Learning Time-Varying Forecast Combinations

Amir Sani*
Centré d’Économie de la Sorbonne
Université Paris 1 Panthéon-Sorbonne
reachme@amirsani.com

Antoine Mandel
Centré d’Économie de la Sorbonne
Université Paris 1 Panthéon-Sorbonne
antoine.mandel@univ-paris1.fr

Abstract

Combining forecasts has been demonstrated as a robust solution to noisy data, structural breaks, unstable forecasters and shifting environmental dynamics. 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 and solves the so-called “Forecast Combination Puzzle”. Theoretical performance bounds are reported along with empirical results.

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. 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 AB-Prod, introduced in Sani et al. (2014), provides a meta-structure that combines the weights of a benchmark algorithm $B$ (the mean in our context) and these of an alternative $A$ 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

*Corresponding Author

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 a 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)]. The paper proceeds as follows. In Section 2, we propose a theoretically guaranteed forecast combination approach that “hedges” performance against the mean with synthetic results. In Section 3, we illustrate the workings of these algorithms with synthetic data. In Section 4, we demonstrate the performance of our approach for the forecast of output and inflation in the framework of [Stock and Watson (2004)]. Section 6 concludes.

2 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 $w_{h|t-h}$ from the decision set $S$ defined by the $K$-dimensional simplex

$$S := \Delta_K := \{ w \in \mathbb{R}_+^K : \sum_{i=1}^K w_i = 1 \}$$

in order to form a combined forecast of the form $\hat{y}_{h|t-h} = \sum_{i=1}^K w_i \hat{y}_{i,t|t-h}$. We also follow the forecast combination literature by using the quadratic loss to measure the performance of a forecast $\hat{y}_{h|t-h}$ as, $l_{i,t} = (y_t - \hat{y}_{i,t|t-h})^2$, with the forecast combination loss from decision weights $w$ as $l_{w,t} = \sum_{i=1}^K w_i l_{i,t}$. Aggregated over time, these losses yield the mean squared forecast error (MSFE) defined for forecaster $i$ by

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

and respectively, for a forecast combination with (fixed) weights $w$,

$$\text{MSFE}_w = \frac{1}{T-h+1} \sum_{t=h+1}^T l_{w,t}.$$ 

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: $\mathcal{H}_t = \{(y_1, \ldots, y_t), (\hat{y}_{1,h+1|t-h}, \ldots, \hat{y}_{1,t|t-h}), \ldots, (\hat{y}_{K,h+1|t-h}, \ldots, \hat{y}_{K,t|t-h})\}$. A forecast combination algorithm $\mathcal{G}$ is then defined as a series of mappings $(g_t)_{t=1,\ldots,T-h}$, where $g_t : \mathcal{H}_t \rightarrow \Delta_K$. The mapping $g_t$ associates to an observation history $h_t \in \mathcal{H}_t$ a vector of weights $g_t(h_t) = (w_{1,t}, \ldots, w_{K,t})$ in the $K$-dimensional simplex $\Delta_K$ 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 $\hat{y}_{h|t-h} = \sum_{i=1}^K w_{i,t} \hat{y}_{i,t|t-h}$. 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 $u_{i,t}$ 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 $\text{Hedge}$ (see e.g. [Freund and Schapire (1997)], [Littlestone and Warmuth (1994)], [Novk...
(1990); Cesa-Bianchi and Lugosi (2006), which exponentially updates the mixture over forecasts according to the gradient of their losses. 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 $l_{G,t}$ the loss of algorithm $G$ in period $t$ and by $L_{G,t} = \sum_{r=1}^{t} l_{G,r}$ its cumulative loss up to period $t$, the regret of an algorithm $G$ with regard to another algorithm $H$ up to time $t$ is defined as $R_{G,t}(H) = L_{G,t} - L_{H,t}$. With some abuse of notation, we can also define the regret of an algorithm $G$ up to time $t$ against a forecaster $i$ as $R_{G,t}(i)$. 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”

Hedge achieves the worst-case optimal regret $O(\sqrt{T \log K})$ 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. 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 $\mu$.

Sani et al. (2014) propose a meta-structure that hedges an algorithm $A$ based on a benchmark algorithm $B$. More precisely, the algorithm $AB$-Prod combines two forecast combination algorithms $A$ (the Alternative) and $B$ (the Benchmark) into a third “hedged” algorithm $C$, where $C$ adapts a weighting over $A$ and $B$ according to the history of their performance. This approach allows any algorithm as a benchmark. Namely, one obtains a fixed constant $2 \log 2$ regret to the benchmark $B$ under any realizations of the loss sequence. It follows from [cf. Theorem 1 in Sani et al. (2014)] that setting $B = \mu$ “solves” the “Forecast Combination Puzzle” in the sense that $AB$-Prod then provides constant and distribution-free theoretical guarantees for the relative performance with respect to the mean combination $\mu$ while exploiting any superior predictive ability of an alternative algorithm $A$, which can be chosen arbitrarily by the decision-maker. Namely, one has:

**Theorem 1.** Let $A$ be any algorithm, $B$ be the mean combination $\mu$ and $D$ be an upper bound on the benchmark losses $L_{B,T}$. Then setting weight $\lambda_B \in (0,1)$, $\lambda_A = 1 - \lambda_B$, Learning rate $\eta = C \sqrt{\frac{1}{T}} < \frac{1}{2}$, where $C = \sqrt{-\log(1 - \lambda_B)}$ simultaneously guarantees $R_{AB$-Prod,$\tau}(i) \leq R_{A,T}(i) + 2C\sqrt{D}$ for any forecaster $i$, and, $R_{AB$-Prod,$t}(i) \leq R_{\mu,T}(i) + 2 \log 2$.

### 3 Empirical Results

In order to illustrate the empirical value of our approach, we compare the performance of $AB$-Prod to this of a set of standard forecast combination algorithms from the macro-economic and machine learning literature on the seven-country forecast combination benchmark dataset from Stock and Watson (2001). The set of algorithms considered include a set of basic combination methods: the mean forecaster (denoted by $\mu$), 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 $AB$-Prod are presented, $AB$-Prod(AdaHedge,$\mu$): $A$ = AdaHedge, $AB$-Prod(Bates-Granger,$\mu$): $A$ = the Bates-Granger method, $AB$-Prod(Recent Best,$\mu$): $A$ = the Recent Best forecaster over the previous round. $\lambda$ weights are set to 0.999 to heavily prefer the Benchmark $B$, which is set to the mean combination approach, $\mu$.

---

\(^2\)Also note that the regret does not characterize the absolute performance of the algorithm $G$. A negative regret is possible if $G$ outperforms the optimal candidate forecaster.
3.1 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)\(^3\). 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.

<table>
<thead>
<tr>
<th></th>
<th>Average RMSE</th>
<th>Min RMSE</th>
<th>Max RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaHedge</td>
<td>1.006743</td>
<td>0.727263</td>
<td>1.424006</td>
</tr>
<tr>
<td>Recent Best</td>
<td>1.288761</td>
<td>0.395634</td>
<td>18.447350</td>
</tr>
<tr>
<td>Bates-Granger</td>
<td>1.026792</td>
<td>0.726393</td>
<td>1.247406</td>
</tr>
<tr>
<td>Median</td>
<td>0.975889</td>
<td>0.723440</td>
<td>1.102208</td>
</tr>
<tr>
<td>Trimmed Mean(alpha=0.05)</td>
<td>0.957247</td>
<td>0.719920</td>
<td>1.024449</td>
</tr>
<tr>
<td>Trimmed Mean(alpha=0.10)</td>
<td>1.663990</td>
<td>0.659524</td>
<td>3.803069</td>
</tr>
<tr>
<td>AB-Prod(AdaHedge,(\mu))</td>
<td>0.952805</td>
<td>0.718049</td>
<td>0.999840</td>
</tr>
<tr>
<td>AB-Prod(Bates-Granger,(\mu))</td>
<td>0.952807</td>
<td>0.718049</td>
<td>0.999842</td>
</tr>
<tr>
<td>AB-Prod(Recent Best,(\mu))</td>
<td>0.952835</td>
<td>0.718046</td>
<td>0.999843</td>
</tr>
<tr>
<td>Random Forecaster</td>
<td>1.051770</td>
<td>0.735312</td>
<td>1.353242</td>
</tr>
<tr>
<td>Ex-Post Optimal</td>
<td>0.798261</td>
<td>0.557397</td>
<td>0.947834</td>
</tr>
</tbody>
</table>

Table 1: Average, Minimum and Maximum Ratio to the mean of the Mean Square Forecast Error over GDP, CPI, Horizons and Countries

The results in Table 1 clearly illustrate the performance advantage and adaptive hedging capabilities of the \(AB\)-Prod meta structure. The three \(AB\)-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.

4 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 \(AB\)-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.

\(^3\)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.
5 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


