Learning Time-Varying Forecast Combinations

Antoine Mandel, Amir Sani

LEARNING TIME-VARYING FORECAST COMBINATIONS

Antoine Mandel and Amir Sani*

Combining forecasts has been demonstrated as a robust solution to noisy data, structural breaks, unstable forecasters and shifting environmental dynamics. In practice, sophisticated combination methods have failed to consistently outperform the mean over multiple horizons, pools of varying forecasters and different endogenous variables. This paper addresses the challenge to "develop methods better geared to the intermittent and evolving nature of predictive relations", noted in Stock and Watson (2001), by proposing an adaptive nonparametric "meta" approach that provides a time-varying hedge against the performance of the mean for any selected forecast combination approach. This approach arguably solves the so-called "Forecast Combination Puzzle" using a meta-algorithm that adaptively hedges weights between the mean and a specific forecast combination algorithm or pool of forecasters augmented with one or more forecast combination algorithms. Theoretical performance bounds are reported and empirical performance is evaluated on the seven-country macroeconomic output and inflation dataset introduced in Stock and Watson (2001) as well as the Euro-area Survey of Professional Forecasters.

Keywords: Forecast Combinations, Forecast Combination Puzzle, Machine Learning, Econometrics.

1. INTRODUCTION

Macroeconomic forecasts provide crucial inputs to decision-makers addressing monetary and fiscal policy issues. Forecast accuracy depends on a selected model's power to extract useful and meaningful information from available macroeconomic time series. Unfortunately, reality limits forecasting models to finite data samples, incomplete information sets and changing environmental dynamics that result in estimation error and model misspecification. In particular, macroeconomic time series

Université Paris 1 Panthéon-Sorbonne, CNRS, Paris School of Economics, Maison de sciences économiques, 106-112 Boulevard de l'hôpital, 75013 Paris France

{antoine.mandel,amir.sani}@univ-pars1.fr

*Corresponding author.

are predominantly composed of a limited number of noisy aggregated samples across varying economic conditions within an unstable forecasting environment. These challenges often result in inconsistent and misspecified models.

Introduced by Bates and Granger (1969), forecast combination methods have demonstrated an advantage in addressing noisy data, structural breaks, forecasters with inconsistent performance and changing environmental dynamics (Timmermann, 2006). Accordingly, a large body of research has focused on the theoretical and empirical development of complex forecast combination procedures that aim to fully exploit the information content within the available pool of forecasts (see Timmermann, 2006, for a recent survey). However, in empirical settings, these complex combination procedures generally fail to consistently outperform the simple mean (see e.g Stock and Watson, 2004, for the case of output and inflation considered in this paper). The theory shows that existing forecast combination methods lack statistical power and subsequently overfit noise in the small number of macroeconomic samples that are generally available for modeling. In contrast, the mean manages this small sample noise consistently, and therefore well. This "Forecast Combination Puzzle" limits the ability to test new forecast combination approaches (especially in real-time settings), without risking underperformance to the mean.

Building on recent advances in online machine learning literature (in particular Sani et al., 2014), the aim of this paper is to propose a solution to the forecast combination puzzle through a meta-algorithm that provides an online hedge to the mean while exploiting the potential superior predictive ability provided by any alternative algorithm. We recast the forecast combination setting to the online (recursive) optimization setting of "Prediction with Expert Advice", where we propose to learn time-varying "meta-weights" directly from the forecasting performance. Namely, the algorithm $\mathcal{AB}$-Prod, introduced in Sani et al. (2014), provides a meta-structure that combines the weights of a benchmark algorithm $\mathcal{B}$ (the mean in our context) and these of an alternative $\mathcal{A}$ in such a way that performance is never worse than a

precomputed constant to the mean while it learns any superior predictive ability of the alternative. Further, the rate at which this approach "learns" is close to optimal in both "easy" and "worst" case environments. This "meta"-algorithm also comes with theoretical guarantees that make no distribution assumptions on the process generating the target series or the losses and results in performance guaranteed for any stochastic, stationary, non-stationary or shifting environment.

Hence, the proposed algorithm proposes theoretical guarantees tailored to address the "Forecast Combination Puzzle" and provides a robust, data-driven procedure to real-time forecasting without the risks associated to testing new algorithms: it allows decision-makers to use novel forecasters in real-time environments while maintaining an hedge to the mean. We illustrate the empirical performance of this approach by showing it systematically outperform the mean in the seven-country output and inflation dataset introduced in Stock and Watson (2004) as well as in the Euro-area output surveys collected in the survey of professional forecasters (SPF).

The paper proceeds as follows. In Section 2, we briefly review the relevant forecast combination and machine learning literature. In Section 3, we propose a theoretically guaranteed forecast combination approach that "hedges" performance against the mean with synthetic results. In Section 4, we illustrate the workings of these algorithms with synthetic data. In Section 5, we demonstrate the performance of our approach for the forecast of output and inflation in the framework of Stock and Watson (2004) and Euro-area SPF. Section 6 concludes.

2. LITERATURE REVIEW

Real macroeconomic data is observed at an aggregate level and often composed of a small sample of time series observations. Traditional least-squares forecasters often fail to forecast such series due to the limited number of samples, noise and model misspecification (Timmermann, 2006; Diebold, 1989; Diebold and Pauly, 1990; Hendry and Clements, 2004). Additionally, macroeconomic models often depend on the configuration of shocks hitting the economy, policy regimes and other

institutional factors. This unstable real-world environment results in inconsistent forecasters.

Forecast combination approaches offer a simple procedure for exploiting the information content of candidate forecasters, while ignoring the need for explicit model selection[1]. Theoretical results from the literature demonstrate that gains achieved through combination weights are caused by this forecast model instability, where increasing instability in individual forecasts results in a larger advantage (see e.g. Hendry and Clements (2004); Diebold and Pauly (1987); Clements and Hendry (1998, 1999, 2006); Pesaran and Timmermann (2005); Timmermann (2006); Aiolfi et al. (2010)). More specifically, they provide a robust solution to small sample sizes, noise, regime shifts, model misspecification, diverse information sets, unstable forecasters and provide an efficient way to improve forecasting performance by diversifying over a pool of forecasts (for a survey, see Diebold and Lopez (1996); Newbold and Harvey (2002); Clements et al. (2002); Clemen (1989); Timmermann (2006); Huang and Lee (2010).). Practical successes in the forecast combination literature include output and inflation (Stock and Watson, 2001), interest rates (Guidolin and Timmermann, 2009), money supply (Granziera et al., 2013), monetary policy (Kapetanios et al., 2008), equity premiums (Rapach et al., 2010), commodities (Chen et al., 2008) and realized volatility (Patton and Sheppard, 2009).

Though forecast combination approaches have demonstrated several successes, theoretical results have not resulted in methods that consistently outperform the simple average. Timmermann (2006); Hsiao and Wan (2014) illustrate the specific conditions where the relative gain from the true ex-post optimal weights over an unbiased mean combination are negligible. With regard to out-of-sample performance, Huang and Lee (2010) showed several simple cases where the mean combination approach even outperforms a linear model set to the data generating process.

[1]This work deals with a finite pool of candidate forecasters. Other works address the case of a very large to infinite pools of forecasters. See Elliott et al. (2013, 2015); Uematsu and Tanaka (2015)

This inability to consistently outperform the mean is referred to as the "Forecast Combination Puzzle" and has been explained as the biased weighting of "optimal" weights due to the low predictive content of candidate forecasts (Huang and Lee, 2010). This underperformance to the mean is further explained as the result of finite sample bias, model misspecification, unobserved variables, noise and changes in the underlying process (Stock and Watson, 2001, 2004; Claeskens et al., 2014; Smith and Wallis, 2005, 2009; Clark and McCracken, 2009; Huang and Lee, 2010). Empirical and theoretical results demonstrate no consistent advantage in alternative means (geometric, trimmed, corrected) or the median over different horizons and endogenous variables (Stock and Watson, 2004). One intuition is that these alternatives tend to smooth the forecast density generated by the pool of forecasters in a way that ignores important information. It's also reasonable to assume that these alternatives bias parts of the forecast density without any consideration for their performance over time.

Forecasts are likely to have varying performance due to changing predictive content in the data, shifting regimes, the choice of tuning methods and the addition or removal of model parameters. Given these temporal dynamics, the unbiased weights of the mean may not always provide the best performance. In fact, time-varying combination weights have demonstrated great potential versus the mean (See e.g. Novales and de Fruto (1997); Hoogerheide et al. (2010); LeSage and Magura (1992); Granger and Ramanathan (1984); Yang (2004); LeSage and Magura (1992); Sessions and Chatterjee (1989); Sánchez (2008); Sancetta (2010); Timmermann (2006)). Unfortunately, assumptions underlying many of these time-varying approaches are often too restrictive to be applied in realistic empirical settings. In particular, many of these time-varying combination approaches assume a model on the temporal dynamics, a known or stationary covariance structure (Yang, 2004; Sancetta, 2010; Granger and Ramanathan, 1984; Sessions and Chatterjee, 1989; LeSage and Magura, 1992) or normality conditions on the residuals (see Claeskens et al., 2014), resulting in

inconsistent performance in real-time data environments.

The machine learning literature on "prediction with expert advice" (Cesa-Bianchi and Lugosi, 2006) provides a complementary perspective. It measures via the notion of regret, the efficiency with which an algorithm can learn from the data and provides distribution-free theoretical guarantees on the performance of these algorithms. Of particular concern for our work are the contributions of Even-Dar et al. (2008) and Sani et al. (2014) that provide dual theoretical guarantees to a benchmark and to the best forecaster in hindsight. By setting the mean as a benchmark, we can leverage on these results to provide an answer to the forecast combination puzzle. Namely, we can guarantee an online hedge to the mean while exploiting the potential superior predictive ability provided by any alternative algorithm

3. THEORETICAL RESULTS

The forecast combination problem consists in a setting where a decision-maker has K forecasts of a real variable of interest at his disposal and aims at aggregating the information contained in the pool of forecasts. More precisely, at a sequence of dates $t = h + 1 \cdots T$, the decision-maker has forecasts $(\hat{y}_{1,t|t-h}, \ldots, \hat{y}_{K,t|t-h}) \in \mathbb{R}^K$ available. These K forecasts are then aggregated into a point forecast at time t, $\hat{y}_{t|t-h} \in \mathbb{R}$, for horizon h. In line with the bulk of the forecast combination literature (Timmermann, 2006) and the "Prediction with Expert Advice" framework of Cesa-Bianchi and Lugosi (2006), we shall restrict attention to convex forecast combinations where the decision-maker chooses decision weights $\mathbf{w}_{t|t-h}$ from the decision set $\mathcal{S}$ defined by the K-dimensional simplex $\mathcal{S} := \Delta_K := \left\{\mathbf{w} \in \mathbb{R}_+^K : \sum_{i=1}^K w_i = 1\right\}$ in order to form a combined forecast of the form $\hat{y}_{t|t-h} = \sum_{i=1}^K w_{i,t|t-h}\hat{y}_{i,t|t-h}$. We also follow the forecast combination literature by using the quadratic loss to measure the performance of a forecast $\hat{y}_{i,t|t-h}$ as,

$$l_{i,t} = (y_t - \hat{y}_{i,t|t-h})^2, \tag{1}$$

with the forecast combination loss from decision weights $\mathbf{w}$ as,

$$l_{\mathbf{w},t} = \sum_{i=1}^{K} w_i \hat{y}_{i,t|t-h}. \tag{2}$$

Aggregated over time, these losses yield the mean squared forecast error (MSFE) defined for forecaster i by,

$$\mathbf{MSFE}_i = \frac{1}{T-h+1} \sum_{t=h+1}^{T} l_{i,t}, \tag{3}$$

and respectively, for a forecast combination with (fixed) weights $\mathbf{w}$,

$$\mathbf{MSFE}_{\mathbf{w}} = \frac{1}{T-h+1} \sum_{t=h+1}^{T} l_{\mathbf{w},t}. \tag{4}$$

A large share of the forecast combination literature has then focused on determining fixed optimal weights that minimize the mean squared forecast error. However, in practice, theoretically determined optimal weights have failed to consistently outperform the mean, i.e. the forecast combination that assigns fixed uniform $1/K$ weights over each of the K forecasts. This negative result is usually referred to as the "Forecast Combination Puzzle." Non-stationarity and regime switches are intuitive explanations for the presence of this puzzle. Time-varying weights are a natural approach to handle these issues, yet existing approaches have failed to overcome this puzzle (see the seminal paper by Bates and Granger (1969) and Timmermann (2006) for a survey). In order to shade new light on this problem, we build on recent advances in the online machine learning literature (in particular Sani et al., 2014).

The building blocks of our approach are forecast combination algorithms that provide (time-varying) forecast combination weights at the sequence of dates $t = 1 \cdots T-h$. More precisely, the information available to the decision-maker at time t is given by the history of forecasts and realizations:

$$\mathcal{H}_t = \{(y_1, \ldots, y_t), (\hat{y}_{1,h+1|h}, \ldots, \hat{y}_{1,t|t-h}), \cdots, (\hat{y}_{K,h+1|h}, \ldots, \hat{y}_{K,t|t-h})\}. \tag{5}$$

A forecast combination algorithm $\mathcal{G}$ is then defined as a series of mappings $(g_t)_{t=1,\ldots,T-h}$, where $g_t : \mathcal{H}_t \to \Delta_K$. The mapping g_t associates to an observation history $h_t \in \mathcal{H}_t$

a vector of weights $g_t(h_t) = (w_{1,t}, \cdots, w_{K,t})$ in the K-dimensional simplex Δ_K and hence aggregates the pool of forecasts available at time t, $(\hat{y}_{1,t|t-h}, \ldots, \hat{y}_{K,t|t-h})$, into a single point forecast $\hat{y}_{t|t-h} = \sum_{i=1}^{K} w_{i,t}\hat{y}_{i,t|t-h}$.

Baseline combination methods from the macro-economic forecast combination literature include the mean, trimmed mean and median. Within the online learning literature, a large class of algorithms have the following structure: each forecaster is characterized by a score $\lambda_{i,t}$ that the decision-maker sequentially updates on the basis of observed losses $l_{i,t}$ and uses to choose his mixture over forecasts (commonly in the form of probability weights which are assigned over the forecasts). A prominent example of such algorithms is the exponentially weighted average forecaster **Hedge** (see e.g. Freund and Schapire (1997); Littlestone and Warmuth (1994); Vovk (1990); Cesa-Bianchi and Lugosi (2006)), which exponentially updates the mixture over forecasts according to the gradient of their losses (see the algorithm protocol in Figure 1).

Input: Learning rate $\eta > 0$, Experts $\{1, \ldots, K\}$, Decision set $\mathcal{S} = \Delta_K$, Rounds T, Losses $\mathcal{L} = [0,1]^K$.

Initialize scores: $\lambda_{i,1} = \frac{1}{K}, \forall i$.

For all $t = 1, \ldots, T$**, repeat**

1. Simultaneously
   - Environment chooses losses $\mathbf{l}_t \in \mathcal{L}$.
   - Learner chooses decision $\mathbf{w}_t \in \mathcal{S}$, where $w_{i,t} = \frac{\lambda_{i,t}}{\sum_{i=1}^{K} \lambda_{i,t}}, \forall i$.
2. Environment reveals losses $\mathbf{l}_t$.
3. Learner suffers loss $\mathbf{w}_t^\top \mathbf{l}_t$.
4. Learner updates scores λ_{t+1}, as $\lambda_{i,t+1} = \lambda_{i,t} \exp(-\eta l_{i,t}), \forall i$.

end for

Figure 1: **Hedge**

The performance of these algorithms is conventionally measured using the notion of regret that accounts for the learning properties of the algorithm over time. Namely, if one denotes by $l_{\mathcal{G},\tau}$ the loss of algorithm $\mathcal{G}$ in period τ and by $L_{\mathcal{G},t} = \sum_{\tau=1}^{t} l_{\mathcal{G},\tau}$ its cumulative loss up to period t, the regret of an algorithm $\mathcal{G}$ with regard to another algorithm $\mathcal{H}$ up to time t is defined as $\mathcal{R}_{\mathcal{G},t}(\mathcal{H}) = L_{\mathcal{G},t} - L_{\mathcal{H},t}$. With some abuse of notation, we can also define the regret of an algorithm $\mathcal{G}$ up to time t against a forecaster i as $\mathcal{R}_{\mathcal{G},t}(i)$. Note that the regret is generally measured with respect to the best forecaster in hindsight $i^* := \arg\min_{i \in K} L_{i,T}$. If the cumulative regret grows at a rate that is less than linear, the algorithm approaches the performance of the best forecaster in hindsight and is said to be "learning"[2]. **Hedge** offers learning properties which are "optimal" in the worst case with respect to the best forecaster in hindsight. Namely, one has:

THEOREM 1 *(Cesa-Bianchi and Lugosi, 2006) For any finite horizon T and forecasters K, the regret upper bound for **Hedge** satisfies,*

$$\mathcal{R}_{\textbf{\textit{Hedge}},T}(i) \leq \frac{T\eta}{8} + \frac{\log K}{\eta},$$

against any forecaster i, and the following regret bound with optimized learning rate $\eta = \sqrt{\frac{8 \log K}{T}}$,

$$\mathcal{R}_{\textbf{\textit{Hedge}},T}(i) \leq \sqrt{\frac{T}{2} \log K}.$$

Hence, **Hedge** achieves the worst-case regret $\mathcal{O}(\sqrt{T \log K})$ to any forecaster i, including the ex-post optimal choice i^* (Cesa-Bianchi and Lugosi, 2006). Note that a "worst-case" guarantee holds in all possible realizations of the loss sequence. Also note that according to Theorem 2.2 in (Cesa-Bianchi and Lugosi, 2006), this bound can not be improved upon by an exponentially weighted average forecaster.

[2]Also note that the regret does not characterize the absolute performance of the algorithm $\mathcal{G}$. A negative regret is possible if $\mathcal{G}$ outperforms the optimal candidate forecaster.

Now, in the context of (macro-economic) forecasting, the best forecaster in hindsight might not be the appropriate benchmark. For example, regime switches may provide an antagonistic realization of the sequence which does not clearly favor a single forecaster in hindsight. This results in an incentive to hedge against the risk that the selected forecaster may not be the best choice over time. According to the forecast combination puzzle, a logical benchmark in this case is the mean combination forecaster μ.

A naive approach to benchmark against the mean would be to add the mean forecast combination as an additional forecaster to the pool for **Hedge**. According to Theorem 1, this would result in an equivalent $\mathcal{O}(\sqrt{T})$ regret guarantee to the mean and other forecasters that compose the pool. Now, Hedge is optimal when it comes to provide a uniform upper bound on the regret with respect to all the forecasters in the pool. Our objective rather is to bound specifically the regret to the mean (or another benchmark), while maintaining a close to optimal $\mathcal{O}(\sqrt{T})$ bound on the regret with respect to the remaining forecasters.

Input: Learning rate $\eta \in (0, 1/2]$, Experts $\{1, \dots, K\}$, Decision set $\mathcal{S} = \Delta_K$, Rounds T and Losses $\mathcal{L} \in [0,1]^K$.

Initialize scores: $\lambda_{i,1} = \frac{1}{K}, \forall i$.

For all $t = 1, \dots, T$**, repeat**

1. Simultaneously
   - Environment chooses losses $\mathbf{l}_t \in \mathcal{L}$.
   - Learner chooses decision $\mathbf{w}_t \in \mathcal{S}$, where $w_{i,t} = \frac{\lambda_{i,t}}{\sum_{i=1}^{K} \lambda_{i,t}}, \forall i$.
2. Environment reveals losses $\mathbf{l}_t$.
3. Learner suffers loss $\mathbf{w}_t^\top \mathbf{l}_t$.
4. Learner updates scores λ_{t+1}, as $\lambda_{i,t+1} = \lambda_{i,t}(1 - \eta l_{i,t}), \forall i$.

end for

Figure 2: **Prod**

From an analytical perspective, this requires determining an analytic expression of the regret bound that is specific to a forecaster. Such an analytic expression can be obtained by replacing the exponential weight update, $e^{\eta x}$, of **Hedge** by its linear approximation, $1 + \eta x$. The resulting algorithm, usually referred to as **Prod** in the online learning literature, is described in Figure 2. It provides an explicit characterization of the regret with respect to a forecaster as a function of its losses. Namely, one has:

THEOREM 2 *(Cesa-Bianchi and Lugosi, 2006; Cesa-Bianchi et al., 2007) For any T and learning rate $\eta \in (0, 1/2]$, **Prod** satisfies the following second-order regret bound,*

$$\begin{aligned}\mathcal{R}_{\textbf{\textit{Prod}},T}(i) &\leq \eta \sum_{t=1}^{T} l_{i,t}^2 + \frac{\log K}{\eta}, \\ &\leq \eta T + \frac{\log K}{\eta},\end{aligned}$$

for any forecaster i, and the following regret bound with optimized learning rate $\eta = \sqrt{\frac{\log K}{T}}$,

$$\mathcal{R}_{\textbf{\textit{Prod}},T}(i) \leq 2\sqrt{T \log K}. \tag{6}$$

REMARK 1 *The regret bound in **Prod** is said to be "second-order" as it is a function of the sum of squared losses.*

REMARK 2 *Note that the linear approximation in **Prod** results in an asymmetric update that recovers slower than the exponential equivalent. Analytically, the higher-order terms that are missing from the **Prod** update (step 4 in Figure 2) imply a lag with regards to the exponential update in **Hedge** (step 4 in Figure 1). Empirically, the update step in **Prod** requires additional reinforcement (i.e. further updates in the same direction) to yield an effect equivalent to this of the update step in **Hedge**.*

Input: Rounds T, Losses $\mathcal{L} \in [0,1]^K$, Decision set $\mathcal{S} = \Delta_K$,
Base forecasters $\{1, \ldots, K\}$ and Fixed distribution $D \in \mathcal{S}$.
Initialize: Learning rate $\eta = \sqrt{\frac{\log K}{T}}$, weight $\lambda_0 = 1 - \eta$
on allocation D, Base forecaster weights $\mu_i = \frac{\eta}{K}$ for $i \in \{1, \ldots, K\}$,
$w_{i,1} = \mu_i, \forall i \in \{0, \ldots, K\}$.
For all $t = 1, \ldots, T$**, repeat**

1. Simultaneously
   - Environment chooses $\mathbf{l}_t \in \mathcal{L}$.
   - Learner chooses decision $\mathbf{w}_t \in \mathcal{S}$, where
   $$w_{i,t} = \frac{\lambda_{i,t}}{\sum_{i=0}^{K} \lambda_{i,t}},$$
   for $i \in \{0, \ldots, K\}$.
2. Environment reveals $\mathbf{l}_t$.
3. Learner suffers loss $\mathbf{w}_t^\top \mathbf{l}_t$.
4. Learner updates weights λ_{t+1}, where for $i \in \{1, \ldots, K\}$,
   $$\lambda_{i,t+1} = \lambda_{i,t}(1 - \eta(l_{i,t} - l_{0,t})).$$

end for

Figure 3: D-Prod(Even-Dar et al., 2008)

Then, in order to obtain dual regret bounds, one with respect to the benchmark and the other with respect to the remaining forecasters, Even-Dar et al. (2008) introduce an alternative normalization rule in **Prod** for the scores of the different forecasts. Namely, the score of the benchmark is kept fixed while the score of the other forecasters are updated as a function of their relative performance with respect to this of the benchmark. The resulting algorithm D-Prod is described in Figure 3. Its regret with respect to the benchmark is minimal because the "benchmark" portion of the algorithm has, by construction, zero regret while the weight on the

other forecasters increase only if they perform better than the benchmark. Namely, one has:

THEOREM 3 (Even-Dar et al. (2008)) *For any rounds T and forecasters K, $D-Prod$ satisfies the following regret bounds,*

$$\mathcal{R}_{D\text{-}Prod,T}(i) = \mathcal{O}\left(\sqrt{T\log K} + \sqrt{\frac{T}{\log K}}\log T\right),$$

for any $\mathbf{x} \in \mathcal{S}$ and

$$\mathcal{R}_{D\text{-}Prod,T}(i) = \mathcal{R}_{D,T}(i) + \mathcal{O}(1).$$

REMARK 3 *Hence, D-Prod achieves constant regret with respect to the benchmark. As emphasized in Sani (2015), this performance can be achieved because the asymmetry in **Prod** emphasized in Remark 2 is directed in favor of the benchmark. The use of an "exact" algorithm like **Hedge** would on the contrary result in equivalent $\mathcal{O}(\sqrt{T})$ regret with respect to the benchmark and other forecasters in the pool.*

Sani et al. (2014) generalize this result further to a meta-structure that hedges an algorithm $\mathcal{A}$ based on a benchmark algorithm $\mathcal{B}$. More precisely, the algorithm $\mathcal{AB}$-Prod combines two forecast combination algorithms $\mathcal{A}$ (the Alternative) and $\mathcal{B}$ (the Benchmark) into a third "hedged" algorithm $\mathcal{C}$, where $\mathcal{C}$ adapts a weighting over $\mathcal{A}$ and $\mathcal{B}$ according to the history of their performance (see e.g. Figure 4). This approach allows to consider any algorithm as a benchmark (while D-Prod only considered benchmark to a fixed distribution) and moreover allows to refine the bound to Benchmark $\mathcal{B}$ by considering a weighted combination of algorithms $\mathcal{A}$ and $\mathcal{B}$, rather than independent algorithms K. Namely, one obtains a fixed constant $2\log 2$ regret to the benchmark $\mathcal{B}$ under any realizations of the loss sequence:

THEOREM 4 (cf. Theorem 1 in Sani et al. (2014)) *Let $\mathcal{A}$ be any algorithm, $\mathcal{B}$ be any benchmark and D be an upper bound on the benchmark losses $L_{\mathcal{B},T}$. Then*

Input: Algorithms $\mathcal{A}$ and $\mathcal{B}$, the convex subset of $\mathcal{S} \in \mathbb{R}^2$, Rounds T, Losses $\mathcal{L} \in [0,1]^2$ and weight $\lambda_{\mathcal{B}} \in (0,1)$.

Initialize: $\lambda_{\mathcal{A}} = 1 - \lambda_{\mathcal{B}}$, Learning rate $\eta = C\sqrt{\frac{1}{T}} < \frac{1}{2}$, where $C = \sqrt{-\log(1-\lambda_{\mathcal{B}})}$.

For all $t = 1, 2, \ldots, T$**, repeat**

1. Simultaneously
   - Environment chooses $\mathbf{l}_t \in \mathcal{L}$.
   - Learner sets $s_t = \dfrac{\lambda_{\mathcal{A},t}}{\lambda_{\mathcal{A},t} + \lambda_{\mathcal{B}}}$.
2. Environment reveals $\mathbf{l}_t$
3. Learner computes respective losses $l_{\mathcal{A},t}$ and $l_{\mathcal{B},t}$.
4. Learner suffers loss $l_{\mathcal{AB}-\text{Prod},t} = s_t l_{\mathcal{A},t} + (1 - s_t) l_{\mathcal{B},t}$.
5. Compute $\delta_t = l_{\mathcal{B},t} - l_{\mathcal{A},t}$
6. Set $\lambda_{\mathcal{A},t+1} = \lambda_{\mathcal{A},t}(1 + \eta\delta_t)$

end for

Figure 4: $\mathcal{AB}$-Prod

setting weight $\lambda_{\mathcal{B}} \in (0,1)$, $\lambda_{\mathcal{A}} = 1 - \lambda_{\mathcal{B}}$, *Learning rate* $\eta = C\sqrt{\frac{1}{T}} < \frac{1}{2}$, *where* $C = \sqrt{-\log(1-\lambda_{\mathcal{B}})}$ *simultaneously guarantees,*

$$\mathcal{R}_{\mathcal{AB}\text{-}Prod,T}(i) \leq \mathcal{R}_{\mathcal{A},T}(i) + 2C\sqrt{D},$$

for any forecaster i *and,*

$$\mathcal{R}_{\mathcal{AB}\text{-}Prod,T}(i) \leq \mathcal{R}_{\mathcal{B},T}(i) + 2\log 2,$$

against any assignment of the loss sequence.

The asymmetric update in *Prod* allows $\mathcal{AB}$-Prod to adaptively trade-off between $\mathcal{A}$ and $\mathcal{B}$ with an asymmetry that provides the necessary momentum to outperform algorithms that rely on a differencing to determine their weights [3]. Further, by

[3] One might consider a similar $\mathcal{AB}$ structure using **Hedge** in place of **Prod** (e.g. $\mathcal{AB}$-Hedge),

initializing the benchmark $\mathcal{B}$ with a large starting weight, $\mathcal{AB}$-Prod requires that the alternative algorithm $\mathcal{A}$ demonstrate an explicit advantage before the $\mathcal{AB}$-Prod updates begin to prefer $\mathcal{A}$ over $\mathcal{B}$. This results in the $\mathcal{O}(\sqrt{T})$ regret to any forecaster i and a constant $\mathcal{O}(2\log 2)$ to the performance of the Benchmark $\mathcal{B}$.

It follows that setting $\mathcal{B} = \mu$ "solves" the "Forecast Combination Puzzle" in the sense that $\mathcal{AB}$-Prod then provides constant and distribution-free theoretical guarantees for the relative performance with respect to the mean combination μ while exploiting any superior predictive ability of an alternative algorithm $\mathcal{A}$, which can be chosen arbitrarily by the decision-maker. Namely, one has:

THEOREM 5 *Let $\mathcal{A}$ be any algorithm, $\mathcal{B}$ be the mean combination μ and D be an upper bound on the benchmark losses $L_{\mathcal{B},T}$. Then setting weight $\lambda_{\mathcal{B}} \in (0,1)$, $\lambda_{\mathcal{A}} = 1 - \lambda_{\mathcal{B}}$, Learning rate $\eta = C\sqrt{\frac{1}{T}} < \frac{1}{2}$, where $C = \sqrt{-\log(1-\lambda_{\mathcal{B}})}$ simultaneously guarantees,*

$$\mathcal{R}_{\mathcal{AB}\text{-}Prod,T}(i) \leq \mathcal{R}_{\mathcal{A},T}(i) + 2C\sqrt{D},$$

for any forecaster i, and,

$$\mathcal{R}_{\mathcal{AB}\text{-}Prod,T}(i) \leq \mathcal{R}_{\mu,T}(i) + 2\log 2.$$

4. SYNTHETIC RESULTS

In order to demonstrate the ability of $\mathcal{AB}$-Prod to effectively and rapidly adapt to the relative performance of the alternative algorithm $\mathcal{A}$, while offering the protection of an explicit benchmark $\mathcal{B}$, we first construct two synthetic scenarios of loss sequences (according to the method proposed in de Rooij et al., 2014) and presented in Sani et al. (2014); Sani (2015), which correspond respectively to situations where the alternative and the benchmark (i.e. the mean in our example) perform poorly. The alternative algorithm we consider in these scenarios is **AdaHedge**, which is

but this would result in $\mathcal{O}(\sqrt{T})$ regret to both $\mathcal{A}$ and $\mathcal{B}$. This was also demonstrated in Sani et al. (2014); Sani (2015)

a variant of **Hedge** with an adaptive learning rate η (see de Rooij et al., 2014). The two scenarios consist of 1000 loss observations for two forecasters[4]. Losses have values in $\{0, 1\}$, i.e. each expert can be right or wrong. Regret and RMSE results are presented for the mean forecast combination μ, **AdaHedge**, and $\mathcal{AB}$-Prod with $\mathcal{A}$ set to AdaHedge and $\mathcal{B}$ set to μ in tables I and II.

	Mean	AdaHedge	AB-Prod(AdaHedge,μ)
Scenario 0	0.5	13.187577	0.507855
Scenario 1	248.5	2.250753	89.459579

Table I: Regret

	Mean	AdaHedge	AB-Prod(AdaHedge,μ)
Scenario 0	1.0	1.025401	1.000016
Scenario 1	1.0	0.507009	0.681601

Table II: Relative MSE

In Scenario 1, the mean combination approach clearly outperforms AdaHedge with regard to its Regret. This is also reflected in its RMSE. Note that the Regret provides a clearer picture of the performance over time. The performance is such that it is impossible to beat the mean combination approach, especially while changing weights over time. This suggests that neither of the forecaster loss sequences shows a substantial advantage. Accordingly, **AdaHedge** is unable to exploit any additional information and underperforms the mean. $\mathcal{AB}$-Prod quickly recognizes the advantage of the mean combination, paying a slight Regret in "learning".

In Scenario 2, the mean combination approach fails and **AdaHedge** offers a substantial performance advantage. This performance is such that it is clear one of the forecaster loss sequences dominates the other. $\mathcal{AB}$-Prod(AdaHedge,μ) recognizes the

[4]The approach generates losses directly for each of the forecasters at each time-step, rather than a prediction for both experts together with a realization.

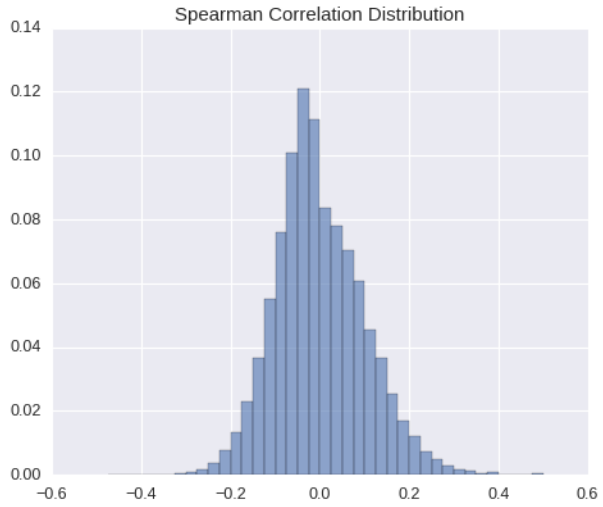

Figure 5: The Spearman correlation distribution corresponding to the loss tuples.

alternative advantage and shifts weights accordingly to perform almost as well as **AdaHedge** (see table I).

To further illustrate the hedging capacity of $\mathcal{AB}$-Prod, we perform a second series of experiments with synthetic data in which we generate $1,000,000$ sequences of 100 losses for two experts by drawing randomly and independently from binomial distributions (with a parameter that varies with the Monte-Carlo simulation). The distribution of the Spearman correlation between the two loss sequences composing the tuples is reported in 5. The aim of this setting is to investigate the performance of four forecast combination algorithms together with their $\mathcal{AB}$-Prod$(\cdot,\mu)$ extensions. More precisely, we consider as baseline algorithms AdaHedge, the time-varying forecast combination approach of Bates-Granger (see Model 1 from Timmermann (2006)), the Recent Best forecaster (which chooses the forecaster that performed best last period) and the random forecaster (ie. choosing one of the forecasters uniformly at random). The corresponding extensions are denoted by $\mathcal{AB}$-Prod(AdaHedge,μ), $\mathcal{AB}$-Prod(Bates-Granger,μ), $\mathcal{AB}$-Prod(Recent Best ,μ) and $\mathcal{AB}$-Prod(Random,μ). In each case, we set $\lambda_{\mathcal{B}} = 0.999$. Results are reported in Figure 6 via the RMSE.

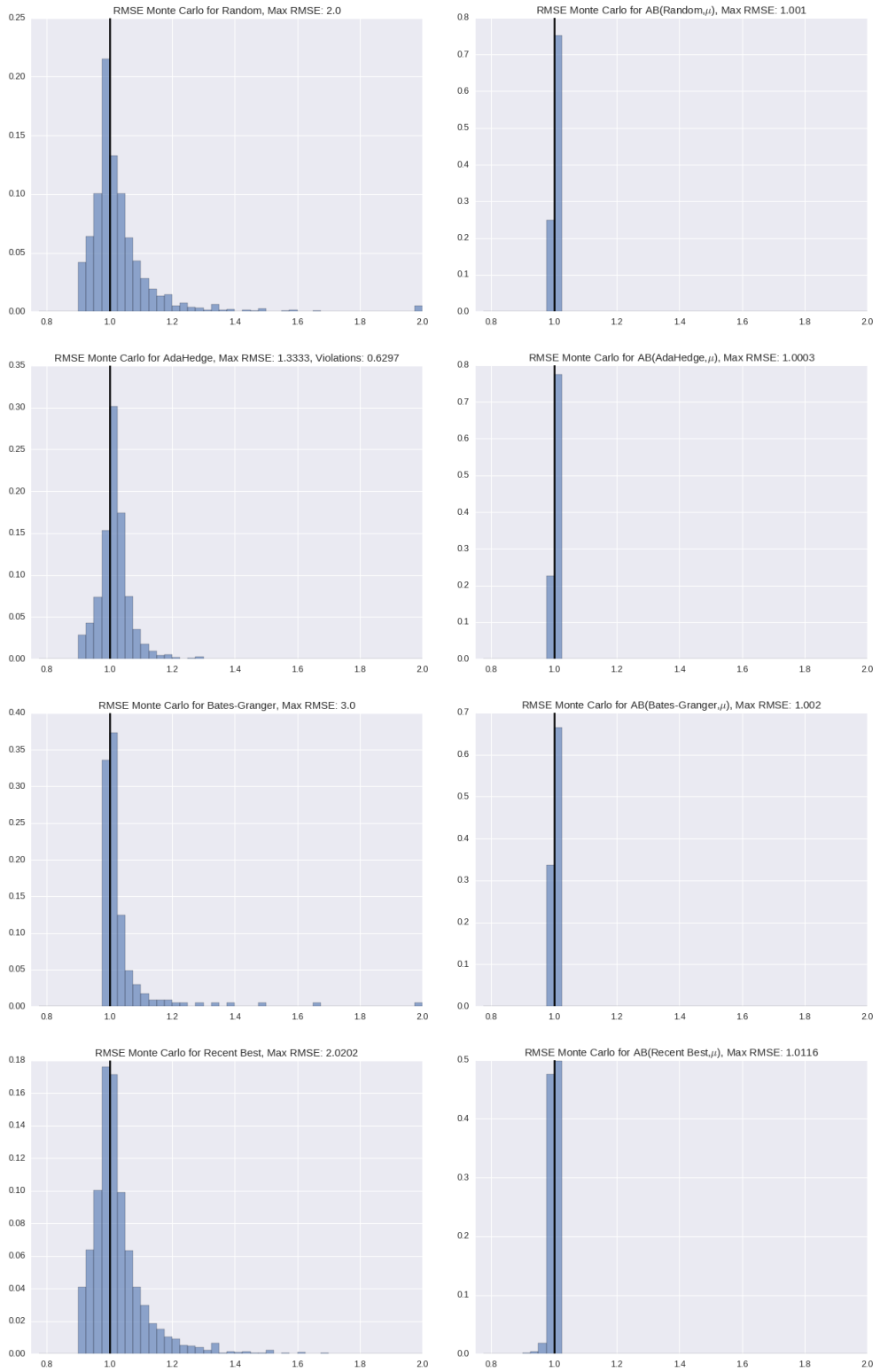

Figure 6: RMSE over synthetic loss sequences for the Random forecaster, AdaHedge, Bates-Granger and the Recent Best algorithms are reported.

For all forecast combination algorithms considered, $\mathcal{AB}$-Prod$(\cdot,\mu)$ demonstrates its ability to offer a real-time hedge to the mean. In particular, for each baseline algorithm the maximum RMSE is notably greater than 1 while in the $\mathcal{AB}$-Prod instantiation, the maximum RSME is negligibly greater than 1. In the latter case, the worst performance over all the algorithms is 1.0116. It corresponds to the upfront "cost" of using $\mathcal{AB}$-Prod$(\cdot,\mu)$, which according to Section 3 is bounded by a $2\log 2$ cost to the Benchmark $\mathcal{B}$. Finally, note that the parameters of $\mathcal{AB}$-Prod were set in order to bias heavily the combination towards the mean, which explains that the advantages of the alternative were very partially capture by the algorithm.

5. EMPIRICAL RESULTS

In order to illustrate the empirical value of our approach, we compare the performance of $\mathcal{AB}$-Prod to this of a set of standard forecast combination algorithms from the macro-economic and machine learning literature in two series of experiments. The first aims at forecasting inflation and output using the seven-country forecast combination dataset from Stock and Watson (2001). The second aims at forecasting the growth-rate in the Euro area using data from the survey of professional forecasters. The set of algorithms considered include:

- A set of basic combination methods: the mean forecaster (denoted by μ), trimmed mean forecasters with $\alpha = 0.05$ and $\alpha = 0.10$ and the median forecaster.
- Three benchmark time-varying combination methods: the AdaHedge algorithm, which is a version of **Hedge** with adaptive learning rate, the Bates Granger time-varying method 1 (BG) introduced in Timmermann (2006) and the Recent Best forecaster, which selects the forecaster with the lowest loss in the last round.
- The ex-post optimal forecaster, which can of course only be determined ex-post but provides a useful benchmark.
- The random forecaster, which selects a single forecaster at random at each

round.

- Three instantiations of the meta-algorithm $\mathcal{AB}$-Prod are presented with $\lambda_{\mathcal{B}} = 0.999$, and the Benchmark $\mathcal{B}$ set to the mean combination approach. The λ weights reflect our desire to heavily prefer the Benchmark μ.
  - AB-Prod(AdaHedge,μ): $\mathcal{A}$ =AdaHedge
  - AB-Prod(Bates-Granger,μ): $\mathcal{A}$ =the Bates-Granger method
  - AB-Prod(Recent Best,μ): $\mathcal{A}$ =the Recent Best forecaster over the previous round

REMARK 4 *Macroeconomic data, and most particularly the SPF, often has missing values , resulting in missing losses that negatively bias otherwise well-performing forecasters. The solution we adopt in the following is to impute missing data with the mean, more precisely to lower-bound the performance of the missing forecaster performance by this of the mean.*

5.1. *Seven country forecast combination dataset*

The seven-country forecast combination dataset from Stock and Watson (2001), consists in 43 quarterly time series of macro-economic indicators available for seven different countries: Canada, France, Germany, Italy, Japan, the United Kingdom and United States. The time-series include asset prices, selected measures of real economic activity and money stock from 1959 to 1999. Each of these time-series is then used to produce independent forecasts of inflation and output by estimating an autoregressive model with one exogenous variable (ARX). These forecasts are then combined using our set of candidate algorithms with a burn-in period of 8 quarters. This experiment is then repeated independently for inflation and output for three different forecast horizons, $h = 2, 4$ and 8 quarters.

REMARK 5 *The ARX forecasts are recursively generated for each exogenous variable using the Python Statsmodels library (Seabold and Perktold, 2010). Coefficients are*

estimated according to the Akaike information criterion (AIC) over 4 *lags, with ARX forecasts generated using a Broyden-Fletcher-Goldfarb-Shanno solver and maximum likelihood estimation on samples up to time* t*. Failed forecasts due to failed maximum likelihood convergence are replaced with the preceding forecast.*

	Average RMSE	Min RMSE	Max RMSE
AdaHedge	1.006743	0.727263	1.424006
Recent Best	1.288761	0.395634	18.447350
Bates-Granger	1.026792	0.726393	1.247406
Median	0.975889	0.723440	1.102208
Trimmed Mean(alpha=0.05)	0.957247	0.719920	1.024449
Trimmed Mean(alpha=0.10)	1.663990	0.659524	3.803069
AB-Prod(AdaHedge,μ)	0.952805	0.718049	0.999840
AB-Prod(Bates-Granger,μ)	0.952807	0.718049	0.999842
AB-Prod(Recent Best,μ)	0.952835	0.718046	0.999843
Random Forecaster	1.051770	0.735312	1.353242
Ex-Post Optimal	0.798261	0.557397	0.974834

Table III: Average, Minimum and Maximum Ratio to the mean of the Mean Square Forecast Error over GDP, CPI, Horizons and Countries

Table III provides a summary of the main results (whose details are in the appendix). The experiment clearly illustrate the performance advantage and adaptive hedging capabilities of the $\mathcal{AB}$-Prod meta structure. The three $\mathcal{AB}$-Prod algorithms outperform the mean for every possible combination of indicator, country and horizon, i.e the maximal RMSE ratio is less than 1. In terms of average performance, they outperform all but the ex-post optimal forecaster, which can only be determined ex-post. Their average performance is also better that this of other time-varying combination algorithms (AdaHedge, Recent Best and Bates-Granger). Moreover, these algorithms do not systematically guarantee better performance than the mean.

A detailed analysis of the results presented in the appendix shows that AB-Prod outperforms AdaHedge, Recent Best and Bates-Granger almost systematically. This suggests that AB-Prod manages, thanks to its meta-structure, to catch faster regime switches in the data and hence alternates between the alternative algorithm when it has a comparative advantage to the mean, while rapidly falling back to the safety of the mean when the alternative algorithm is unable to exploit further available information.

5.2. *Survey of professional forecasters*

The Euro-area Survey of Professional Forecasters has been conducted by the European Central Bank at a quarterly frequency since the inception of the European Monetary Union (see Bowles et al., 2007, 2010; Garcia, 2003, for a detailed description). There are around 75 survey participants, who are experts affiliated with financial and non-financial European institutions. The average number of respondents per survey is 59. Each participant[5] receives a survey of growth expectations for 1 and 2 year rolling horizons[6], with one week to reply. Survey results are published the following month. Further, the target forecast changes depending on the specific criterion by which the GDP is measured. These experts are asked to provide point forecasts for GDP and inflation at different horizons (we focus on the 1-year rolling GDP forecast horizon). Their answers provide a time-series of forecasts that are natural inputs for a forecast combination approach.

REMARK 6 *The SPF suffers from a large number of missing values with less than 60 respondents in average. We have considered two approaches to overcome this issue: the reduction to a balanced panel, as is common in the SPF literature (see e.g. Elliott*

[5]Note that the specific expert at each institution is not necessarily the same in each survey and the data has many missing values.

[6]Note that the rolling horizons are set one and two years ahead of the latest period for which the variable in question is observed when the survey is conducted and not one or two years ahead of the survey date.

and Timmermann (2005); Aiolfi et al. (2010); Genre et al. (2013)), and imputation of the missing values through the mean of available forecasts at the specific time step. Both approaches give similar results. Due to these gaps and the frequency of mean imputations, forecaster performance is expected to be close to the mean of existing forecasts. $\mathcal{AB}$ instantiations are set to $\lambda_{\mathcal{B}} = 0.999$.

	Balanced	Imputed
AdaHedge	0.969944	0.969086
Recent Best	0.810353	0.808181
Bates-Granger	1.021676	1.010879
Median	0.995093	0.995307
Trimmed Mean(alpha=0.05)	0.997080	0.996030
Trimmed Mean(alpha=0.10)	0.985286	1.006134
AB-Prod(AdaHedge,μ)	0.995106	0.995319
AB-Prod(Bates-Granger,μ)	0.995108	0.995316
AB-Prod(RB,Median)	0.995831	0.996091
Random Forecaster	0.969570	0.963621
Ex-Post Optimal	0.862678	0.860968

Table IV: SPF Data: Imputed and Balanced

Table IV reports the performance of the forecast combination algorithms in this setting. The three $\mathcal{AB}$-Prod algorithms are close to or outperform the mean combination forecast. The cost of protection is apparent in the gap between the $\mathcal{AB}$ instantiations and non-$\mathcal{AB}$ forms of the algorithms. This is most clear in the gap in performance between the $\mathcal{AB}$ and non-$\mathcal{AB}$ forms of the Recent Best algorithm. Given the lack of structure in the data resulting from gaps, changing forecasters and inconsistent forecast histories per forecaster, the problem of "learning" is expected. In each case, the $\mathcal{AB}$ instantiations learn to prefer the safety and consistency of the

mean. If a larger margin of error was acceptable, one might consider reducing the weight $\lambda_{\mathcal{B}}$.

6. DISCUSSION AND CONCLUSIONS

The paper presented a novel meta-hedging approach to adaptively combining candidate forecasts over time. More specifically, an algorithm was proposed based on several modifications to the state-of-the-art $\mathcal{AB}$-Prod algorithm from the online machine learning literature that provides an intuitive imputation strategy over an augmented pool and explicit protection against the forecast combination puzzle. In the later, the proposed algorithm actively and adaptively "hedges" performance to the Benchmark, while providing dual distribution-free theoretical regret guarantees that the performance will never be worse by a fixed constant against the benchmark with additional dual guarantees against a pool of forecasters augmented by the man and any other forecast combination algorithm. In addition to providing outstanding performance, the proposed methods provide a simple, consistent and theoretically guaranteed procedure for hedging against the so-called Forecast Combination Puzzle, while also giving access to state-of-the-art tools for combining forecasts.

7. ACKNOWLEDGMENTS

We gratefully acknowledge the support of H2020-FET proactive project DOLFINS and the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. We also thank Gilles Stoltz for his useful comments.

REFERENCES

Aiolfi, M., Capistrán, C., and Timmermann, A. G. (2010). Forecast combinations. *CREATES research paper*, (2010-21).

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Or*, pages 451–468.

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., and Rautanen, T. (2007). The ecb survey of professional forecasters (spf)-a review after eight years' experience. *ECB Occasional Paper*, (59).

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., and Rautanen, T. (2010). An evaluation of the growth and unemployment forecasts in the ecb survey of professional forecasters. *OECD Journal. Journal of Business Cycle Measurement and Analysis*, 2010(2):63.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. (2007). Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352.

Chen, Y.-C., Rogoff, K., and Rossi, B. (2008). Can exchange rates forecast commodity prices? Technical report, National Bureau of Economic Research.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2014). The forecast combination puzzle: A simple theoretical explanation.

Clark, T. E. and McCracken, M. W. (2009). Combining forecasts from nested models*. *Oxford Bulletin of Economics and Statistics*, 71(3):303–329.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

Clements, M. and Hendry, D. (1998). *Forecasting economic time series*. Cambridge University Press.

Clements, M. and Hendry, D. (1999). Forecasting non-stationary economic time series: the zeuthen lectures on economic forecasting.

Clements, M., Hendry, D. F., et al. (2002). *Pooling of forecasts*. Nuffield College.

Clements, M. P. and Hendry, D. F. (2006). Forecasting with breaks. *Handbook of economic forecasting*, 1:605–657.

de Rooij, S., van Erven, T., Grünwald, P. D., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. *Accepted to the Journal of Machine Learning Research*.

Diebold, F. X. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, 5(4):589–592.

Diebold, F. X. and Lopez, J. A. (1996). Forecast evaluation and combination.

Diebold, F. X. and Pauly, P. (1987). Structural change and the combination of forecasts. *Journal*

of Forecasting, 6(1):21–40.

Diebold, F. X. and Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, 6(4):503–508.

Elliott, G., Gargano, A., and Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2):357–373.

Elliott, G., Gargano, A., and Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54:86–110.

Elliott, G. and Timmermann, A. (2005). Optimal forecast combination under regime switching. *International Economic Review*, 46(4):1081–1102.

Even-Dar, E., Kearns, M., Mansour, Y., and Wortman, J. (2008). Regret to the best vs. regret to the average. *Machine Learning*, 72(1-2):21–37.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139.

Garcia, J. A. (2003). An introduction to the ecb's survey of professional forecasters. *ECB Occasional Paper*, (8).

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

Granger, C. W. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204.

Granziera, E., Luu, C., St-Amant, P., et al. (2013). The accuracy of short-term forecast combinations. *Bank of Canada Review*, 2013(Summer):13–21.

Guidolin, M. and Timmermann, A. (2009). Forecasts of us short-term interest rates: A flexible forecast combination approach. *Journal of Econometrics*, 150(2):297–311.

Hendry, D. F. and Clements, M. P. (2004). Pooling of forecasts. *The Econometrics Journal*, 7(1):1–31.

Hoogerheide, L., Kleijn, R., Ravazzolo, F., Van DijK, H. K., and Verbeek, M. (2010). Forecast accuracy and economic gains from bayesian model averaging using time-varying weights. *Journal of Forecasting*, 29(1-2):251–269.

Hsiao, C. and Wan, S. K. (2014). Is there an optimal forecast combination? *Journal of Econometrics*, 178:294–309.

Huang, H. and Lee, T.-H. (2010). To combine forecasts or to combine information? *Econometric Reviews*, 29(5-6):534–570.

Kapetanios, G., Labhard, V., and Price, S. (2008). Forecast combination and the bank of england's suite of statistical forecasting models. *Economic Modelling*, 25(4):772–792.

LeSage, J. P. and Magura, M. (1992). A mixture-model approach to combining forecasts. *Journal*

of Business & Economic Statistics, 10(4):445–452.

Littlestone, N. and Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, 108:212–261.

Newbold, P. and Harvey, D. I. (2002). Forecast combination and encompassing. *A companion to economic forecasting*, pages 268–283.

Novales, A. and de Fruto, R. F. (1997). Forecasting with periodic models a comparison with time invariant coefficient models. *International Journal of Forecasting*, 13(3):393–405.

Patton, A. J. and Sheppard, K. (2009). Optimal combinations of realised volatility estimators. *International Journal of Forecasting*, 25(2):218–238.

Pesaran, M. H. and Timmermann, A. (2005). Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics*, 129(1):183–217.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies*, 23(2):821–862.

Sancetta, A. (2010). Recursive forecast combination for dependent heterogeneous data. *Econometric Theory*, 26(02):598–631.

Sánchez, I. (2008). Adaptive combination of forecasts with application to wind energy. *International Journal of Forecasting*, 24(4):679–693.

Sani, A. (2015). *Machine Learning for Decision Making*. PhD thesis, Université de Lille 1.

Sani, A., Neu, G., and Lazaric, A. (2014). Exploiting easy data in online optimization. In *Advances in Neural Information Processing Systems*, pages 810–818.

Seabold, J. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*.

Sessions, D. N. and Chatterjee, S. (1989). The combining of forecasts using recursive techniques with non-stationary weights. *Journal of Forecasting*, 8(3):239–251.

Smith, J. and Wallis, K. F. (2005). Combining point forecasts: The simple average rules, ok? *manuscript, University of Warwick*.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle*. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Stock, J. H. and Watson, M. W. (2001). Forecasting output and inflation: the role of asset prices. Technical report, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, 1:135–196.

Uematsu, Y. and Tanaka, S. (2015). Macroeconomic forecasting and variable selection with a very

large number of predictors: A penalized regression approach. *arXiv preprint arXiv:1508.04217*.

Vovk, V. (1990). Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory (COLT)*, pages 371–386.

Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20(01):176–222.

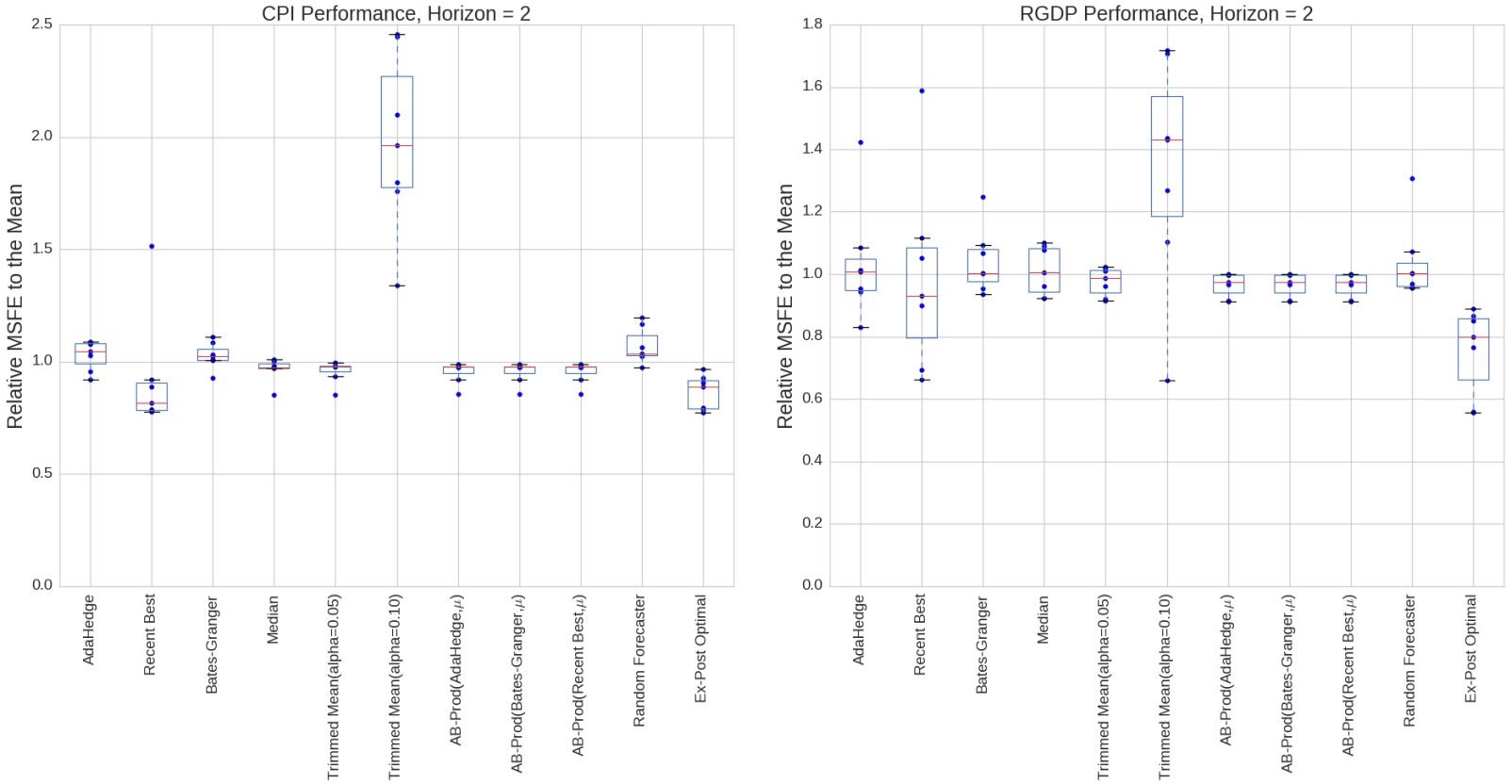

Figure 7: Relative MSFE, Horizon = 2

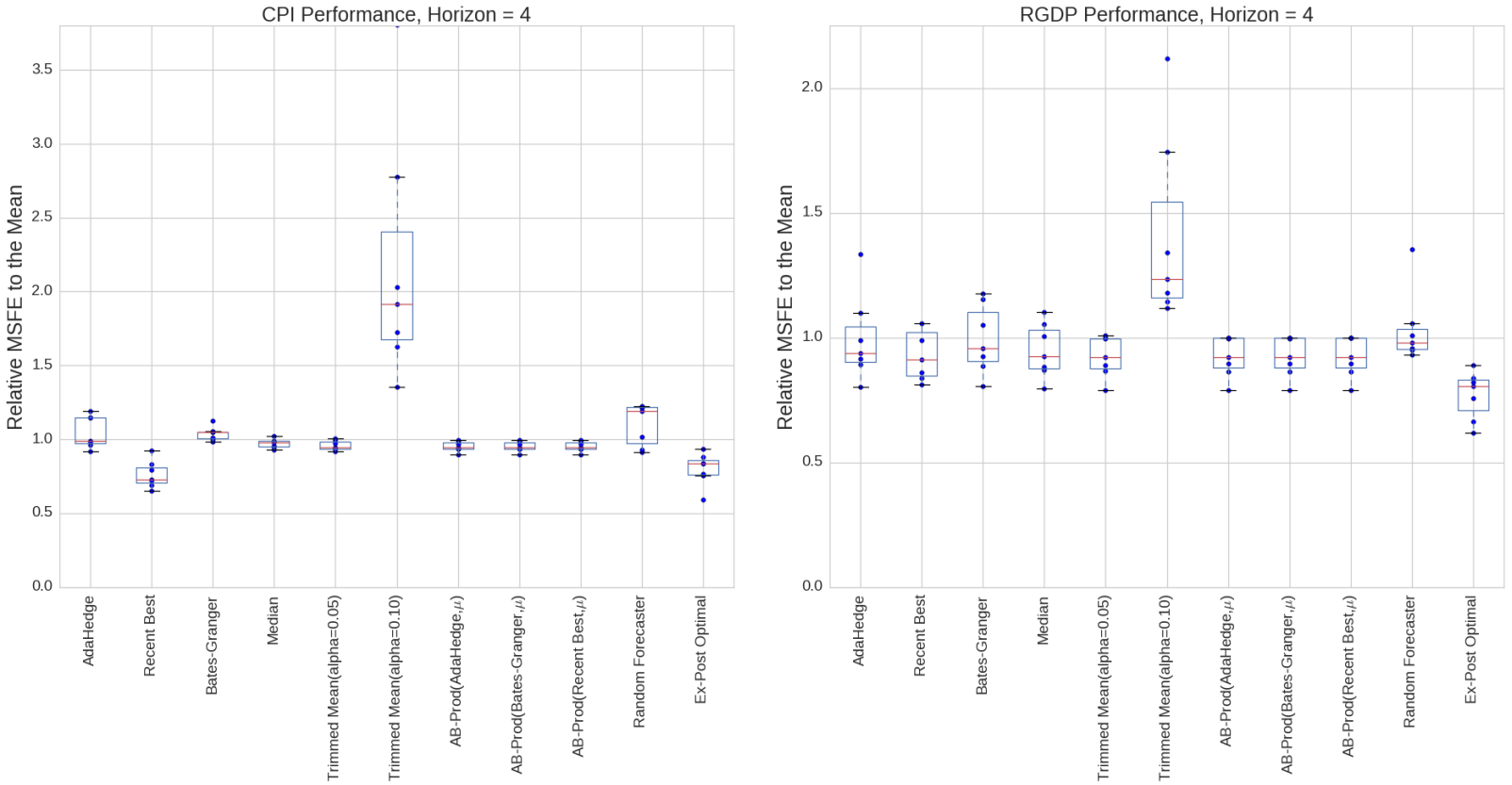

Figure 8: Relative MSFE, Horizon = 4

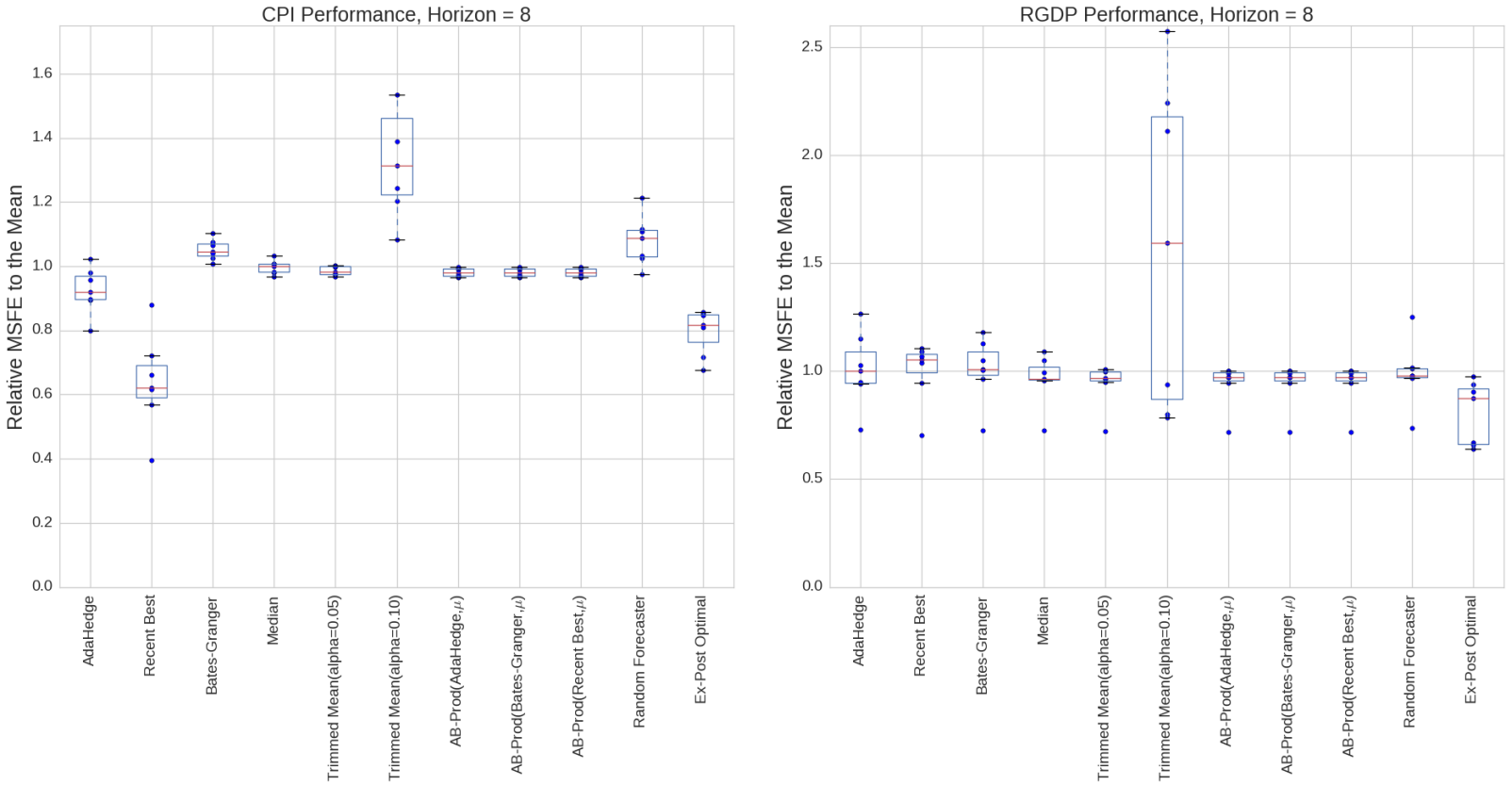

Figure 9: Relative MSFE, Horizon = 8

	China	France	Germany	Italy	Japan	UK	US
AdaHedge	1.027692	0.917716	0.955901	1.089603	1.078820	1.043915	1.079786
Recent Best	0.887602	0.785163	0.813654	0.778690	0.775533	0.919000	1.515963
Bates-Granger	1.028942	0.926949	1.006539	1.109468	1.007305	1.023833	1.084799
Median	0.967980	0.850313	0.979326	1.003207	0.971819	0.974722	1.009402
Trimmed Mean(alpha=0.05)	0.977863	0.850787	0.979695	0.979157	0.935158	0.975582	0.994010
Trimmed Mean(alpha=0.10)	1.759095	2.099013	1.337272	2.456080	1.962812	1.796089	2.445601
AB-Prod(AdaHedge,μ)	0.977713	0.854275	0.975541	0.975923	0.918655	0.974841	0.986960
AB-Prod(Bates-Granger,μ)	0.977713	0.854276	0.975546	0.975925	0.918648	0.974839	0.986961
AB-Prod(Recent Best,μ)	0.977699	0.854261	0.975526	0.975892	0.918625	0.974829	0.987004
Random Forecaster	1.023256	1.165608	1.032905	1.064479	1.193638	0.974331	1.034971
Ex-Post Optimal	0.903505	0.795443	0.788056	0.885872	0.774109	0.926796	0.965566

Table V: CPI, Horizon = 2, Relative MSFE Results

	China	France	Germany	Italy	Japan	UK	US
AdaHedge	0.942731	1.008187	1.424006	1.084366	0.829337	1.013956	0.953742
Recent Best	0.898264	1.051995	1.587237	0.692809	0.663120	1.116999	0.931004
Bates-Granger	0.935480	1.002353	1.247406	1.092291	1.068171	1.002029	0.953713
Median	0.922007	1.005610	1.101678	1.089201	1.078190	0.962154	0.923201
Trimmed Mean(alpha=0.05)	0.915359	0.985940	1.018093	1.009618	1.024449	0.962236	0.919850
Trimmed Mean(alpha=0.10)	1.102148	1.435502	1.716145	1.705908	0.659524	1.430964	1.267201
AB-Prod(AdaHedge,μ)	0.912228	0.975225	0.997529	0.997485	0.999730	0.966550	0.916113
AB-Prod(Bates-Granger,μ)	0.912227	0.975224	0.997513	0.997486	0.999755	0.966549	0.916113
AB-Prod(Recent Best,μ)	0.912223	0.975229	0.997543	0.997445	0.999709	0.966560	0.916111
Random Forecaster	0.955954	1.001357	1.306822	1.072049	0.968677	1.001748	0.955778
Ex-Post Optimal	0.865022	0.850251	0.764778	0.559406	0.557397	0.887923	0.799327

Table VI: RGDP, Horizon = 2, Relative MSFE Results

	China	France	Germany	Italy	Japan	UK	US
AdaHedge	0.987378	1.148972	0.962878	1.188571	0.915973	0.981786	1.148653
Recent Best	0.690108	0.725732	0.831245	0.651087	0.923325	0.719829	0.791346
Bates-Granger	1.048280	0.997643	0.983153	1.047659	1.053514	1.008424	1.126311
Median	0.989902	0.944505	0.931294	0.956771	0.985518	0.975252	1.020092
Trimmed Mean(alpha=0.05)	0.995129	0.932900	0.946988	0.938937	0.915762	0.971114	1.005232
Trimmed Mean(alpha=0.10)	1.914956	1.725628	1.623579	2.029354	3.803069	1.351599	2.776898
AB-Prod(AdaHedge,μ)	0.990434	0.933375	0.946516	0.936971	0.894786	0.964997	0.996970
AB-Prod(Bates-Granger,μ)	0.990440	0.933360	0.946518	0.936957	0.894800	0.964999	0.996968
AB-Prod(Recent Best,μ)	0.990404	0.933333	0.946503	0.936917	0.894787	0.964970	0.996934
Random Forecaster	1.222143	0.910372	0.928015	1.211898	1.188929	1.015007	1.220682
Ex-Post Optimal	0.877543	0.835150	0.837161	0.753022	0.591105	0.762880	0.932993

Table VII: CPI, Horizon = 4, Relative MSFE Results

	China	France	Germany	Italy	Japan	UK	US
AdaHedge	0.892596	0.800332	1.099527	1.333724	0.987369	0.913697	0.938322
Recent Best	0.859455	0.809850	1.055864	18.447350	0.837273	0.987821	0.911919
Bates-Granger	0.885777	0.805800	1.174391	1.154563	1.048829	0.924879	0.955719
Median	0.868024	0.794325	1.004354	1.102208	1.053814	0.883106	0.923204
Trimmed Mean(alpha=0.05)	0.865283	0.790131	0.996797	0.995448	1.007904	0.888014	0.921476
Trimmed Mean(alpha=0.10)	1.143301	1.117660	1.744260	2.117454	1.233383	1.178624	1.339899
AB-Prod(AdaHedge,μ)	0.863956	0.789163	0.997695	0.995876	0.999673	0.894583	0.921555
AB-Prod(Bates-Granger,μ)	0.863955	0.789163	0.997703	0.995858	0.999679	0.894584	0.921557
AB-Prod(Recent Best,μ)	0.863953	0.789164	0.997691	0.997647	0.999658	0.894590	0.921553
Random Forecaster	0.955566	0.949688	1.007614	1.353242	1.057964	0.931288	0.979254
Ex-Post Optimal	0.821905	0.755434	0.887838	0.616751	0.663060	0.836903	0.806403

Table VIII: RGDP, Horizon = 4, Relative MSFE Results

	China	France	Germany	Italy	Japan	UK	US
AdaHedge	0.799612	0.895583	0.919928	0.980815	0.956147	1.021483	0.896565
Recent Best	0.395634	0.568066	0.721221	0.621968	0.615896	0.879640	0.661627
Bates-Granger	1.044667	1.025635	1.007630	1.064509	1.102054	1.039898	1.076220
Median	1.006699	0.980125	0.967675	0.982552	1.008216	0.999470	1.033027
Trimmed Mean(alpha=0.05)	1.002034	0.968371	0.983280	0.973659	0.975715	0.997962	1.000458
Trimmed Mean(alpha=0.10)	1.244203	1.202384	1.312435	1.534758	2.803616	1.083087	1.388801
AB-Prod(AdaHedge,μ)	0.990934	0.964579	0.980968	0.971414	0.966193	0.994973	0.997491
AB-Prod(Bates-Granger,μ)	0.990959	0.964592	0.980977	0.971422	0.966207	0.994975	0.997509
AB-Prod(Recent Best,μ)	0.990894	0.964546	0.980948	0.971378	0.966159	0.994959	0.997467
Random Forecaster	0.975942	1.214248	1.025708	1.107808	1.087838	1.115666	1.032259
Ex-Post Optimal	0.716703	0.816942	0.852993	0.846011	0.675670	0.857574	0.809316

Table IX: CPI, Horizon = 8, Relative MSFE Results

	China	France	Germany	Italy	Japan	UK	US
AdaHedge	0.940745	0.727263	1.151296	1.264624	1.027481	1.001673	0.950453
Recent Best	0.946763	0.704056	1.037182	1.088804	1.053447	1.106216	1.068228
Bates-Granger	0.964636	0.726393	1.179088	1.128691	1.048486	1.009904	1.003223
Median	0.961519	0.723440	0.991709	1.090088	1.050152	0.956663	0.964825
Trimmed Mean(alpha=0.05)	0.949375	0.719920	0.995852	0.999659	1.008941	0.964136	0.966100
Trimmed Mean(alpha=0.10)	0.939316	0.800743	1.591848	2.573644	0.783951	2.112149	2.243635
AB-Prod(AdaHedge,μ)	0.945688	0.718049	0.999464	0.990341	0.999840	0.972429	0.970134
AB-Prod(Bates-Granger,μ)	0.945690	0.718049	0.999467	0.990327	0.999842	0.972430	0.970139
AB-Prod(Recent Best,μ)	0.945689	0.718046	0.999453	0.990323	0.999843	0.972440	0.970145
Random Forecaster	0.967656	0.735312	0.976700	1.249795	1.010907	1.014356	0.976907
Ex-Post Optimal	0.873424	0.668203	0.974834	0.638040	0.656253	0.936141	0.903260

Table X: RGDP, Horizon = 8, Relative MSFE Results