation methods have to be implemented in real time just like the combination methods. This requirement presents no problem for the first imputation method. But for the second method, the excessive amount of missing data even after imposing the participation requirement makes it infeasible sometimes to estimate all the β_is in real time. When this is the case, we use the most recent estimate of β_i when such an estimate is available.

Missing data creates an additional challenge to the estimation of the weights associated with some algorithms. However, it may be noted that the on-line algorithms may have a relative advantage over BG in this regard. First, as pointed out before, they do not need estimates of error covariances. In addition, the weights of s-AFTER defined in Section 2 can be written recursively as

$$\hat{\omega}_{jt+1}^{s-AFTER} = \frac{\hat{\omega}_{jt}^{s-AFTER}\hat{\sigma}_{jet}^{-1}\exp\left(-\frac{e_{jt}^2}{2\hat{\sigma}_{jet}^2}\right)}{\sum_{j=1}^{n}\left[\hat{\omega}_{jt}^{s-AFTER}\hat{\sigma}_{jet}^{-1}\exp\left(-\frac{e_{jt}^2}{2\hat{\sigma}_{jet}^2}\right)\right]} \text{ for } t \geq t_o + 1 \tag{4}$$

from which it is evident that forecast errors from the past, which affect $\hat{\omega}_{jt+1}^{s-AFTER}$ through $\hat{\omega}_{jt}^{s-AFTER}$, play an equally important role in determining $\hat{\omega}_{jt+1}^{s-AFTER}$ as the latest forecast error. Indeed, the natural logarithm of $\hat{\omega}_{jt+1}^{s-AFTER}$ behaves like a unit root process. Thus a large forecast error tends to have a permanent effect on the weights of s-AFTER. For our purposes, it may be an advantage since the long memory property of $\hat{\omega}_{jt+1}^{s-AFTER}$ may help to alleviate the problem of missing data in unbalanced panel as the impact of past errors do not decay at all.

4. Performance of Combination Methods

4.1. Measuring the Performance of Combination Methods

To evaluate the performance of the AFTER algorithms and the machine learning algorithm MLS, we conduct a real time forecast combination exercise using SPF forecasts

of four important macroeconomic variables. There are subtle differences in the assumptions underlying AFTER and MLS algorithms: AFTER requires the existence of a moment generating function of the forecast errors, but the errors need not be bounded. MLS does not make any assumption on the nature of the forecasts and the actuals, or the stability of the system besides a tail condition on the error distribution. However, its performance bounds are rather weak. In addition to the AFTERs and MLS, we consider several other combination methods, including Bates and Granger (BG), median (ME), recent best (RB), trimmed mean (TM), as well as simple average (SA). SA is such that the combined forecast is the simple (equally-weighted) average of individual forecasts. ME uses the median of individual forecasts as the combined forecast. RB takes as combined forecast for the current period the forecast made by the individual forecaster who has the smallest MSE over all previous periods. TM selects the mean of the pool of individual forecasts after the maximum and the minimum forecasts are removed.[10].

The MLS method is implemented according to Algorithm 1 in Sancetta (2010). The core step in the algorithm is to compute the current-period weight (before shrinkage) $\omega_{jt+1}^{MLS'}$ for each forecaster, based on her previous-period weight ω_{jt}^{MLS} and current-period loss $l_t(\omega_t^{MLS})$. Let $\nabla l_t(\omega_t^{MLS})$ be the gradient of the loss function with respect to (previous-period) weight ω_t^{MLS}, and $\nabla_j l_t(\omega_t^{MLS})$ be its jth element. The current-period weight is calculated as:

$$\omega_{jt+1}^{MLS'} = \frac{\omega_{jt}^{MLS} \exp\left[-\eta t^{-\alpha} \nabla_j l_t(\omega_{jt}^{MLS})\right]}{\sum_{j=1}^{n} \left\{\omega_{jt}^{MLS} \exp\left[-\eta t^{-\alpha} \nabla_j l_t(\omega_{jt}^{MLS})\right]\right\}} \tag{5}$$

where η is the learning rate, and α is a parameter that controls the speed of learning. In the shrinkage step that gives the current-period weight used for combination ω_{jt+1}^{MLS}, all the $\omega_{jt+1}^{MLS'}$s that are lower than a predetermined small threshold (γ/η, which is controlled by parameter γ, given a choice of η) is replaced by the threshold weight γ/n, and the remaining

[10]We have also considered the Winsorized mean method where the top and bottom 5% of individual forecasts are Winsorized. Although this maintains the variability of individual forecasts more than trimming, its performance is very similar to that of TM, and is therefore omitted.

weights are scaled such that all weights add up to 1. This step eliminates extremely big or small weights and shrinks all weights towards SA (or equal) weights. In the MLS method, the gradient of the loss function $\nabla l_t(\omega_t^{MLS})$, together with learning rate parameter η and α, are used in generating the ex post combination weights, which are then projected on a pre-specified subset in the shrinkage step to ensure that all the weights are bounded by some threshold constraint.

We compare the performance of each combination method against that of SA using the relative MSE measure. For any combination method, the relative MSE is the ratio between the MSE of the combined forecasts produced by that method and the MSE of the averages of the individual forecasts. The relative MSE provides information on the relative forecast accuracy that is independent of the absolute accuracy (i.e., the actual MSEs). The latter often vary greatly depending on the variable, horizon, and sample periods. Therefore, it is entirely possible that in certain cases, even methods with relatively better performance produce poor forecasts. Such forecasts are of no practical use, and comparisons of these forecasts are spurious. Therefore, while still reporting the results for longer horizon forecasts, we focus our analysis on current-quarter and one-quarter-ahead forecasts.

As we carry out the forecast combination exercises in real time, following now a standard practice in forecast evaluation, we use the first vintage (initial release) of a variable as actual value when calculating the MSEs. Pair-wise comparison of combination methods can be performed using the Diebold and Mariano (1995) test of equal forecast accuracy. However, statistical significance of such tests across target variables and forecast horizons are not be directly compared, cf. Capistrán and Timmermann (2009). To guard against possible data snooping, White (2000)'s reality check tests can also be conducted. However, as Clark and McCraken (2013) have pointed out, for recursive MSPE comparisons in nested models, as in our case, the Diebold-Mariano test is advisable even though it tends to be unduly conservative. Thus, the true significance of the relative forecast performance could be higher than suggested by the test results. More importantly, our aim here is not to identify the universally superior method via pair-wise comparisons, but rather to iden-

tify situations in which one method seems to have an edge over others. As indicated in our theoretical discussion, the on-line algorithms like s-AFTER has scope for enhanced performance in real time because they use not only estimated long-run variance but also individual forecasts errors from recent past to select forecasters. Thus, the equal-RMSE tests are less meaningful in our context. Therefore, while we report the test results, we caution against over-interpreting them.

4.2. Comparison of the Accuracy of Combined Forecasts

We implement the aforementioned combination algorithms in real time on imputation-balanced panels and compare their performance. MLS is implemented with $\alpha = 0.5$ and $\gamma = 0.05$, as in Sancetta (2010). The learning rate parameter η is chosen *ex post* from values in the set $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 1\}$ based on performance of the combined forecasts.[11] Table 2 gives the MSEs of the combined forecasts for CPI Inflation, PGDP, RGDP and UNEMP produced by different combination methods.[12] Figure 2 shows the relative MSEs for PGDP and UNEMP based on the MSE values reported in Table 2.

For PGDP, performance-based combination methods, especially the AFTER algorithms and the MLS method, often perform better in subsample 1 (1968:IV to 1990:IV), presumably because the volatility and heterogeneity in individual forecasters were more substantial in the early 1980s. For current-quarter forecasts, both MLS and AFTER produce combined forecasts with MSEs lower than 80% of that of the simple average forecasts. Improvements in forecast accuracies are noted for one-quarter and two-quarter horizons as well

[11] For choosing the optimum learning rate, Sancetta (2010) proposes two methods in addition to choosing the learning rate *ex post*. Based on our results, in most cases, the performance of MLS algorithm is found to be insensitive to how the learning rate is chosen, similar to the findings in Sancetta (2010) (p. 613). Results obtained using automatically-determined, data-dependent learning rate are available upon request.

[12] In what follows, we report only the results from using the second imputation method. Also, we omit results related to the median, trimmed mean, and Winsorized mean combination methods because they did not contribute any additional insight to our findings in the paper.

with decreasing efficacy. MSE of the current-quarter forecasts produced by MLS, the best performer for this subsample, is 1.250, far less than the benchmark MSE of 1.687.

For UNEMP, performance-based algorithms contribute the most to forecast accuracy in subsample 2 (2000:I to 2013:I), when unemployment rate was drastically and unexpectedly affected by the most recent recession beginning 2007:IV. The combined forecasts produced by the L_1-AFTER algorithm have about 10% lower MSE than those based on the simple average for most horizons. The MLS algorithm also provides slightly more accurate forecasts. The h-AFTER and the s-AFTER algorithms noticeably outperform the benchmark at one-quarter and two-quarter horizons.

From Table 2, we find that the MSEs for CPI inflation associated with MLS, s-AFTER, L_1-AFTER and h-AFTER for current quarter forecasts during the second subsample period are less than those for SA by substantial margins. For RGDP, none of the combination methods shows clear-cut superiority over SA. We also note that for all variables and in both subsamples, BG and MLS never produce combined forecasts that are much less accurate than simple average, while the AFTER algorithms sometimes show inferior performance. Overall, as the horizon increases, relative contribution of the combination methods relative to the simple average becomes smaller.

Since L_1-AFTER and h-AFTER are derived using loss functions other than the squared loss, it is necessary to evaluate their performance using appropriate loss functions. Table 3 provides such a comparison in two panels corresponding to Huber loss and absolute loss for PGDP and UNEMP. As expected, the losses, when calculated using the appropriate loss functions, are smaller than those reported in Table 2.[13] More appropriately, in Figure 3, we show the relative MAE and Huber loss for L_1-AFTER and h-AFTER respectively. The figure shows that the relative performance of L_1-AFTER and h-AFTER under appropriate loss functions is similar to the results in Figure 2, and does not contradict our conclusions based on squared error losses.

[13]Only exceptions are a few L_1-AFTER losses for UNEMP for current and one-quarter-ahead forecasts where the errors are very small fractions, so that their squares become much less that their absolute values.

4.3. Imputation Methods and Missing Data Re-Examined

The second imputation method (i.e., using imputation regressions) preserves an individual's tendency to make forecasts that deviate from the cross-sectional averages. This method is advantageous when the forecasters in the sample do indeed have such a tendency, i.e., idiosyncratic biases. By regressing deviations from the contemporaneous averages on lagged deviations, we can check to what extent such a tendency exists. Results of these checks are reported in Table 4. For each variable and each subsample, we report the number of regressions we ran for each forecaster, as well as the average proportion, across all forecasters, of these regressions in which the past deviation is statistically significant.

For PGDP subsample 1, in only 9% of the regressions, past deviations are significant in explaining future deviations, when current-quarter forecasts are examined. But for three-quarter-ahead forecasts, in 35% of the regressions, past deviations are significant. For PGDP subsample 2, past deviations are significant in a considerably larger proportion of regressions at all horizons, increasing from 16% to as high as 55%. The situation is the opposite for UNEMP, where in subsample 2, past deviations are significant in a smaller number of regressions compared to subsample 1. Still, as horizon increases, for UNEMP, past deviations become significant in increasing percentage of cases.

However, note that the difference between the missing forecasts imputed by the second imputation method and those imputed by simple average is generally rather small. Even for some of the forecasters whose forecasts consistently deviate from the mean, since the deviations and/or $\hat{\beta}_i$ may be small, the imputed values are often close to the mean. A similar observation is reported in Genre et al. (2013).

5. Behavior of Selected Forecast Combination Algorithms

5.1. A Closer Look at s-AFTER, MLS, and BG Methods

The results in the previous section suggest possible advantage of the newly developed AFTER and MLS methods in certain cases. Even though many studies have documented the performance of a wide range of combination algorithms using actual and/or simulated

data, few studies looked at how the combination methods work in practice in small samples period-by-period on an individual level. It is therefore particularly interesting and informative to compare their behavior to that of the familiar BG method. By observing and documenting the small sample behavior of these combination algorithms, we hope to gain additional insights into their inner working mechanisms that we may not easily obtain by relying exclusively on asymptotics. In each of Figures 4, 5 and 6, from top to bottom, for each forecaster in the panel, we show the squared forecast errors, the evolution of cumulative MSEs, the weights estimated using BG method, the weights estimated using s-AFTER (or MLS) method, as well as the squared errors of the combined forecasts produced by the two methods.

Figure 4 compares the MLS method with the BG method using current-quarter forecasts of PGDP for subsample 1. In this case, the MLS method performs better than the BG method with 25% lower MSE. As individual squared errors and MSEs show, individual performance is rather stable and clearly heterogeneous, with the exception of a few quarters at the beginning of the sample period. The BG method produces stable weights after the first year or two, essentially weighting most of the forecasters equally around 8%. This is the so-called portfolio diversification logic of BG emphasized by Timmermann (2006). On the contrary, the MLS method puts extremely high weights on the best forecaster who shows persistently good performance. In the beginning of 1978, the previously identified best forecaster showed a small uptick in MSE which led to a drastic down weighting of the forecaster with the second best picking up the share of the weight drastically. Subsequently, the weights assigned by the MLS method dropped by nearly 50%, while the weights assigned to this forecaster by the BG method dropped very little. As shown in the comparison of the squared errors of the combined forecasts, the squared error of BG combined forecast is significantly larger than that of the MLS combined forecast. A similar event happened again in early 1981, where the MLS combined forecasts showed a smaller error. In addition, we note that starting from 1978, after the deterioration in the performance of the previous top forecaster, the MLS method gave almost equal weights to the next two best

forecasters, until after 1981 when the previous top forecaster re-established her edge. During this period, no significant change happened to the weights assigned to these forecasters by the BG method.

In Figure 5, we compare the s-AFTER method with the BG method using one-quarter-ahead forecasts of UNEMP for subsample 2. In this case, apart from the relatively large errors seen after 2007, for most of the quarters, individual forecast errors are small and individual performance is clearly heterogeneous, especially for the best and worst forecasters. This persistence in ranking is similar to the evidence in Aiolfi and Timmermann (2006). Both the BG and the s-AFTER methods successfully identify the best forecaster and assign relatively high weight. Still, the weights assigned by the BG method are around 10% while the weights assigned by the s-AFTER method varied from 20% to 95%. This behavior of s-AFTER is clearly anticipated by our asymptotic theory presented in Proposition 3 where we express s-AFTER weights as normalised power transform of BG weights. In two quarters around early 2008 and early 2009, the best forecaster made relatively big mistakes that led to notable increases in her MSE. Similar to the behavior of the MLS method in the previous case, the s-AFTER method drastically decreased the weights as a result of the performance hit – about a 40% decrease in weight in early 2008 and about a 20% decrease in early 2009. Yang (2004) has emphasized this property of the s-AFTER algorithm wherein a small error by a well-established good forecaster produces significant weight adjustment. Aggressively weighting the good forecaster and penalizing poor performance, s-AFTER displays superior performance with more than 20% reduction in MSE compared to the benchmark. The superior performance of the s-AFTER method for this period, compared to the BG method, comes mostly because of smaller errors during the period from late 2008 to early 2009.

From the above two cases, we see that s-AFTER and MLS methods behave more aggressively in adjusting the weights than the familiar BG method. This makes the algorithms adapt to changes in individual forecasters' performance and adjust their weights in an agile manner, so that changes in performance get quickly reflected in weights. However, if the

changes in performance do not persist into the future, the adjustments made by these algorithms may even worsen the situation. For example, during periods with high volatility, a poor forecaster may produce a highly accurate forecast purely by chance rather than due to forecasting skill. When a change in performance is less likely to persist, the weights should arguably be adjusted cautiously rather than aggressively. A psychological support for this logic can be found in Denrell and Fang (2010). From a diversification perspective, aggressively adjusting weights creates increased risk. If structural break happens or a top forecaster happens to behave poorly in one period, the combined forecasts may suffer a huge unexpected loss. Figure 6 provides such an example. In forecasting current quarter RGDP, the best forecaster, forecaster 44, who was receiving nearly 90% of the weight from s-AFTER, made a big mistake in 1979:IV. Even though immediately after this mistake, the weight assigned to this forecaster by s-AFTER method dropped to below 1%, there was no chance to avoid a big forecast error in that period. This mistake alone made the s-AFTER inferior on average over the whole sample compared to BG even though for all other quarters in the sample, the two sets of forecasts are very close. Interestingly, after the 1979:IV mistake, forecaster 44 vanished completely from combining even though his/her performance continued to be in the middle range. Unlike model-based forecasts that tend to be more stable and rank preserving (cf, Aiolfi and Timmermann (2006)), the relatively large psychological component in these subjective survey forecasts makes s-AFTER-type algorithms particularly susceptible to such outliers.

In order to see how outliers are accommodated in L_1-AFTER, we re-examine Figure 5 where the dynamics of BG and s-AFTER are presented in forecasting UNEMP in subsample 2. In Figure 7, we show the individual MAEs, L_1-AFTER weights and squared error from L_1-AFTER forecasts during the same episode. Due to the huge forecast errors during 2008-2009 period, whereas s-AFTER hesitated with its prime forecaster by scaling down the weight temporarily (see Figure 5), L_1-AFTER downplayed the importance of the initial error and steadily increased the weight of the original best forecaster. The result was that in the post-2008 period, L_1-AFTER had notably smaller forecast errors than both BG and

s-AFTER. This example clearly illustrates how, compared to s-AFTER, L_1-AFTER dealt with and minimized the influence of forecast outliers generated by the latest recession.

5.2. Requirements for Successful Combinations

Based on the results in the previous sections, we are able to identify the following conditions, under which the performance-based weighting algorithms are likely to outperform the simple average method.

Firstly, it is crucial that performance of individual forecasters are relatively stable over time and rank preserving. A stricter requirement is that past performance of a forecaster is a good predictor of this forecaster's future performance. This condition is necessary because the performance-based weighting methods rely on past performance to predict future performance. In our experiments, we found the predictive power to be very low, and the widespread incidences of missing forecasts in subjective survey forecasts like the SPF make the value of the combination weights even more tenuous.

Secondly, in order to produce a series of optimally-combined forecasts that outperform simple benchmarks, the differences in forecasters' performance should be sufficiently large together with widely different correlations in forecast errors between forecasters. Much of these requirements have been discussed in Aiolfi and Timmermann (2006), and have been independently corroborated in the burgeoning psychological literature.[14] However, these prerequisites assume special significance while analyzing the performance of aggressive on-line combination algorithms facing many missing forecasts. Since the on-line algorithms depend more seriously on the correct identification and stability of the few good forecasters, frequent gaps in the data create more serious problems for these algorithms than the BG method. Too many missing data create a similar problem with the hot deck imputation scheme suggested by Genre et al (2013).

Thirdly, for weighting methods that generously weight the best forecaster based on past

[14]See for example, Larrick and Soll (2006), Yaniv and Milyavsky (2007), Vul and Pashler (2008), Herzog and Hertwig (2009), and Soll and Larrick (2009).

performance, e.g., the AFTER or MLS methods, it is necessary that the best forecaster does not make big mistakes. Otherwise, such sparse combining based on a very few forecasters would produce combined forecasts that can suffer greatly from such mistakes, and will be more risky than the BG approach.

6. Conclusions

This study examines the performance and dynamic behavior of the newly developed AFTER and MLS methods for forecast combination in unbalanced panels. For monitoring large surveys like SPF or Blue Chip forecasts, these on-line algorithms can be automated such that learning and adaptation to time-varying relative usefulness of forecasters, old and new, can take place without user intervention and discretion. Our aim here is not to run a horse race among alternative combination schemes, but rather to understand the conditions under which alternative forecast combination algorithms work better than the simple equally-weighted average.

To have a better understanding of how these alternative algorithms work, we establish the large sample limits or approximations of the s-AFTER algorithm. We show that with differentially declining heteroscadasticity of forecast errors over time across survey respondents, s-AFTER algorithm allocates nonzero weights to a group of good forecasters whose forecasts approach long run efficiency faster than others. This justifies many of the suggested forecast combination approaches like trimming, clustering of a homogenous group of forecasters, optimal weighting, and shrinkage. Under some other reasonable assumptions, the s-AFTER weights are shown to be approximately a normalized power transform of BG weights.

We then show that when implementing different forecast combination methods on unbalanced panels, each method implicitly imputes the missing forecasts differently. This makes the performance of the combined forecasts produced by different combination algorithms not directly comparable. To address this issue, we explicitly impute missing forecasts using a regression method that incorporates individual idiosyncrasies as well as

the average forecast of others, and use the same data to evaluate alternative combination methods.

Furthermore, we evaluate these newly developed forecast combination procedures and examine in details the inner mechanics characterizing the algorithms. Empirical evidence confirms our analytical results on the behavior of the combination algorithms. Our results suggest that these robust on-line algorithms can potentially reduce the MSE of the combined forecasts, when persistent forecaster heterogeneity and outliers are prevalent in the forecasts. This is achieved mostly because the algorithms are very agile in adapting to recent changes in individual performance and weighting good forecasters aggressively. We find that the on-line algorithms tend to perform well at shorter horizons, especially when other algorithms fail due to volatility clustering, structural breaks, and outliers. In particular, the performance of individual forecasters need to be sufficiently persistent and heterogeneous for the newly developed pattern recognition and machine learning algorithms to deliver maximum improvement in forecast accuracy. Unfortunately, situations in which these conditions prevail are difficult to determine a priori. Moreover, since the on-line algorithms depend critically on the correct identification of a few good forecasters in real time, frequent gaps in the forecast data, as typical in most expectational surveys, create special problems for these methods. Thus, on balance, our evidence suggests that the simple average scheme continues to be a safe and dependable combination method in summarizing survey forecasts provided by large panels with frequent entry and exit of experts.

References

Aiolfi M, Timmermann A. 2006. Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics* **135**: 31–53.

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

Andridge RC, Little RA. 2010. A review of hot deck imputation for survey non-response. *International Statistical Review* **78**: 40–64.

Bates JM, Granger CWJ. 1969. The Combination of Forecasts. *Operations Research Quarterly* **20**: 451–468.

Capistrán C, Timmermann A. 2009. Forecast Combination With Entry and Exit of Experts. *Journal of Business & Economic Statistics* **27**: 428–440.

Chen Z, Yang Y. 2007. Time series models for forecasting: Testing or combining. *Studies in Nonlinear Dynamics and Econometrics* **11**: Article 3.

Clark T, McCraken M. 2013. Advances in Forecast Evaluation. In Elliott G, Granger C, Timmermann A (eds.) *Handbook of Economic Forecasting*, volume 2. Elsevier, 1107–1201.

Denrell J, Fang C. 2010. Predicting the Next Big Thing: Success as a Signal of Poor Judgment. *Management Science* **56**: 1653–1667.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**: 253–263.

Fan S, Chen L, Lee WJ. 2008. Short-term load forecasting using comprehensive combination based on multi-meteorological information. *Industrial and Commercial Power Systems Technical Conference. ICPS. IEEE/IAS.* .

Genre V, Kenny G, Meyler A, Timmermann A. 2013. Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting* **29**: 108–121.

Hall P, Heyde CC. 1980. *Martingale limit theory and its application*. Academic Press.

Herzog SM, Hertwig R. 2009. The Wisdom of Many in One Mind Improving Individual Judgments With Dialectical Bootstrapping. *Psychological Science* **20**: 231–237.

Inoue A, Kilian L. 2008. How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association* **103**: 511–522.

Issler JV, Lima LR. 2009. A panel data approach to economic forecasting: The bias-corrected average forecast. *Nonparametric and Robust Methods in Econometrics* **152**: 153–164.

Lahiri K, Sheng X. 2010. Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics* **25**: 514–538.

Larrick RP, Soll JB. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science* **52**: 111–127.

Leung G, Barron A. 2006. Information theory and mixing least-squares regressions. *Information Theory, IEEE Transactions* **52**: 3396–3410.

Massart P. 2007. *Concentration inequalities and model selection*. Springer, Berlin.

McLeish DL. 1975. Dependent central limit theorems and invariance principles. *The Annals of Probability* **2**: 620–628.

Phillips CB, Solo V. 1992. Asymptotics for linear processes. *Annals of Statistics* **20**: 971–1001.

Poncela P, Rodrigues J, Sanchez-Mangas R, Senra E. 2011. Forecast combination through dimension reduction techniques. *International Journal of Forecasting* **27**: 224–237.

Rapach DE, Strauss JK. 2005. Forecasting employment growth in Missouri with many potentially relevant predictors: An analysis of forecast combination methods. *Federal Reserve Bank of St. Louis Regional Economic Development* **1**: 97–112.

Rapach DE, Strauss JK. 2007. Forecasting real housing price growth in the eighth district states. *Federal Reserve Bank of St. Louis Regional Economic Development* **3**: 33–42.

Sancetta A. 2010. Recursive forecast combination for dependent heterogeneous data. *Econometric Theory* **26**: 598–631.

Sanchez I. 2008. Adaptive combination of forecasts with application to wind energy. *International Journal of Forecasting* **24**: 679–693.

Schmidt P. 1977. Estimation of seemingly unrelated regressions with unequal numbers of observations. *Journal of Econometrics* **5**: 365–377.

Smith J, Wallis KF. 2009. A Simple Explanation of the Forecast Combination Puzzle. *Oxford Bulletin of Economics and Statistics* **71**: 331–355.

Soll JB, Larrick RP. 2009. Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **35**: 780–805.

Stock J, Watson M. 2004. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* **23**: 405–430.

Timmermann A. 2006. Forecast Combinations. In Elliott G, Granger C, Timmermann A (eds.) *Handbook of Economic Forecasting*, volume 1. Elsevier, 135–196.

Vovk VG. 1990. Aggregating strategies. *Proceedings of the Third Workshop on Computational Learning Theory* : 371–386.

Vul E, Pashler H. 2008. Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science* **19**: 645–647.

Wei X, Yang Y. 2012. Robust forecast combination. *Journal of Econometrics* **166**: 224–236.

White H. 1984. *Asymptotic Theory for Econometricians*. New York: Academic Press.

White H. 2000. A Reality Check for Data Snooping. *Econometrica* **68**: 1097–1126.

Yang Y. 2004. Combining forecasting procedures: some theoretical results. *Econometric Theory* **20**: 176–222.

Yaniv I, Milyavsky M. 2007. Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* **103**: 104–120.

Zou H, Yang Y. 2004. Combining time series models for forecasting. *International Journal of Forecasting* **20**: 69–84.

Appendix

Proof of Lemma 1(i). Write

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}(e_{j\tau}^2-\sigma_{\varepsilon t}^2)=\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}(e_{j\tau}^2-\sigma_{\varepsilon\tau}^2)$$
$$+\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\left(\sigma_{\varepsilon\tau}^2-\sigma_{\varepsilon}^2\right)+\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\left(\sigma_{\varepsilon}^2-\sigma_{\varepsilon t}^2\right). \quad (1)$$

Using Assumption A, we obtain,

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\left(\sigma_{\varepsilon}^2-\sigma_{\varepsilon t}^2\right)=O_p(t^{\frac{1}{2}-\alpha}),$$

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\left(\sigma_{\varepsilon\tau}^2-\sigma_{\varepsilon}^2\right)=\begin{cases}O_p((t-1)^{(\frac{1}{2}-\alpha)}) & \text{if } \frac{1}{2}<\alpha<1\\ O_p(\frac{\log(t-1)}{(t-1)^{1/2}}) & \text{if } \alpha=1\\ O_p(\frac{1}{(t-1)^{1/2}}) & \text{if } \alpha>1\end{cases}$$

Since $\alpha > \frac{1}{2}$, the last two terms in the RHS of (1) converges in probability to zero as $t \to \infty$. Therefore, it suffices to establish the asymptotic limit of $\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}(e_{j\tau}^2-\sigma_{\varepsilon\tau}^2)$. Since $e_{j\tau}=\varepsilon_\tau+\varepsilon_{j\tau}$, where $\varepsilon_{j\tau}=(m_\tau-\hat{y}_{j\tau})$, we write

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}(e_{j\tau}^2-\sigma_{\varepsilon\tau}^2)=\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}(\varepsilon_\tau^2-\sigma_{\varepsilon\tau}^2)$$
$$+\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\varepsilon_{j\tau}^2+\frac{2}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\varepsilon_{j\tau}\varepsilon_\tau.$$

Under the assumption $E(\varepsilon_\tau^2|\mathcal{F}_{\tau-1}) = \sigma_{\varepsilon\tau}^2$, we have $E(\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2|\mathcal{F}_{\tau-1}) = 0$ for $\tau > 1$. Hence $\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2$ is a martingale difference sequence relative to the filtration $\{\mathcal{F}_\tau, \tau \geq 0\}$. Note that, for some $\delta > 0$ and any $0 < \epsilon < \infty$,

$$\begin{aligned}
& t^{-1}\sum_{\tau=1}^{t} E\left\{(\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2)^2 1_{\{|\varepsilon_\tau^2-\sigma_{\varepsilon\tau}^2|\geq\epsilon t^{1/2}\}}\right\} \\
\leq\ & t^{-1}\sum_{\tau=1}^{t} E\left\{(\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2)^2 1_{\{|\varepsilon_\tau^2-\sigma_{\varepsilon\tau}^2|\geq\epsilon t^{1/2}\}}(\frac{|\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2|}{\epsilon t^{1/2}})^{\frac{\delta}{2}}\right\} \\
\leq\ & \epsilon^{-\frac{\delta}{2}} t^{-\frac{\delta}{4}} \underset{\sup_\tau}{E}\left\{|\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2|^{2+\frac{\delta}{2}}\right\} \\
\leq\ & 2^{1+\frac{\delta}{2}}\epsilon^{-\frac{\delta}{2}} t^{-\frac{\delta}{4}} M \\
\to\ & {}_p\, 0 \quad \text{as } t \to \infty.
\end{aligned}$$

where the third inequality is obtained by Loeve's inequality. Next, using the fact $0 \leq E(\varepsilon_{\tau+1}^4|\mathcal{F}_\tau) - \sigma_{\varepsilon\tau+1}^4 \leq E(\varepsilon_{\tau+1}^4|\mathcal{F}_\tau)$ almost surely, Jensen inequality and the law of iterated expectations, we get

$$\underset{\sup_t}{E}\left\{\frac{1}{t}\sum_{\tau=1}^{t} E\left[(\varepsilon_{\tau+1}^2 - \sigma_{\varepsilon\tau+1}^2)^2|\mathcal{F}_\tau\right]\right\}^{1+\frac{\delta}{4}} \leq M < \infty.$$

It follows immediately that

$$t^{-1}\sum_{\tau=1}^{t} E\left\{(\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2)^2 1_{\{|\varepsilon_\tau^2-\sigma_{\varepsilon\tau}^2|\geq\epsilon t^{1/2}\}}|\mathcal{F}_{\tau-1}\right\} \to_p 0, \tag{2}$$

as $t \to \infty$. (2), together with Assumption A, yields

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}(\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2) \to_d N(0,\omega) \quad \text{as} \quad t \to \infty, \tag{3}$$

by virtue of central limit theorem for martingale difference sequence (see, for example, Corollary 3.1 in Hall and Heyde (1980)). To prove Lemma 1(i), it remains to show

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\varepsilon_{j\tau}^2 = o_p(1), \tag{4}$$

and

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\varepsilon_{j\tau}\varepsilon_\tau = o_p(1). \tag{5}$$

Observe that

$$\begin{aligned}\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}E\varepsilon_{j\tau}^2 &= \frac{1}{E(\sigma_\varepsilon^{-2})(t-1)^{1/2}}\sum_{\tau=1}^{t-1}E\Big(\frac{\varepsilon_{j\tau}}{\sigma_{\varepsilon\tau}}\Big)^2 \\ &\quad + \frac{1}{E(\sigma_\varepsilon^{-2})(t-1)^{1/2}}\sum_{\tau=1}^{t-1}E(\sigma_\varepsilon^{-2}-\sigma_{\varepsilon\tau}^{-2})E\varepsilon_{j\tau}^2 \\ &= O(t^{\frac{1}{2}-\beta_j}), \end{aligned} \tag{6}$$

where the first equality is obtained by virtue of Assumption C, and the second equality follows readily from Taylor's expansion, Assumption A and B(i). Using Assumption B(ii), we have

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}[\varepsilon_{j\tau}^2 - E(\varepsilon_{j\tau}^2)] = O_p(t^{-\frac{\gamma_j}{2}}), \tag{7}$$

as $t \to \infty$. (6) and (7) jointly yield

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\varepsilon_{j\tau}^2 = O_p\Big(t^{\max\{\frac{1}{2}-\beta_j,-\frac{\gamma_j}{2}\}}\Big), \tag{8}$$

verifying (4) provided $\beta_j > \frac{1}{2}$ and $\gamma_j > 0$. For (5), using Assumption C and the fact that ε_τ is a martingale difference sequence, we see immediately that

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}E(\varepsilon_{j\tau}\varepsilon_\tau) = 0, \tag{9}$$

and

$$\begin{aligned} var\Big(\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\varepsilon_{j\tau}\varepsilon_\tau\Big) &= \frac{1}{(t-1)}\sum_{\tau=1}^{t-1}E\varepsilon_{j\tau}^2 E\varepsilon_\tau^2 \\ &\le M^{\frac{2}{4+\delta}}\frac{1}{(t-1)}\sum_{\tau=1}^{t-1}E\varepsilon_{j\tau}^2 \\ &= O(t^{-\beta_j}). \end{aligned}$$

Combining (9) and (10), we have

$$\frac{1}{(t-1)^{1/2}}\sum_{\tau=1}^{t-1}\varepsilon_{j\tau}\varepsilon_\tau = O_p(t^{-\frac{\beta_j}{2}}), \tag{10}$$

which confirms the validity of (5) as long as $\beta_j > 0$. We now proceed to prove Lemma 1(ii). Observe that $\frac{1}{(t-1)^{1/2}} \sum_{\tau=1}^{t-1} \varepsilon_{j\tau}^2$ will not vanish even as $t \to \infty$ in view of (8) when $0 < \beta_j \leq \frac{1}{2}$. It is thus necessary to remove the asymptotic bias to derive the asymptotic normality. Note that by using (7) and (10), we have

$$\begin{aligned} \frac{1}{(t-1)^{1/2}} \sum_{\tau=1}^{t-1} (e_{j\tau}^2 - \sigma_{\varepsilon\tau}^2 - E\varepsilon_{j\tau}^2) &= \frac{1}{(t-1)^{1/2}} \sum_{\tau=1}^{t-1} (\varepsilon_\tau^2 - \sigma_{\varepsilon\tau}^2) + o_p(1) \\ &\to_d N(0, \omega) \end{aligned}$$

whenever $0 < \beta_j \leq \frac{1}{2}$ and $\gamma_j > 0$. □

Proof of Proposition 1. We first prove (i) and (ii). Let $\sigma_{j\varepsilon\tau}^2 = \frac{1}{(\tau-1)} \sum_{s=1}^{\tau-1} E\varepsilon_{js}^2$, and $Z_{j\tau} = \frac{1}{(\tau-1)^{1/2}} (\sum_{s=1}^{\tau-1} (e_{j\tau}^2 - \sigma_{\varepsilon\tau}^2 + \sigma_{j\varepsilon\tau}^2)$. Then

$$\hat{\sigma}_{ej\tau}^2 = \sigma_{\varepsilon\tau}^2 + \sigma_{j\varepsilon\tau}^2 + (\tau - 1)^{-1/2} Z_{j\tau}$$

since $\sigma_{j\varepsilon\tau}^2 = O(\tau^{-\beta_j})$ by Assumption B(i) and $Z_{j\tau}$ is of order $O_p(1)$ by lemma 1(ii), with $0 < \beta_j < \frac{1}{2}$, we can write, by neglecting the term of smaller stochastic order,

$$\hat{\sigma}_{je\tau}^2 = \sigma_{\varepsilon\tau}^2 + \sigma_{j\varepsilon\tau}^2$$

so that

$$\hat{\sigma}_{je\tau}^{-1} = \sigma_{\varepsilon\tau}^{-1} (1 - \frac{1}{2} \sigma_{\varepsilon\tau}^{-2} \sigma_{j\varepsilon\tau}^2), \tag{11}$$

and

$$\hat{\sigma}_{je\tau}^{-2} = \sigma_{\varepsilon\tau}^{-2} (1 - \sigma_{\varepsilon\tau}^{-2} \sigma_{j\varepsilon\tau}^2),$$

approximately by Taylor's expansion. So we have

$$\Big(\prod_{\tau=t_o+1}^{t} \hat{\sigma}_{je\tau}^{-1} \Big)^{\frac{1}{t-t_o}} = \Big(\prod_{\tau=t_o+1}^{t} \sigma_{\varepsilon\tau}^{-1} \Big)^{\frac{1}{t-t_o}} \exp\Big(\frac{-1}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\sigma_{j\varepsilon\tau}^2}{\sigma_{\varepsilon\tau}^2} \Big), \tag{12}$$

and

$$\begin{aligned} \frac{1}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{e_{j\tau}^2}{\hat{\sigma}_{je\tau}^2} &= \frac{1}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\varepsilon_\tau^2}{\sigma_{\varepsilon\tau}^2} + \frac{1}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2} \\ &\quad - \frac{1}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\varepsilon_\tau^2 \sigma_{j\varepsilon\tau}^2}{\sigma_{\varepsilon\tau}^4}, \end{aligned}$$

for sufficiently large t_o. Next, using Assumption A, we have

$$\frac{1}{(t-t_o)}\sum_{\tau=t_o+1}^{t-1}\frac{\sigma^2_{j\varepsilon\tau}}{\sigma^2_{\varepsilon\tau}} = \frac{1}{(t-t_o)\sigma^4_{\varepsilon t}}\sum_{\tau=t_o+1}^{t-1}\sigma^2_{\varepsilon\tau}\sigma^2_{j\varepsilon\tau} + o_p(1),$$

and

$$\frac{1}{(t-t_o)}\sum_{\tau=t_o+1}^{t-1}\frac{\varepsilon^2_\tau\sigma^2_{j\varepsilon\tau}}{\sigma^4_{\varepsilon\tau}} = \frac{1}{(t-t_o)\sigma^4_{\varepsilon t}}\sum_{\tau=t_o+1}^{t-1}\varepsilon^2_\tau\sigma^2_{j\varepsilon\tau} + o_p(1),$$

so that

$$\begin{aligned}
&\frac{1}{(t-t_o)}\sum_{\tau=t_o+1}^{t-1}\Big(\frac{\sigma^2_{j\varepsilon\tau}}{\sigma^2_{\varepsilon\tau}} - \frac{\varepsilon^2_\tau\sigma^2_{j\varepsilon\tau}}{\sigma^4_{\varepsilon\tau}}\Big)\\
= \quad & \frac{1}{(t-t_o)\sigma^4_{\varepsilon_t}}\sum_{\tau=t_o+1}^{t-1}\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)\sigma^2_{j\varepsilon\tau} + o_p(1).
\end{aligned}$$

Observe that

$$\frac{1}{(t-1)}\sum_{\tau=1}^{t-1}\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)\sigma^2_{j\varepsilon\tau} = O_p(t^{-\frac{1}{2}-\beta_j}),$$

because

$$\frac{1}{(t-1)}\sum_{\tau=1}^{t-1}E\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)\sigma^2_{j\varepsilon\tau} = 0, \tag{13}$$

and

$$\begin{aligned}
\mathrm{var}\Big(\frac{1}{(t-1)}\sum_{\tau=1}^{t-1}\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)\sigma^2_{j\varepsilon\tau}\Big) &= \frac{1}{(t-1)^2}\sum_{\tau=1}^{t-1}E\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)^2(\sigma^2_{j\varepsilon\tau})^2\\
&\le M\frac{1}{(t-1)^2}\sum_{\tau=1}^{t-1}(\sigma^2_{j\varepsilon\tau})^2 = O(t^{-1-2\beta_j}).
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{1}{(t-t_o)}\sum_{\tau=t_o+1}^{t-1}\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)\sigma^2_{j\varepsilon\tau} &= \frac{t-1}{(t-t_o)}\frac{1}{(t-1)}\sum_{\tau=1}^{t-1}\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)\sigma^2_{j\varepsilon\tau}\\
&\quad -\frac{t_o}{(t-t_o)}\frac{1}{t_o}\sum_{\tau=1}^{t_o}\Big(\sigma^2_{\varepsilon\tau}-\varepsilon^2_\tau\Big)\sigma^2_{j\varepsilon\tau}\\
&= O_p(t^{-\frac{1}{2}-\beta_j})\\
&\to 0
\end{aligned} \tag{14}$$

as $t \to \infty$, $t_o \to \infty$ and $\frac{t_o}{t} \to 0$. (12), (13) and (14) together gives rise to

$$\begin{aligned}
&\Big\{ \prod_{\tau=t_o+1}^{t} \hat{\sigma}_{je\tau}^{-1} \exp\Big(-\frac{1}{2} \sum_{\tau=t_o+1}^{t-1} \frac{e_{j\tau}^2}{\hat{\sigma}_{je\tau}^2} \Big) \Big\}^{\frac{1}{t-t_o}} \\
= \; & \Big(\prod_{\tau=t_o+1}^{t} \sigma_{\varepsilon\tau}^{-1} \Big)^{\frac{1}{t-t_o}} \exp\Big(-\frac{1}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\varepsilon_\tau^2}{\sigma_{\varepsilon\tau}^2} \Big) \\
& \times \exp\Big(-\frac{1}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2} \Big) \qquad (15)
\end{aligned}$$

Then, it follows from (15) that

$$\begin{aligned}
\left(\frac{\hat{\omega}_{jt}^{s-AFTER}}{\hat{\omega}_{kt}^{s-AFTER}} \right)^{\frac{t^{\beta_k}}{t-t_o}} &= \frac{\exp\Big(-t^{(\beta_k-\beta_j)} \frac{t^{\beta_j}}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2} \Big)}{\exp\Big(-\frac{t^{\beta_k}}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} \frac{\varepsilon_{k\tau}^2}{\sigma_{\varepsilon\tau}^2} \Big)} \\
&= \frac{\exp\Big(-t^{(\beta_k-\beta_j)} \frac{t^{\beta_j}}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} E\Big(\frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2}\Big) \Big)}{\exp\Big(-\frac{t^{\beta_k}}{2(t-t_o)} \sum_{\tau=t_o+1}^{t-1} E\Big(\frac{\varepsilon_{k\tau}^2}{\sigma_{\varepsilon\tau}^2}\Big) \Big)} \\
&\to \begin{cases} 0 & \text{if } \beta_j < \beta_k, \\ \exp(-\frac{1}{2} d_{jk}) & \text{if } \beta_j = \beta_k, \\ \exp(-\frac{1}{2} s_k) & \text{if } \beta_j > \beta_k, \end{cases} \qquad (16)
\end{aligned}$$

as $t_o \to \infty$, $t \to \infty$ and $\frac{t_o}{t} \to \infty$. Note that the second equality in (16) is obtained by utilizing the fact that $t^{-1+\beta_j} \sum_{\tau=1}^{t} \Big[\frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2} - E\Big(\frac{\varepsilon_{j\tau}^2}{\sigma_{\varepsilon\tau}^2}\Big) \Big] = O_p(t^{\beta_i - \frac{1+\gamma_j}{2}})$, a result implied by Assumption B(ii), and by neglecting the terms of small orders. From (16), we see immediately that when $0 < \beta_j < \beta_k < \frac{1}{2}$,

$$\left(\frac{\hat{\omega}_{jt}^{s-AFTER}}{\hat{\omega}_{kt}^{s-AFTER}} \right) \to 0,$$

and hence

$$\hat{\omega}_{jt}^{s-AFTER} \to 0, \qquad (17)$$

as $t_o \to \infty$, $t \to \infty$ and $\frac{t_o}{t} \to \infty$. By the same argument, we have

$$\hat{\omega}_{kt}^{s-AFTER} \to 0,$$

provided $\beta_j > \beta_k$ for all $k \neq j$ so that

$$\hat{\omega}_{jt}^{s-AFTER} \to 1, \qquad (18)$$

as $t_o \to \infty$, $t \to \infty$ and $\frac{t_o}{t} \to \infty$ since $\sum_{k=1}^{n} \hat{\omega}_{kt}^{s-AFTER} = 1$. Note that (17) also holds when $\beta_j = \beta_k$ with $d_{jk} > 0$ because $0 < \exp(-\frac{1}{2}d_{jk}) < 1$. By contrast, we will obtain (18) when $\beta_j = \beta_k$ and $d_{jk} < 0$ for all $k \neq j$. This follows because

$$\left(\frac{\hat{\omega}_{jt}^{s-AFTER}}{\hat{\omega}_{kt}^{s-AFTER}}\right) \to \infty,$$

in view of (16) and the fact that $\exp(-\frac{1}{2}d_{jk}) > 1$, and hence

$$\hat{\omega}_{kt}^{s-AFTER} \to 0,$$

for all $k \neq j$ as $t_o \to \infty$, $t \to \infty$ and $\frac{t_o}{t} \to \infty$. Now It remans to verify Proposition (iii). Observe that $\beta_{(1)} = \cdots = \beta_{(m)} > \beta_{(m+1)} \geq \cdots \geq \beta_{(n)}$, and hence for any $\beta_j \in B_m^c$,

$$\hat{\omega}_{jt}^{s-AFTER} \to 0,$$

by virtue of Proposition 1(i) as $t_o \to \infty$, $t \to \infty$ and $\frac{t_o}{t} \to \infty$. Therefore for any $\beta_j, \beta_k \in B_m$ such that $j \neq k$ and $d_{jk} = 0$, we have

$$\hat{\omega}_{jt}^{s-AFTER} = \frac{\prod_{\tau=t_o+1}^{t-1} \hat{\sigma}_{je\tau}^{-1} \exp(-\frac{1}{2}\sum_{\tau=t_o+1}^{t-1} \frac{e_{j\tau}^2}{\hat{\sigma}_{je\tau}^2})}{\sum_{\beta_k \in B_m} \prod_{\tau=t_o+1}^{t-1} \hat{\sigma}_{ke\tau}^{-1} \exp(-\frac{1}{2}\sum_{\tau=t_o+1}^{t-1} \frac{e_{k\tau}^2}{\hat{\sigma}_{ke\tau}^2})}, \tag{19}$$

approximately for large t_o and $t >> t_o$.

□

Proof of Proposition 2. The proof of Proposition 2 is straightforward and similar to the proof of Proposition 1(i) and (ii) and hence neglected here. It can be obtained upon request.

□

Proof of Proposition 3. Let $\sigma_{je\tau}^2 = \sigma_{\varepsilon\tau}^2 + \sigma_{j\varepsilon}^2$. Then using (3), Theorem 5.15 in White (1984), and Assumption C, we can obtain

$$\frac{1}{t^{1/2}} \sum_{\tau=1}^{t} (e_{j\tau}^2 - \sigma_{ej\tau}^2) \to_d N(0, \lambda)$$

for some $\lambda > 0$ as $t \to \infty$. Hence $\hat{\sigma}^2_{je\tau} = \frac{1}{\tau}\sum_{s=1}^{t} e^2_{js} = \sigma^2_{je\tau} + O_p(\frac{1}{\tau^{1/2}})$, which, together with Assumption A, yields

$$\frac{1}{t}\sum_{\tau=1}^{t}\frac{e^2_{j\tau}}{\hat{\sigma}^2_{je\tau}} \to_p 1$$

as $t \to \infty$ so that

$$\begin{aligned}
\frac{1}{t-t_o}\sum_{\tau=t_o+1}^{t}\frac{e^2_{j\tau}}{\hat{\sigma}^2_{je\tau}} &= \frac{t}{t-t_o}\frac{1}{t}\sum_{\tau=1}^{t}\frac{e^2_{j\tau}}{\hat{\sigma}^2_{je\tau}} - \frac{t}{t-t_o}\frac{1}{t}\sum_{\tau=1}^{t_o}\frac{e^2_{j\tau}}{\hat{\sigma}^2_{je\tau}} \\
&= \frac{1}{t}\sum_{\tau=1}^{t}\frac{e^2_{j\tau}}{\hat{\sigma}^2_{je\tau}} + o_p(1) \\
&\to_p 1
\end{aligned}$$

as $t \to \infty$, $t_o \to \infty$ and $\frac{t_o}{t} \to 0$. Therefore

$$\left\{\prod_{\tau=t_o+1}^{t-1}\hat{\sigma}^{-1}_{ej\tau}\exp\left(-\frac{1}{2}\sum_{\tau=t_o+1}^{t}\frac{e^2_{j\tau}}{\hat{\sigma}^2_{je\tau}}\right)\right\}^{\frac{1}{t-t_o}} = \left(\prod_{\tau=t_o+1}^{t-1}\sigma^{-1}_{je\tau}\right)^{\frac{1}{t-t_o}}\exp\left(-\frac{1}{2}\right)$$

approximately by Taylor's expansion as $t \to \infty$, $t_o \to \infty$ and $\frac{t_o}{t} \to 0$. Then

$$\begin{aligned}
\left(\frac{\hat{\omega}^{s-AFTER}_{kt}}{\hat{\omega}^{s-AFTER}_{jt}}\right)^{\frac{1}{t-t_o}} &= \left(\frac{\prod_{\tau=t_o+1}^{t-1}\sigma^{-1}_{ke\tau}}{\prod_{\tau=t_o+1}^{t-1}\sigma^{-1}_{je\tau}}\right)^{\frac{1}{t-t_o}} \\
&= \frac{\sigma^{-1}_{ket}}{\sigma^{-1}_{jet}} \\
&= \left(\frac{\omega^{BG}_{kt}}{\omega^{BG}_{jt}}\right)^{\frac{1}{2}}
\end{aligned}$$

where the second equality is approximated by using Assumption A, the last one is obtained because $\omega^{BG}_{jt} = \frac{\sigma^{-2}_{jet}}{\sum_{j=1}^{n}\sigma^{-2}_{jet}}$ roughly provided $\sigma^2_{\varepsilon\tau} << \sigma^2_{j\varepsilon}$ for all j and all τ. □

Table 1. Simulation results

This table gives the average of the MSEs of the combined forecasts produced by SA, and the mean percentage deviations of the MSE of the method of interest from that of SA. In the Data column, "Imputed" denotes balanced data with missing values imputed by taking the average of available values contemporaneously.

Sample Size	Number of Forecasters	Data	MSE of SA	(Relative MSE - 1) * 100%			
				BG	L_1-AFTER	h-AFTER	s-AFTER
Experiment 1: Time-Invariant Heterogeneity in Forecasting Performance							
		Balanced	1.007	0.00%	1.47%	0.55%	0.50%
50	5	Unbalanced	1.016	0.25%	3.37%	3.89%	2.96%
		Imputed	1.016	0.00%	0.57%	0.20%	0.18%
		Balanced	1.021	-0.02%	1.51%	0.57%	0.51%
50	10	Unbalanced	1.025	0.20%	3.35%	2.60%	2.53%
		Imputed	1.025	-0.01%	0.64%	0.15%	0.12%
		Balanced	1.001	-0.02%	1.03%	0.47%	0.47%
100	5	Unbalanced	1.008	0.05%	1.97%	1.90%	1.88%
		Imputed	1.008	-0.01%	0.51%	0.16%	0.16%
		Balanced	0.996	0.00%	1.74%	0.54%	0.54%
100	20	Unbalanced	0.997	0.03%	3.45%	2.86%	2.98%
		Imputed	0.997	0.00%	0.89%	0.25%	0.25%
		Balanced	0.996	0.00%	2.05%	0.94%	1.00%
200	20	Unbalanced	0.998	0.02%	3.66%	3.28%	3.38%
		Imputed	0.998	0.00%	1.16%	0.39%	0.43%
		Balanced	1.003	0.00%	2.12%	0.89%	0.98%
300	40	Unbalanced	1.003	0.00%	4.03%	3.89%	3.93%
		Imputed	1.003	0.00%	1.24%	0.44%	0.49%
Experiment 2: Time-Varying Heterogeneity in Forecasting Performance							
		Balanced	1.023	-0.02%	1.40%	0.51%	0.49%
50	5	Unbalanced	1.033	0.13%	2.69%	2.67%	2.30%
		Imputed	1.033	0.01%	0.73%	0.22%	0.22%
		Balanced	0.995	-0.01%	1.24%	0.36%	0.33%
50	10	Unbalanced	0.999	0.11%	3.26%	2.19%	2.18%
		Imputed	0.999	-0.01%	0.36%	0.05%	0.03%
		Balanced	1.014	0.00%	1.99%	0.95%	0.97%
100	5	Unbalanced	1.023	0.09%	2.61%	2.20%	2.30%
		Imputed	1.023	0.01%	0.90%	0.35%	0.36%
		Balanced	1.000	0.00%	1.79%	0.53%	0.55%
100	20	Unbalanced	1.002	0.04%	3.51%	3.25%	3.19%
		Imputed	1.002	0.00%	0.88%	0.19%	0.20%
		Balanced	1.013	0.00%	2.46%	0.92%	0.95%
200	20	Unbalanced	1.015	0.02%	3.75%	3.33%	3.34%
		Imputed	1.015	0.00%	1.22%	0.39%	0.41%
		Balanced	0.997	0.00%	2.54%	1.01%	1.06%
300	40	Unbalanced	0.997	0.01%	4.12%	3.84%	3.91%
		Imputed	0.997	0.00%	1.29%	0.43%	0.45%

Table 2. MSEs of forecasts made by different combination methods

Method	Current Quarter Forecasts		1-Quarter Ahead Forecasts		2-Quarter Ahead Forecasts		3-Quarter Ahead Forecasts	
	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I
CPI Inflation (CPI)								
BG		**2.854**		8.764		9.088		8.922
ME		3.221		8.675		9.201		8.802
RB		**1.279**		**11.389**		9.950		8.895
SA		3.175		8.773		9.034		8.859
TM		3.169		8.745		9.048		8.978
MLS		**1.702**		8.786		9.031		7.407
L_1-AFTER		1.220		10.519		10.021		8.782
h-AFTER		**1.483**		9.591		9.758		9.301
s-AFTER		**1.466**		9.549		9.742		9.263
GDP Price Deflator Inflation (PGDP)								
BG	1.605	0.944	**3.403**	0.940	**4.212**	0.987	4.840	1.046
ME	1.639	0.986	3.412	0.946	**4.202**	0.994	4.778	**1.098**
RB	1.494	0.962	**3.140**	1.137	**4.706**	**1.273**	6.255	**1.304**
SA	1.687	0.970	3.481	0.940	4.286	0.974	4.869	1.025
TM	1.650	1.001	3.430	0.940	4.254	0.985	4.821	1.048
MLS	**1.250**	0.970	**3.358**	0.939	4.251	0.938	4.872	1.026
L_1-AFTER	**1.291**	0.984	**3.129**	1.098	4.483	**1.240**	**7.521**	**1.308**
h-AFTER	**1.325**	0.875	3.142	**1.074**	4.151	**1.095**	5.491	1.260
s-AFTER	**1.326**	0.891	**3.135**	1.086	4.149	**1.084**	5.460	1.257
Real GDP Growth (RGDP)								
BG	5.890	1.356	**11.657**	3.429	14.696	5.478	18.320	6.883
ME	6.100	1.324	11.231	3.365	14.541	5.449	17.801	6.839
RB	**9.335**	1.410	**14.222**	4.163	16.067	5.875	22.689	6.300
SA	6.038	1.350	11.374	3.423	14.662	5.476	17.971	6.879
TM	5.982	1.344	11.457	3.342	14.471	5.468	17.503	6.868
MLS	6.096	1.352	11.416	3.429	14.701	5.485	**18.170**	6.887
L_1-AFTER	8.300	1.344	**12.520**	**3.868**	**17.370**	5.636	21.615	7.036
h-AFTER	7.350	1.336	12.096	3.475	15.069	5.980	**21.300**	6.887
s-AFTER	7.297	1.338	12.173	3.456	15.022	6.140	21.543	6.915
Unemployment Rate (UNEMP)								
BG	0.049	**0.026**	0.241	0.170	0.490	0.434	0.812	0.938
ME	**0.047**	0.026	0.245	0.173	0.494	0.435	0.810	0.940
RB	0.051	0.027	0.227	0.125	0.501	0.371	0.793	0.871
SA	0.050	0.027	0.242	0.174	0.496	0.448	0.822	0.960
TM	0.050	0.026	0.238	0.173	0.504	0.447	0.814	0.951
MLS	0.049	**0.026**	0.242	0.172	**0.496**	0.444	0.822	0.958
L_1-AFTER	0.053	0.026	0.240	0.136	0.492	0.364	0.812	0.815
h-AFTER	0.051	0.028	0.248	0.153	0.486	0.411	0.798	0.967
s-AFTER	0.052	0.029	0.250	0.153	0.486	0.411	0.801	0.976

* Numbers in bold denote one-sided DM test rejection at 10%, i.e., the method performs significantly better than SA.

Table 3. Performance of h-AFTER and L_1-AFTER using alternative loss functions

Panel I. Performance of h-AFTER and SA benchmark measured by Huber loss

Method	Current Quarter Forecasts		1-Quarter Ahead Forecasts		2-Quarter Ahead Forecasts		3-Quarter Ahead Forecasts	
	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I
GDP Price Deflator Inflation (PGDP)								
SA	1.187	0.830	1.959	0.820	2.277	0.857	2.527	0.914
h-AFTER	0.998	0.750	1.755	0.892	2.239	0.938	2.702	1.067
Unemployment Rate (UNEMP)								
SA	0.050	0.027	0.235	0.174	0.435	0.378	0.665	0.662
h-AFTER	0.051	0.028	0.242	0.153	0.438	0.353	0.657	0.657

Panel II. Performance of L_1-AFTER and SA benchmark measured by MAE

Method	Current Quarter Forecasts		1-Quarter Ahead Forecasts		2-Quarter Ahead Forecasts		3-Quarter Ahead Forecasts	
	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I	1968:IV to 1990:IV	2000:I to 2013:I
GDP Price Deflator Inflation (PGDP)								
SA	0.974	0.795	1.402	0.820	1.575	0.850	1.711	0.874
L_1-AFTER	0.888	0.796	1.267	0.876	1.599	0.929	2.078	0.966
Unemployment Rate (UNEMP)								
SA	0.176	0.125	0.381	0.296	0.523	0.456	0.675	0.618
L_1-AFTER	0.174	0.129	0.379	0.274	0.535	0.422	0.696	0.560

* L_1-AFTER implemented using d_{ji}.

Table 4. Percentage of significant imputation regressions at each horizon

Variable	Subsample	Number of Regressions Run for Each Forecaster	Percentage of Significant Regressions at Each Horizon			
			Current-Quarter Forecasts	One-quarter Ahead Forecasts	Two-Quarter Ahead Forecasts	Three-Quarter Ahead Forecasts
PGDP	1	88	0.09	0.21	0.28	0.35
	2	52	0.16	0.49	0.57	0.55
UNEMP	1	88	0.17	0.42	0.57	0.61
	2	52	0.24	0.40	0.46	0.48

Figure 1. Overview of data patterns

This figure shows the overall data patterns for all forecasters (using PGDP current-quarter forecasts as example, since the patterns are similar for all other variables and horizons). A black square in the figure represents a data point. Blank represents missing. As forecasters with no forecasts are suppressed in producing the figure, the figure shows fewer forecasters than there are in SPF.

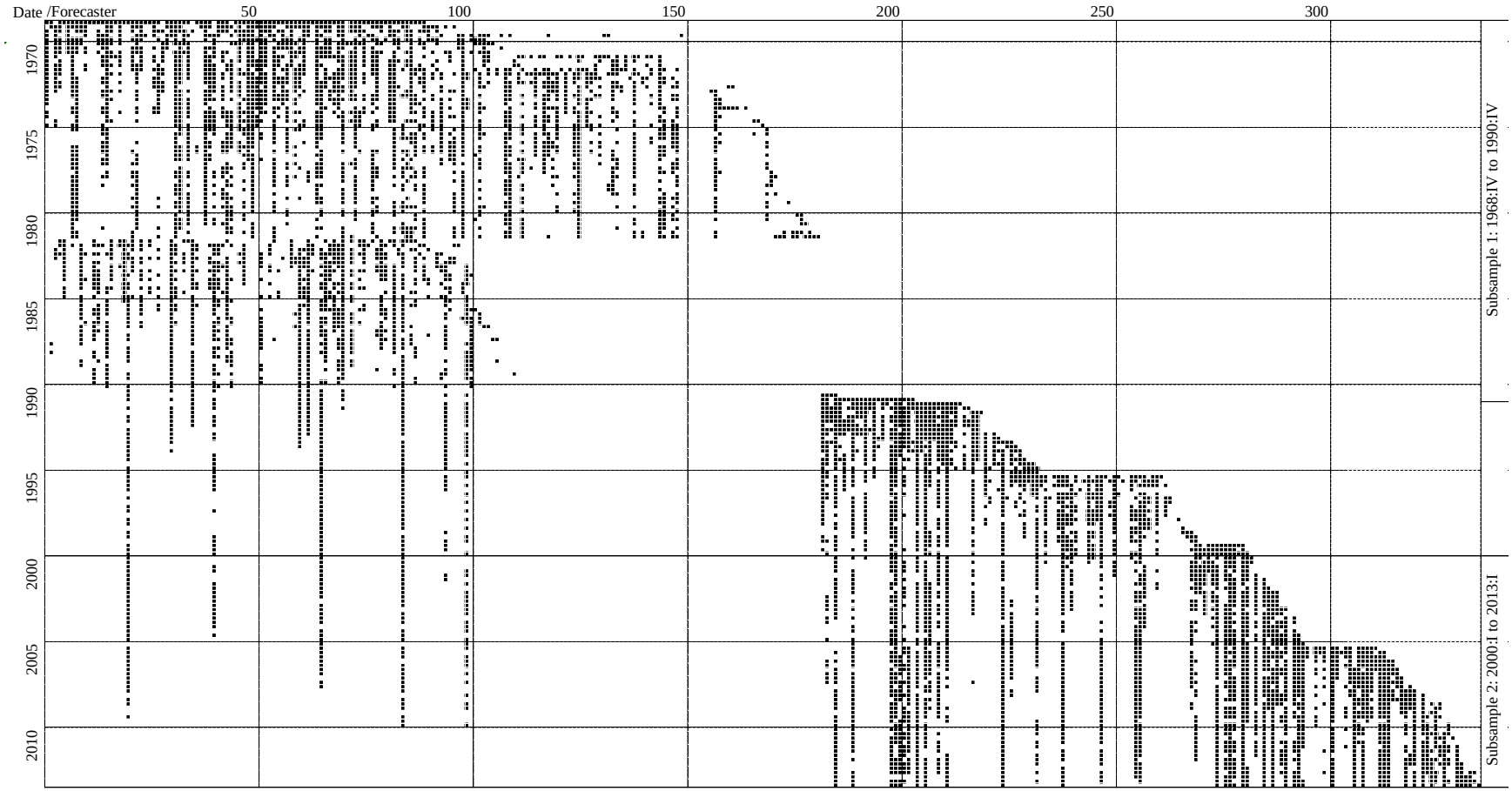
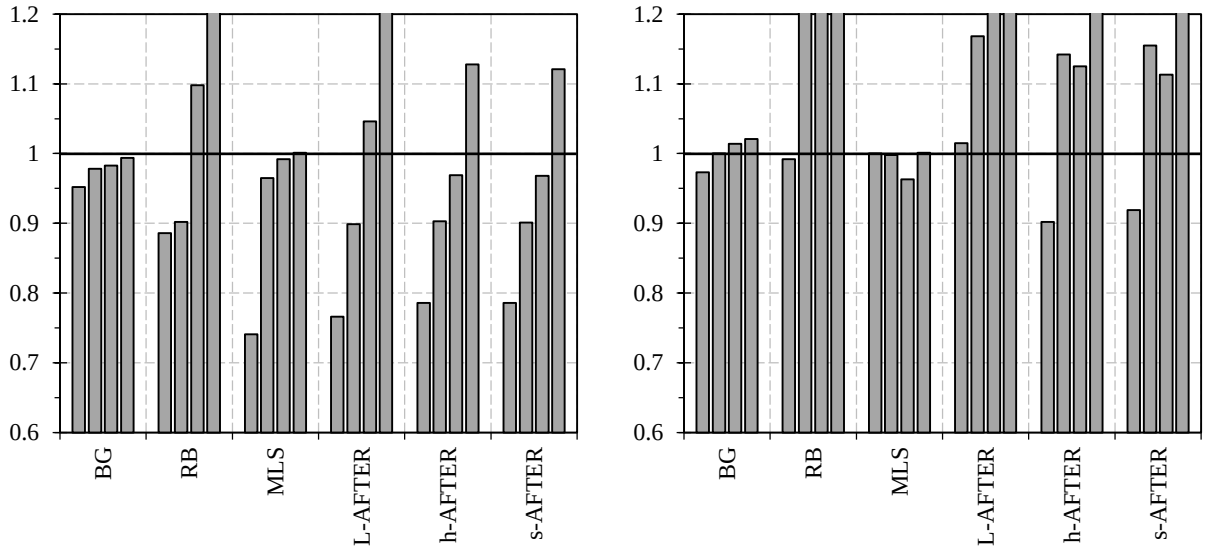


Figure 2. Relative forecast performance of alternative combination methods

Performance reported in this set of figures is MSE of the method of interest relative to MSE of simple average method. Each group of bars represents one method (denoted under the group). The four bars in each group represent current quarter forecasts to 3-quarter ahead forecasts (left to right).

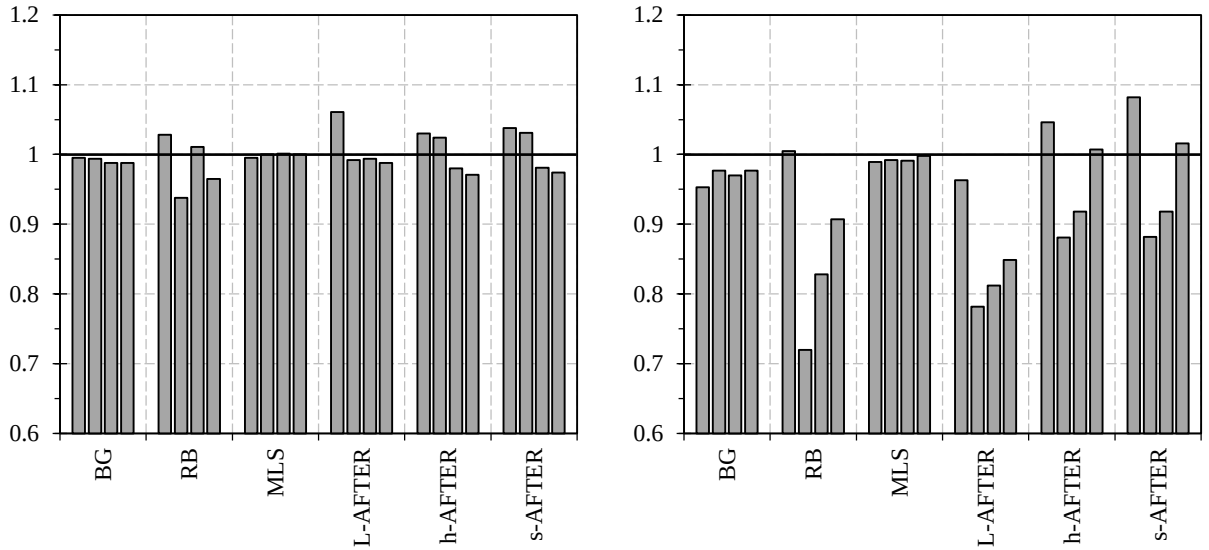
GDP Price Deflator Inflation (PGDP)



Subsample 1: 1968:IV to 1990:IV

Subsample 2: 2000:I to 2013:I

Unemployment Rate (UNEMP)



Subsample 1: 1968:IV to 1990:IV

Subsample 2: 2000:I to 2013:I

Figure 3. L_1-AFTER and h-AFTER evaluated using appropriate loss functions

Performance reported in this set of figures is MAE (for L_1-AFTER) or Huber loss (for h-AFTER) relative to the respective loss of simple average method. The four bars in each group of bars represent current quarter forecasts to 3-quarter ahead forecasts (left to right). The name of the variable, method, and subsample are denoted under the bars.

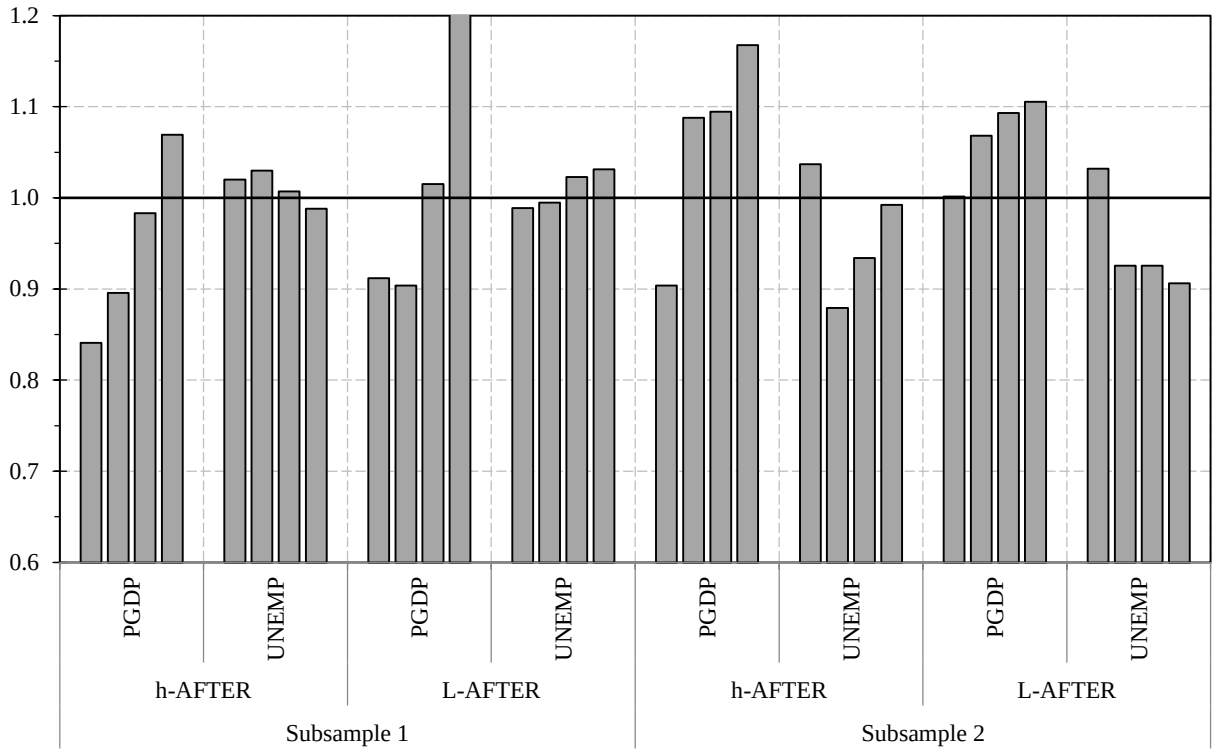


Figure 4. Evolution of weights and performance of individual forecasters - PGDP

GDP Price Deflator Inflation (PGDP), subsample 1, current-quarter forecasts

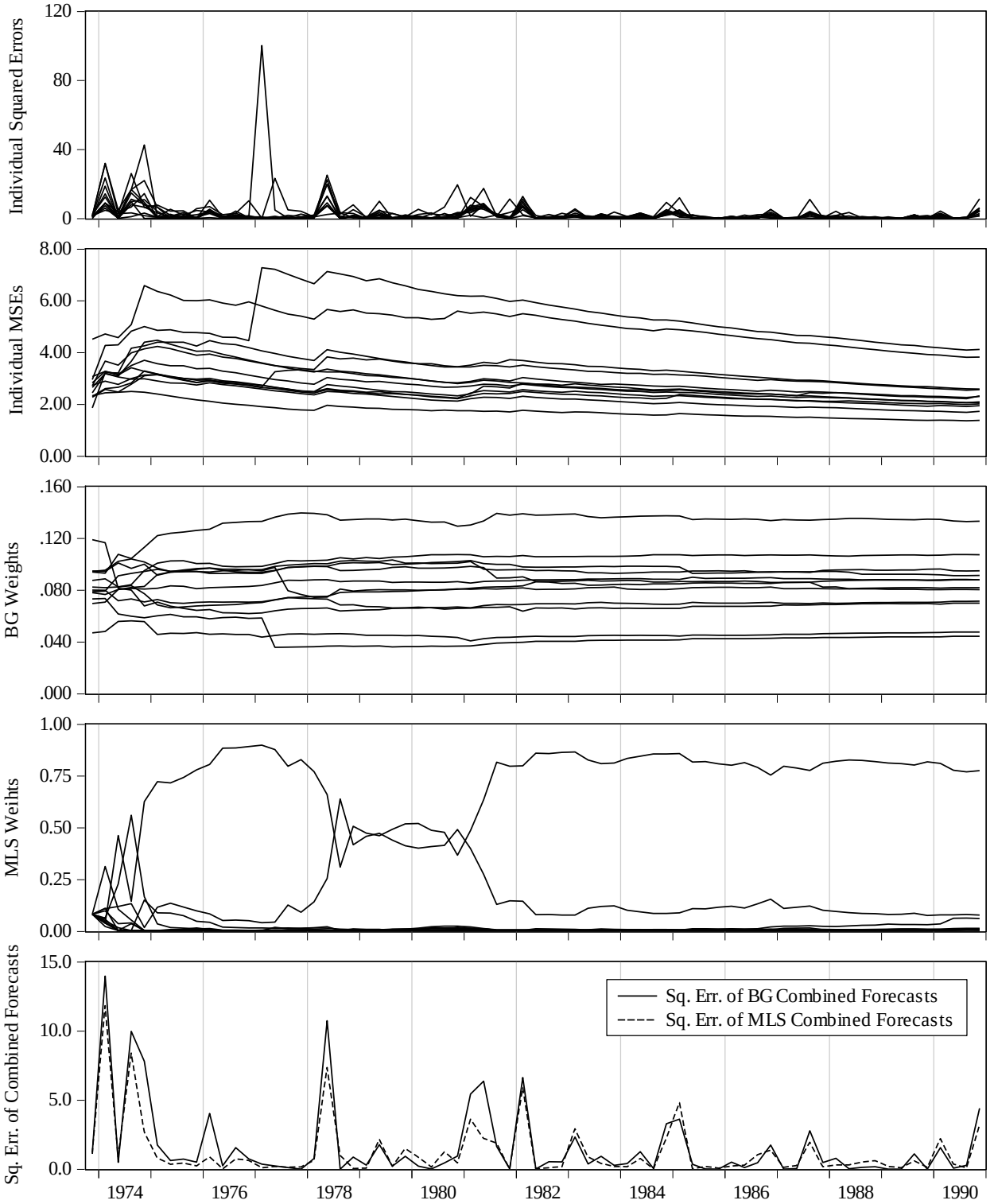


Figure 5: Evolution of weights and performance of individual forecasters - UNEMP

Unemployment Rate (UNEMP), subsample 2, one-quarter ahead forecasts

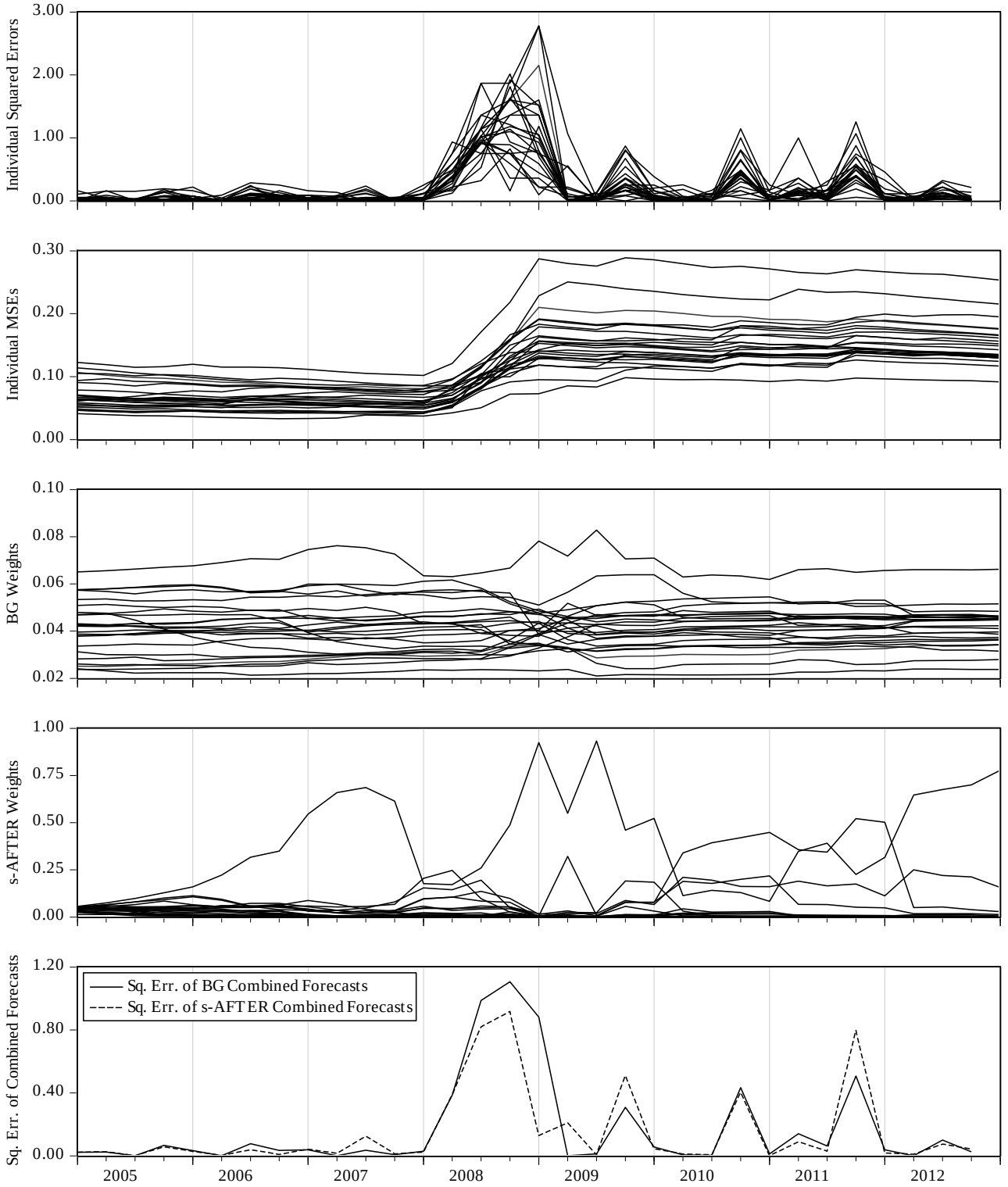


Figure 6. Reacting to Outlier – Behavior of BG and s-AFTER method

Real GDP (RGDP), subsample 1, current-quarter forecasts. In the top four panels, solid line represents forecaster 44, dashed line represents forecaster 65. The outlier is forecaster 44's forecast for the fourth quarter of 1974.

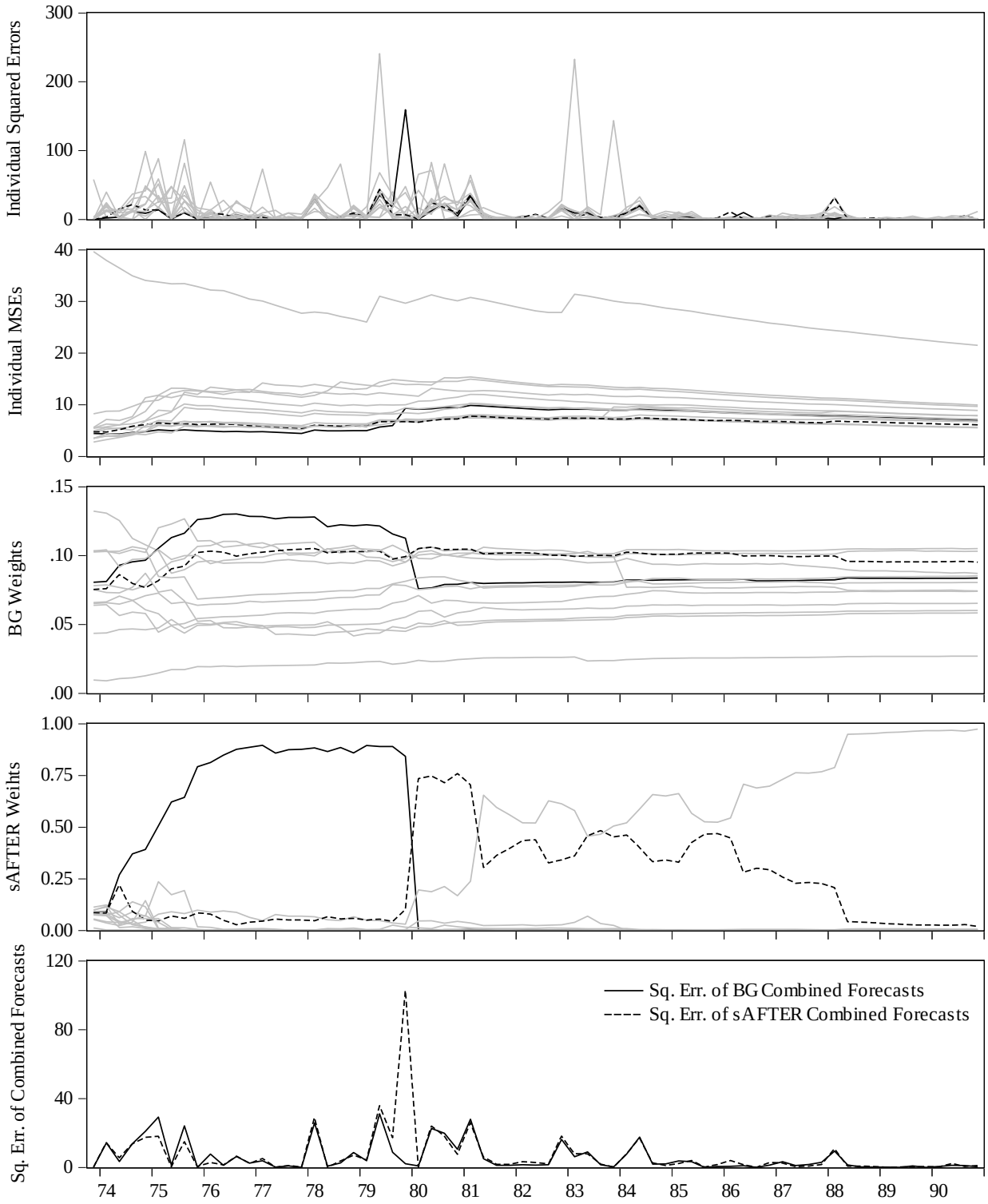


Figure 7. Reacting to Structural Break– Behavior of L-AFTER method

Unemployment rate (UNEMP), subsample 2, one-quarter-ahead forecasts. Shown from top to bottom are individual absolute errors, individual MAEs, weights assigned to individual forecasters by L-AFTER method, and the absolute errors of the combined forecasts produced by L-AFTER method. Behavior of s-AFTER and BG methods for this case can be seen in subfigure 2 of Figure 3. The individual (forecaster 483) receiving the highest weight is the same in BG, s-AFTER, and L-AFTER.