

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

The Impact of Sampling Variability on Estimated Combinations of Distributional Forecasts

Ryan Zischke
Gael M. Martin
David T. Frazier
D.S. Poskitt

The Impact of Sampling Variability on Estimated Combinations of Distributional Forecasts*

Ryan Zischke^{†‡§}, Gael M. Martin[†], David T. Frazier[†], D. S. Poskitt[†]

June 6, 2022

Abstract

We investigate the performance and sampling variability of estimated forecast combinations, with particular attention given to the combination of forecast *distributions*. Unknown parameters in the forecast combination are optimized according to criterion functions based on proper scoring rules, which are chosen to reward the form of forecast accuracy that matters for the problem at hand, and forecast performance is measured using the out-of-sample expectation of said scoring rule. Our results provide novel insights into the behavior of estimated forecast combinations. Firstly, we show that, asymptotically, the sampling variability in the performance of standard forecast combinations is determined solely by estimation of the constituent models, with estimation of the combination weights contributing no sampling variability whatsoever, at first order. Secondly, we show that, if computationally feasible, forecast combinations produced in a single step – in which the constituent model and combination function parameters are estimated jointly – have superior predictive accuracy and lower sampling variability than standard forecast combinations – where constituent model and combination function parameters are estimated in two steps. These theoretical insights are demonstrated numerically, both in simulation settings and in an extensive empirical illustration using a time series of S&P500 returns.

Keywords: Forecast Combination, Forecast Combination Puzzle, Probabilistic Forecasting, Scoring Rules, S&P500 Forecasting, Two-Stage Estimation

JEL Classification: C18, C51, C53

*This research has been supported by Australian Research Council (ARC) Discovery Grant DP200101414. Frazier was also supported by ARC Early Career Researcher Award DE200101070; and Martin and Frazier were provided support by the ARC Centre of Excellence in Mathematics and Statistics.

[†]Department of Econometrics and Business Statistics, Monash University.

[‡]Methodology Division, Australian Bureau of Statistics. The views expressed in this paper are those of the authors alone, and do not in any way represent the Methodology Division, the Australian Bureau of Statistics, the Australian Public Service, or the Australian Government.

[§]Corresponding author

E-mail address: ryan.zischke@monash.edu.

1 Introduction

Since the publication of the seminal papers by [Stone \(1961\)](#) and [Bates and Granger \(1969\)](#), the forecasting literature has seen an explosion of interest in the production of point and/or distributional forecasts via weighted combinations of forecasts from distinct models (see [Timmermann 2006](#), [Aastveit et al. 2019](#), and [Wang et al. 2022](#) for relevant reviews). Forecast combinations (of both types) have attracted such attention in part because of their highly competitive performance out-of-sample ([Makridakis et al. 2018, 2020](#); [Thorey et al. 2018](#); [Wang et al. 2018](#); [Taylor 2020](#)). At the same time, however, there remains considerable disagreement about the relative merits of different combination approaches, including the role played by estimation error in the performance of such methods.

Two popular methods of producing *distributional* forecast combinations, for instance, are the linear opinion pool ([Stone 1961](#); [Hall and Mitchell 2007](#); [Geweke and Amisano 2011](#); [Opschoor et al. 2017](#); [Martin et al. 2022](#)) and the beta-transformed linear opinion pool ([Ranjan and Gneiting 2010](#); [Gneiting and Ranjan 2013](#); [Satopää et al. 2014](#); [Baran and Lerch 2018](#)). One popular method of producing *point* forecast combinations is by taking a weighted average of the constituent forecasts ([Bates and Granger 1969](#); [Stock and Watson 2004](#); [Timmermann 2006](#); [Smith and Wallis 2009](#); [Claeskens et al. 2016](#)). All three of these approaches are examples of *combination functions*, which are (often parameterized) functions that produce a single forecast using forecasts from constituent models, for each future time point.

Given a combination function, one is confronted with the choice of whether to estimate the weight given to each constituent model, or to fix the weights at given values, usually a vector of values that assigns equal weight to each model. Experimentation with equally weighted combinations, and those that are “optimized” according to some criterion, has revealed that choosing the weights to optimize some reward function does not necessarily lead to better out-of-sample performance in practice, as measured by that reward function ([Clemen 1989](#)); a finding known colloquially as the “forecast combination puzzle”. Empirical studies have provided support for both possible positions. The analysis of [Stock and Watson \(2004\)](#) and [Smith and Wallis \(2009\)](#) suggest that estimating forecast combination weights has little impact on the accuracy of forecast combinations, with the equally

weighted combinations marginally outperforming combinations with optimized weights. On the other hand, there is also empirical evidence to suggest that optimal weights can outperform the equally weighted combination (e.g. [Genre et al. 2013](#); [Hsiao and Wan 2014](#); [Martin et al. 2022](#)).

Investigations have, to our knowledge, only focused on the impact of combination weight estimation (e.g. [Smith and Wallis 2009](#); [Claeskens et al. 2016](#); [Elliott 2017](#); [Chan and Pauwels 2018](#)), with the impact of sampling variability due to the estimation of the constituent model parameters largely being neglected. In this paper, we explore the impact of sampling variability in forecast combinations, comprising variability in estimates of the constituent model *and* combination parameters, on the accuracy of combinations of *distributional forecasts*. Herein, we measure the accuracy of distributional forecast combinations through expected (proper) scoring rules ([Gneiting and Raftery 2007](#)), and rigorously analyze how sampling variability in the full vector of estimated parameters impacts (distributional) forecast accuracy of estimated combinations.

Our first key finding is that, in general contexts, and under weak regularity conditions, estimation of the combination weights imparts no bias or variability into the performance of standard forecast combinations, at least as far as first order asymptotic behavior is concerned. This finding lies in opposition to the analysis of [Claeskens et al. \(2016\)](#), who show that in certain stylized scenarios estimation of the weights imparts additional bias and variance into the forecasts, but fits squarely within the findings of [Smith and Wallis \(2009\)](#), who argue that the failure of the optimized combination to perform best is due to finite-sample error. This key result is used to demonstrate that the sampling variability of our chosen forecast accuracy measure depends only on the variability of the estimated constituent model parameters, and underscores the significant role played by the estimated constituent models used in forecast combinations. Since the variability in the estimated constituent models is often much larger than the sampling variability of the estimated weights, our results suggest that not incorporating the variability of constituent model estimates within the analysis of forecast combinations may lead to conclusions that have little practical relevance.

The second key contribution made in this paper is to compare and contrast the behavior of forecast combinations that have been produced in the standard manner, against forecast

combinations that are produced in a single step. Historically, forecast combinations have been produced by first producing (estimated) forecast distributions from the constituent models, and then estimating, or fixing, the corresponding combination weights. This two-step procedure allows us to view commonly applied forecast combination approaches as *two-stage* extremum estimators (Pagan 1986; Newey and McFadden 1994; Frazier and Renault 2017), whereby the constituent models are estimated in the first stage, and the combination parameters are estimated in the second stage, conditional on the first stage estimates. In contrast, if the estimated weights and constituent model parameters were to be estimated in a single step, the resulting (vector) parameter estimator could be viewed as a *one-stage* estimator, defined by a joint optimization program. Our results demonstrate that, when feasible, one-stage forecast combinations are preferable to standard two-stage forecast combinations. In particular, forecast combinations produced in a single step have higher accuracy, and lower sampling variability, than those produced in the standard two-stage manner, as measured by our chosen forecast performance measure.¹

The paper proceeds as follows. In Section 2 we introduce generic notation for the distributional forecast combinations we consider, and discuss how to assess the performance of these combinations. Section 3 uses this framework to demonstrate that the sampling variability of standard (two-stage) forecast combination methods is entirely driven by the sampling variability in the constituent models, with no contribution from estimation of the weights; furthermore, we show that one-stage combination methods have superior forecasting performance relative to the standard forecast combination. In Section 4, we illustrate all findings numerically in a simulation setting, using a distributional forecast combination based on two constituent models. Asymptotic results on forecast performance are shown to apply in a finite-sample empirical setting in Section 5, using a time series of daily S&P500 returns from January 5th, 1988 to December 31st, 2021, with the dominance of the one-stage estimator seen to be robust to an increase in volatility. It is also shown that optimizing for forecast accuracy in the tails is of benefit during a period of high volatility, no matter what measure of out-of-sample forecast accuracy is used. Section 6 concludes. Proofs for all theoretical results can be found in the supplementary appendix, and software

¹We acknowledge the extensive Bayesian literature on combining forecast distributions (e.g. Billio et al. 2013; Casarin et al. 2015; Casarin et al. 2016; Pettenuzzo and Ravazzolo 2016; Aastveit et al. 2018; Bassetti et al. 2018; Baştürk et al. 2019; Casarin et al. 2019; McAlinn and West 2019; Loaiza-Maya et al. 2021). The Bayesian treatment of this problem is, of course, fundamentally different from the frequentist approach adopted here.

for reproducing numerical results is available at <https://github.com/zisc/samp-var-comb>.

The following notation is used throughout the paper and the supplementary appendix. Given variables y_1, y_2, \dots, y_t , we denote by $y_{1:t}$ the column vector $[y_1 \ y_2 \ \dots \ y_t]'$. For a sequence a_n converging to zero, the terms $\mathcal{O}_p(a_n)$ and $o_p(a_n)$ are used to describe the convergence of a random variable relative to a_n ; see [van der Vaart \(1998\)](#) for a textbook treatment. The symbols \Rightarrow and \xrightarrow{p} denote convergence in distribution and convergence in probability, respectively, and $\text{plim}_n X_n$ denotes the probability limit as $n \rightarrow \infty$ of a random sequence X_n . A dot is used to represent “the function whose argument replaces the dot”. For example, $f = g(\cdot, y)$ defines a function $f(x) = g(x, y)$ for some fixed y .

2 Accuracy of Distributional Forecast Combinations

We first discuss the use of scoring rules to assess the accuracy of distributional forecasts (Section 2.1), then formalize the class of distributional forecast combinations (Section 2.2), and discuss practical implementation details (Section 2.3), before describing the precise manner in which we measure the forecast performance of estimated combinations (Section 2.4).

2.1 Scoring Rules

We follow [Gneiting and Raftery \(2007\)](#) and measure the accuracy of distributional forecasts using scoring rules. A *scoring rule* (or *score*) $S(F, y)$ measures the accuracy of a predictive cumulative distribution function (CDF) F , when the variable we are trying to predict, Y , achieves realization $Y = y$. All scoring rules in our analysis are taken to be positively orientated, so that $S(F, y)$ can be interpreted as the reward the forecaster achieves for quoting distribution F when the realization $Y = y$ is observed.

Further, we consider that the scoring rules S are *strictly proper*: for a random variable Y , with true distribution function $H = \Pr(Y \leq \cdot)$, a scoring rule is proper if the expected score under H , $\mathbb{E}_H[S(F, Y)] = \int S(F, y) dH$, is maximized at $F = H$; and the score is called *strictly proper* if it is also maximized nowhere else. Throughout the remainder we take $S(\cdot, \cdot)$ to represent an arbitrary scoring rule that is strictly proper, and positively-oriented. See [Gneiting and Raftery \(2007\)](#) for a thorough introduction to use of scoring rules in measuring the accuracy of forecast distributions.

Arguably the most well-known scoring rule is the logarithmic score (log score) (Hall and Mitchell 2007; Geweke and Amisano 2011), which is defined by

$$S^{\text{LS}}(F, y) = \log f(y), \quad (1)$$

where f is the probability density function (PDF) of the CDF F . Another popular scoring rule is the censored log score (Diks et al. 2011; Opschoor et al. 2017), which prioritizes accurate forecasts in a region of the support B , potentially at the expense of forecast accuracy on B^c , the region outside B . The censored log score is defined by

$$S^{\text{CS}_B}(F, y) = \begin{cases} \log f(y) & y \in B \\ \log \left(\int_{B^c} f(y) \, dy \right) & y \in B^c \end{cases}. \quad (2)$$

Both of these scoring rules feature in our numerical work in Sections 4 and 5.

2.2 Defining Distributional Forecast Combinations

In many cases, the forecaster entertains several possible models that can be used to predict a variable of interest. Rather than choosing a single model, she can combine several models to form a distributional forecast combination: a distributional forecast combination is formed from the combination of several constituent forecast distributions. Distributional forecast combinations are thus formed from two pieces: the (predictive) distributions from the constituent models, and the method by which the distributions are combined, i.e. the combination function.

Consider Y_1, \dots, Y_t generated from the probability triplet (Ω, \mathcal{A}, H) ; given $Y_{1:t-1}$, our goal is to produce a forecast for the distribution of the random variable $Y_t|Y_{1:t-1}$. Let $F_1(\cdot | Y_{1:t-1}; \gamma_1)$ denote a predictive CDF for $Y_t|Y_{1:t-1}$ that depends on unknown parameters $\gamma_1 \in \Gamma_1 \subseteq \mathbb{R}^{d_{\gamma_1}}$.² We use the summary notation $F_{1,t}^{\gamma_1} \equiv F_1(\cdot | Y_{1:t-1}; \gamma_1)$ hereafter.

While the practitioner can use $F_{1,t}^{\gamma_1}$ to produce a forecast distribution for $Y_t|Y_{1:t-1}$, in the majority of forecasting settings there exists a collection of K possible models that can be used to produce forecast distributions. We suppose that each of these forecast distributions is indexed by their constituent model parameters, $\gamma_1, \dots, \gamma_K$, with $\gamma_j \in \Gamma_j \subseteq \mathbb{R}^{d_{\gamma_j}}$, $j =$

²To apply the results of the current paper to h -step-ahead forecasts, simply replace “ $1:t-1$ ” with “ $1:t-h$ ” throughout.

1, 2, \dots, K, which gives the practitioner access to K forecast distributions $\{F_{1,t}^{\gamma_1}, \dots, F_{K,t}^{\gamma_K}\}$ to choose amongst. Rather than adopt just one of these models, we can consider combining them to produce a single forecast distribution. In this way, we can stack the unknown parameters into a single vector that contains all the parameters from the constituent models: $\gamma \in \Gamma := \Gamma_1 \times \Gamma_2 \times \dots \times \Gamma_K$.

Following [Gneiting and Ranjan \(2013\)](#), we consider a combination function Q_η indexed by the parameter vector $\eta \in \mathcal{E} \subseteq \mathbb{R}^{d_\eta}$. Collecting the unknown parameters as $\theta = [\eta' \ \gamma']' \in \Theta := \mathcal{E} \times \Gamma$, at each time-period t , the combination function Q_η produces a single one-step-ahead predictive distribution $F_{c,t}^\theta \equiv F_c(\cdot \mid Y_{1:t-1}; \eta, \gamma)$, where F_c denotes the composition of Q_η and the K predictive distributions $F_{j,t}^{\gamma_j}$, $j = 1, 2, \dots, K$ from the constituent models under consideration. As noted earlier, and with selected references, common choices for the combination function include the linear pool and the beta-transformed linear pool, with the vector η comprising the familiar set of weights on the unit simplex in the former case. At no stage do we assume that the true data generating process (DGP) is spanned by the forecast combination.

2.3 Producing Forecast Combinations

The single predictive CDF $F_{c,t}^\theta$, which is the composition of Q_η and $F_{1,t}^{\gamma_1}, \dots, F_{K,t}^{\gamma_K}$, can be interpreted as a statistical model, P_θ , indexed by an unknown parameter vector $\theta \in \Theta$, which we use to produce a predictive distribution for Y_t given a random sample Y_1, Y_2, \dots, Y_{t-1} . With this construction, we aim to choose parameter estimates $\hat{\theta}_n \equiv \hat{\theta}_n(Y_1, Y_2, \dots, Y_n)$ using the observed data so that $P_{\hat{\theta}_n}$ is, in some sense, a good approximation to the true unknown distribution of $Y_t \mid Y_{1:t-1}$.

A clear choice then is to follow [Gneiting and Raftery \(2007\)](#) (see also [Martin et al. 2022](#)), and produce $\hat{\theta}_n$ by maximizing the in-sample average scoring rule in which we wish to measure the accuracy of our predictive distributions: for $\mathcal{S}_n(\theta) = \frac{1}{n} \sum_{t=1}^n S(F_{c,t}^\theta, Y_t)$, the unknown parameters can be estimated as

$$\hat{\theta}_n := \operatorname{argmax}_{\theta \in \Theta} \mathcal{S}_n(\theta). \tag{3}$$

Once $\hat{\theta}_n$ has been obtained, the forecast combination function $F_{c,n}^{\hat{\theta}_n}$ can be used to produce a forecast distribution for the random variable Y_{n+1} . Such an approach has been referred

to as an “optimal prediction approach” by [Gneiting and Raftery \(2007\)](#) since maximizing the in-sample criterion we wish to evaluate should, *a priori*, produce forecast distributions that attain high accuracy in that chosen score.

In the case of a predictive combination, however, the dimensionality of θ is often high. Hence, and in the spirit of the general forecast combination literature (e.g. [Hall and Mitchell 2007](#); [Geweke and Amisano 2011](#); [Gneiting and Ranjan 2013](#)), the most common approach to producing forecast combinations is to proceed in two steps (e.g. [Clemen 1989](#); [Stock and Watson 2004](#); [Geweke and Amisano 2011](#); [Genre et al. 2013](#); [Makridakis et al. 2020](#); [Martin et al. 2022](#)); first, estimate the parameters of the constituent models, $\gamma_1, \dots, \gamma_K$, then condition on these estimates to estimate the combination parameters, η .

More particularly, given our sample $Y_{1:n}$, we consider that the two-stage forecast distribution is produced using an estimator for θ that is constructed in two steps. First, each γ_j , $j = 1, 2, \dots, K$ is estimated by maximizing the in-sample average scoring rule: $\tilde{\gamma}_{j,n} := \operatorname{argmax}_{\gamma_j \in \Gamma_j} \frac{1}{n} \sum_{t=1}^n S(F_{j,t}^{\gamma_j}, Y_t)$, and we collect $\tilde{\gamma}_{j,n}$, $j = 1, \dots, K$, as $\tilde{\gamma}_n := [\tilde{\gamma}'_{1,n} \ \cdots \ \tilde{\gamma}'_{K,n}]'$. In the second stage, the combination parameters η are estimated via

$$\tilde{\eta}_n := \operatorname{argmax}_{\eta \in \mathcal{E}} \mathcal{S}_n(\eta, \tilde{\gamma}_n), \quad \mathcal{S}_n(\theta) \equiv \mathcal{S}_n(\eta, \gamma) := \frac{1}{n} \sum_{t=1}^n S(F_{c,t}^\theta, Y_t), \quad (4)$$

and the estimated parameters are stacked as

$$\tilde{\theta}_n := [\tilde{\eta}'_n \ \tilde{\gamma}'_n]'. \quad (5)$$

Remark 1. For treatments of two-stage estimation, see e.g. [Pagan \(1986\)](#); [Newey and McFadden \(1994\)](#); [Frazier and Renault \(2017\)](#). To paraphrase [Pagan \(1986\)](#), in the context of predictive combinations we can view the parameters that underlie the constituent predictive models as nuisance parameters and note that “estimation would generally be easy if the nuisance parameters were known.” In this setting, our chosen strategy for dealing with these nuisance parameters is to replace each of them by some value that is estimated from the data. As we shall see later, this seemingly innocuous practice increases the sampling variability of the forecast performance of our estimated model $P_{\tilde{\theta}_n}$.

We will refer to forecast distributions produced through a single estimation step as one-stage forecast combinations in order to distinguish them from standard (two-stage) forecast

combinations.

2.4 Assessing Accuracy of Forecast Combinations

In the case of single models, e.g., $F_{j,t}^{\gamma_j}$, accuracy of forecast combinations can be measured through the limiting expected average score

$$\lim_{n \rightarrow \infty} \mathbb{E}_H \left[\frac{1}{n} \sum_{t=1}^n S(F_{j,t}^{\gamma_j}, Y_t) \right]. \quad (6)$$

In the case of forecast combinations, accuracy of forecast distributions can be measured via

$$\mathcal{S}_0(\theta) := \lim_{n \rightarrow \infty} \mathbb{E}_H [\mathcal{S}_n(\theta)], \quad \mathcal{S}_0(\theta) \equiv \mathcal{S}_0(\eta, \gamma), \quad (7)$$

with $\mathcal{S}_n(\theta)$ as defined in (4), and where we remind the reader that \mathbb{E}_H denotes expectation under the true, unknown, distribution H of Y_1, \dots, Y_n, \dots . For the remainder of the paper we drop the subscript for brevity, and note that all expectations and variances are to be taken with respect to the true distribution H .

Rather than attempting to draw conclusions using a conventional in-sample versus out-of-sample split, we instead consider the performance of forecast combinations through the theoretically important construct of $\mathcal{S}_0(\theta)$.³ We choose this approach for measuring forecast performance for two main reasons: firstly, it obviates the need to choose the in-sample/out-of-sample evaluation mechanism, which has important ramifications for the performance of forecast evaluation and which, depending on the sample splitting, may ultimately obscure the sampling variability of the forecast combinations; secondly, this approach will allow us to easily link the sampling variability of different forecast combination schemes to their performance.

3 Implications for Forecast Performance

In this section, we show that one-stage combinations have superior forecast performance over standard two-stage combinations, and that this performance measure also has less sampling variability in the one-stage case. For the sake of brevity, and to keep technical details to a minimum, we only present the key results. Regularity conditions, technical dis-

³While, strictly speaking, \mathcal{S}_0 is infeasible to construct in practice, it can be estimated consistently (as $n \rightarrow \infty$) using a sequence of out-of-sample forecast evaluations.

cussion of said conditions, and proofs of all stated results are collected in the supplementary appendix.

As discussed in Section 2.4, we measure forecast performance via $\mathcal{S}_0(\theta)$. Therefore, the forecasting performance associated with the one- and two-stage forecast combinations can be assessed via $\mathcal{S}_0(\hat{\theta}_n)$ and $\mathcal{S}_0(\tilde{\theta}_n)$. To establish the behavior of these quantities, we first note that, under standard regularity conditions, see, e.g., [Newey and McFadden \(1994\)](#), as $n \rightarrow \infty$, $\hat{\theta}_n \xrightarrow{p} \theta^0 = [\eta^{0'} \ \gamma^{0'}]'$, and $\tilde{\theta}_n \xrightarrow{p} \theta^* = [\eta^{*'} \ \gamma^{*'}]'$, with $\gamma^* = [\gamma_1^{*'} \ \dots \ \gamma_K^{*'}]'$ where each γ_j^* maximizes the corresponding limiting expected average score in (6), η^* maximizes $\mathcal{S}_0(\eta, \gamma^*)$, and θ^0 maximizes $\mathcal{S}_0(\theta)$.

As a measure of forecast performance, the out-of-sample expected average score is in the same spirit as the performance measures considered by [West \(1996\)](#), [Hansen \(2005\)](#) and [Giacomini and White \(2006\)](#). However, unlike the above references, we are not concerned with testing the accuracy of different forecasting methods. Our main goal is to understand and document the impact of sampling variability on different forecast combination methods, and we are not interested in ascertaining which forecasting combination method provides superior forecasting performance; at least not within a conventional, i.e., formal, hypothesis testing framework.

We can summarize the first-order implications for the forecast performance of the one- and two-stage approaches as follows.

Theorem 1. *If Assumptions 1-3 in Appendix A.1 are satisfied, then the following results hold.*

1. *With probability converging to one, as $n \rightarrow \infty$, $\mathcal{S}_0(\hat{\theta}_n) > \mathcal{S}_0(\tilde{\theta}_n)$.*

2. *$\|\mathcal{S}_0(\tilde{\eta}_n, \tilde{\gamma}_n) - \mathcal{S}_0(\eta^*, \tilde{\gamma}_n)\| = o_p(n^{-1/2})$.*

Remark 2. Theorem 1 has implications for the practical performance of forecast combinations. Firstly, Part 1. of Theorem 1 shows that the two-stage nature of standard forecast combinations ensures that, in any strictly proper scoring rule, the one-stage combination will achieve a (weakly) higher performance than the two-stage approach, asymptotically. Secondly, Part 2. of Theorem 1 shows that, at first-order, sampling variability in the performance of two-stage combinations is unaffected by sampling variability from optimizing the combination weights, and is therefore driven entirely by sampling variability from estimating the constituent models. In particular, we see in Theorem 1, Part 2. that knowing

η^* *a priori* (see the right-hand \mathcal{S}_0 term) or estimating it with $\tilde{\eta}_n$ (see the left-hand \mathcal{S}_0 term) makes no difference (see the subtraction) to the first-order behavior of the forecast performance measure (see the $o_p(n^{-1/2})$ term on the right-hand-side).

Remark 3. Part 2. of Theorem 1 has implications for the so-called forecast combination puzzle.⁴ Namely, optimizing the combination in a two-stage fashion implies that $\mathcal{S}_0(\tilde{\eta}_n, \tilde{\gamma}_n) \xrightarrow{p} \mathcal{S}_0(\eta^*, \gamma^*) > \mathcal{S}_0(\bar{\eta}, \gamma^*)$, for all $\bar{\eta} \in \mathcal{E} - \{\eta^*\}$ (and in particular, for $\bar{\eta}$ being the vector of equal weights), where the result follows by continuity of \mathcal{S}_0 , the convergence $\tilde{\theta}_n \xrightarrow{p} \theta^*$, and since η^* maximizes $\mathcal{S}_0(\eta, \gamma^*)$. Critically, however, Theorem 1, Part 2. demonstrates that this optimization imparts no additional variability, at first-order, into the asymptotic distribution of $\mathcal{S}_0(\tilde{\eta}_n, \tilde{\gamma}_n)$. Hence, Theorem 1 lends support to the optimal linear pool over its equally weighted counterpart as it pertains to forecast performance evaluations conducted via scoring rules, and in cases where η^* is not the equally-weighted combination. However, when η^* is the equally weighted combination, Theorem 1, Part 2. implies that estimated forecast combinations will perform very similarly to the equally weighted combination, and that there will be little difference between the two forecast distributions in finite-samples; at least as measured by \mathcal{S}_0 or a consistent estimator thereof.

The following result compares, in our chosen measure of forecast performance, the sampling variability of one- and two-stage forecast combinations.

Theorem 2. *If Assumptions 1-3 in Appendix A.1 are satisfied and we have $\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\gamma \neq 0$, then for W^0 and W^* as defined in Appendix A.2:*

(One-stage) $\mathcal{S}_0(\hat{\theta}_n) = \mathcal{O}_p(n^{-1})$. More specifically, for $X \sim N(0, W^0)$,

$$n\{\mathcal{S}_0(\hat{\theta}_n) - \mathcal{S}_0(\theta^0)\} \Rightarrow -\frac{1}{2}X'[-\partial^2\mathcal{S}_0(\theta^0)/\partial\theta\partial\theta']X.$$

(Two-stage) $\mathcal{S}_0(\tilde{\theta}_n) = \mathcal{O}_p(n^{-1/2})$. More specifically, for $W_{\gamma\gamma}^*$ denoting the $\gamma\gamma$ -block of W^* ,

$$\sqrt{n}\{\mathcal{S}_0(\tilde{\theta}_n) - \mathcal{S}_0(\theta^*)\} \Rightarrow N(0, [\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\gamma]'W_{\gamma\gamma}^*[\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\gamma]).$$

Remark 4. Theorem 2 implies that not only does the one-stage approach yield *better* forecast performance (see Theorem 1, Part 1.), it also produces forecasts with *lower sampling*

⁴The literature on the forecast combination puzzle focuses almost exclusively on linear combinations of point forecasts (Martin et al. 2022, is an exception). Recall from Section 2.2 that our results pertain to a generic combination function, and that linear combinations are indeed such a function. We conjecture that our results remain valid when the scoring rule S is replaced by an appropriate reward function for point forecasts, which would imply that our results and conclusions apply in the point forecasting context.

variability than the two-stage combination, as measured by our forecasting accuracy measure.

Remark 5. Part 2. of Theorem 2 demonstrates that the out-of-sample forecast performance for the two-stage combination does not depend on the sampling variability of the parameter estimates for the combination parameters η . This is consistent with the results of Theorem 1, Part 2. in that, at first order, the sampling variability of the out-of-sample forecast performance is entirely driven by the sampling variability of the estimated constituent models.

Remark 6. Theorem 2 relies on the assumption that the one- and two-stage estimators converge to different limit optimizers, that is, $\theta^* \neq \theta^0$. If the limit optimizers coincide, and $\theta^* = \theta^0$, Theorem 2 does not apply.⁵ For this case we require more intricate structures to systematically compare the forecast performance of the one- and two-stage estimators, and we leave this challenge for future research. For an example of a two-stage estimator that coincides with its one-stage counterpart in the limit, see Pagan (1986, sec. 2).

4 Monte Carlo Analysis

4.1 DGP, Forecast Combination and Scoring Rules

Figures 1 through 4 illustrate Theorems 1 and 2 in the case of a simple example. Data Y_1, Y_2, \dots, Y_n is generated from the following left- and right-censored⁶ first-order autoregressive (AR(1)) process with conditionally Gaussian errors and with a first-order autoregressive conditional heteroscedasticity (ARCH(1)) structure:

$$Y_t = \min(\max(X_t, -5), 5), \quad X_t = 0.5X_{t-1} + V_t Z_t, \quad V_t^2 = 0.2 + 0.75V_{t-1}^2 Z_{t-1}^2, \quad Z_t \stackrel{i.i.d.}{\sim} N(0, 1).$$

The true DGP is stationary, with an unconditional mean of zero and an unconditional standard deviation of approximately 0.93.

Two constituent models are used to define the forecast combination: a Gaussian AR(1) model,

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \sigma Z_t, \quad Z_t \stackrel{i.i.d.}{\sim} N(0, 1), \tag{8}$$

⁵In view of their role as expected-score-maximizers, $\theta^* = \theta^0$ if and only if $\gamma^* = \gamma^0$.

⁶Censoring removes rare but extreme outliers that corrupt simulation results. By drawing a sample of 10^7 draws from Y_t , we found that $\Pr(Y_t \text{ censored}) \approx 0.0035$.

and a constant-mean model, with normally distributed innovations and an ARCH(1) structure for the conditional variance,

$$Y_t = \mu + V_t Z_t, \quad V_t^2 = \beta_0 + \beta_1 V_{t-1}^2 Z_{t-1}^2, \quad Z_t \stackrel{i.i.d.}{\sim} N(0, 1). \quad (9)$$

These two models have, respectively, the following one-step-ahead predictive CDFs for Y_t conditional on $Y_{1:t-1}$:

$$F_1(y_t | Y_{1:t-1}; \gamma_1) = \Phi\left(\frac{y_t - \alpha_0 - \alpha_1 Y_{t-1}}{\sigma}\right), \quad F_2(y_t | Y_{1:t-1}; \gamma_2) = \Phi\left(\frac{y_t - \mu}{\sqrt{\beta_0 + \beta_1 (Y_{t-1} - \mu)^2}}\right),$$

where $\gamma_1 = [\alpha_0 \ \alpha_1 \ \sigma^2]'$, $\gamma_2 = [\mu \ \beta_0 \ \beta_1]'$ and Φ is the CDF of the standard normal distribution. Our forecast combination is a linear pool of these two models, so the one-step-ahead predictive CDF of the combination is

$$F_c(y_t | Y_{1:t-1}; \theta) = \eta F_1(y_t | Y_{1:t-1}; \gamma_1) + (1 - \eta) F_2(y_t | Y_{1:t-1}; \gamma_2), \quad (10)$$

where $\eta \in [0, 1]$ is the weight assigned to the AR(1) forecast, $1 - \eta$ is the weight assigned to the ARCH(1) forecast, and $\theta = [\eta \ \gamma']'$. Let $F_{c,t}^\theta = F_c(\cdot | Y_{1:t-1}; \theta)$.

Since our forecast combination is supported on \mathbb{R} and our DGP is supported on $[-5, 5]$, our forecast combination is clearly misspecified. Censoring occurs rarely however, so it is pertinent to discuss the similarity of our forecast combination to X_t , the stochastic process of the DGP before censoring. Whereas X_t has a normal forecast distribution, our forecast combination is a mixture of two normal distributions, and is therefore normal if and only if $\eta = 0$ or $\eta = 1$. If $\eta = 0$, the forecast combination fails to capture the AR characteristics of X_t , and if $\eta = 1$, the forecast combination fails to capture the ARCH characteristics of X_t . Even without censoring, our forecast combination is therefore unable to exactly represent the true DGP, no matter how the parameters are estimated, nor how large is the number of observations available.

We make use of the log score and the censored log score defined in (1) and (2), respectively. The log score is a ‘‘local’’ scoring rule, returning a high value if the realized observation is in the high-density region of the distributional forecast combination, and a low value otherwise. Our chosen censored log score, on the other hand, prioritizes accurate

forecasts on the set $B = (-\infty, F^{-1}(0.2)]$ defined by the left 20% tail of the stationary distribution of the true DGP, where F^{-1} is the quantile function of this distribution. The quantile function F^{-1} is estimated using 10^7 draws from the DGP. In what follows, we abbreviate the log score as LS, and the censored log score as $\text{CS}_{<20\%}$.

4.2 Simulation Design

To illustrate Theorems 1 and 2, we require estimates of the expectation and variance of the forecast performance $\mathcal{S}_0(\hat{\vartheta}_n)$ in (7), with respect to the sampling distribution of some parameter estimator denoted generically by $\hat{\vartheta}_n$, for any given score S and sample size n . The nature of $\hat{\vartheta}_n$ is determined by the approach used to estimate the forecast combination. We also require estimates of the expected average score of the true DGP, $\mathcal{S}_{DGP} := \lim_{n \rightarrow \infty} \mathbb{E}[\frac{1}{n} \sum_{t=1}^n S(F_{0,t}, Y_t)]$, and of the limit optimizer, θ^* , where we use the notation $F_{0,t}$ to indicate the true one-step-ahead predictive based on data up to time $t - 1$.

We will detail our precise simulation design shortly. For now, we make the following observations. In general, none of the required moments are available in closed form; nor is θ^* . Our performance measure $\mathcal{S}_0(\cdot)$ is an expectation, however, as is \mathcal{S}_{DGP} , so if some standard regularity conditions hold we can evaluate these to any desired degree of accuracy as the sample mean from a sufficiently long realization drawn from the true DGP. An “error-free” estimate of θ^* is then obtained by optimizing the estimate of $\mathcal{S}_0(\cdot)$.

To characterize the sampling variation of the expectation $\mathcal{S}_0(\cdot)$ as a function of the parameters $\hat{\vartheta}_n$ of the estimated forecast combination, we need estimates of the mean and variance of $\mathcal{S}_0(\hat{\vartheta}_n)$, which we obtain by using sampling in three ways. First, to produce a single $\hat{\vartheta}_n$. Second, to estimate $\mathcal{S}_0(\cdot)$ (as described above) at the given $\hat{\vartheta}_n$. Third, to replicate this process, in order to estimate the mean and variance of the performance measure $\mathcal{S}_0(\hat{\vartheta}_n)$ by the sample mean and sample variance, respectively, over M replicated estimates, $\hat{\vartheta}_n$. We cannot estimate the mean and variance of $\mathcal{S}_0(\hat{\vartheta}_n)$ with a large enough value of M to render the sampling error negligible, because of the need to optimize to produce $\hat{\vartheta}_n$ at each iteration. We therefore quantify this error by producing confidence intervals for the mean and variance of $\mathcal{S}_0(\hat{\vartheta}_n)$ via standard asymptotics for i.i.d. draws. Below, we use a pre-subscript i on an estimator ${}_i\hat{\vartheta}_T$ to indicate that the corresponding forecast combination is optimized according to the score S^i , for some $i \in \{\text{LS}, \text{CS}_{<20\%}\}$

(i.e., for some $S^i \in \{S^{\text{LS}}, S^{\text{CS}_{<20\%}}\}$). We also use a post-superscript j on a score S^j used to measure performance. In this way, the indices i and j distinguish between the score S^i used to estimate the parameters, and the score S^j used to measure performance, which are not always the same.

In detail, we perform the following steps:

1. Draw 10^6 observations $y_{1:10^6}^{(0)}$ from the true DGP, given in the previous subsection.
2. Produce the two-stage estimates ${}_i\tilde{\theta}_{10^6}$ in (5), via the scores $i = \text{LS}, \text{CS}_{<20\%}$, and using $y_{1:10^6}$. With this number of observations being very large, we view the sample criterion in (4) as an error-free estimate of $\mathcal{S}_0(\cdot)$, and ${}_i\tilde{\theta}_{10^6}$ as thus an error-free estimate of θ^* . In Step 5 we will extract the combination function parameters ${}_i\tilde{\eta}_{10^6}$ from ${}_i\tilde{\theta}_{10^6}$ as an essentially exact representation of η^* given score i .
3. Construct the true one-step-ahead predictives $F_{0,t}, t = 1, 2, \dots, 10^6$ and compute

$$\frac{1}{10^6} \sum_{t=1}^{10^6} S^j(F_{0,t}, y_t^{(0)}) \approx \lim_{n \rightarrow \infty} \mathbb{E} \left[\frac{1}{n} \sum_{t=1}^n S^j(F_{0,t}, Y_t) \right] = \mathcal{S}_{DGP} \quad (11)$$

for each $j \in \{\text{LS}, \text{CS}_{<20\%}\}$, where S^{LS} and $S^{\text{CS}_{<20\%}}$ are the log score and censored log score, respectively, defined in the previous subsection. This quantity approximates the highest possible forecast performance \mathcal{S}_{DGP} – that attained by the true DGP – which we use to benchmark the performance of our estimators.

4. Draw 250000 observations $y_{1:250000}^{(1)}$ from the true DGP, independently of Step 1.
5. Produce the one- and two-stage estimates, ${}_i\hat{\theta}_n^{(1)}$ and ${}_i\tilde{\theta}_n^{(1)}$, for the forecast combination given in the previous subsection, using n of the observations drawn in Step 4 ($y_{1:n}^{(1)}$), according to the scores $i \in \{\text{LS}, \text{CS}_{<20\%}\}$, for sample sizes $n = 500, 501, \dots, 2000$. Also produce the two-stage estimator with the combination parameters fixed at their limiting values, ${}_i\tilde{\theta}_n^{(1)}|_{\eta^*} := [\eta^* \quad {}_i\tilde{\gamma}_n^{(1)'}]'$, for the same scores i and sample sizes n .
6. Approximate the out-of-sample forecast performance using the average $\hat{\mathcal{S}}_0(\hat{\vartheta}_n^{(1)}) = \frac{1}{100n} \sum_{t=250001-100n}^{250000} S^j(\hat{F}_{c,t}, y_t^{(1)})$ for all $\hat{\vartheta}_n^{(1)} \in \{{}_i\hat{\theta}_n^{(1)}, {}_i\tilde{\theta}_n^{(1)}, {}_i\tilde{\theta}_n^{(1)}|_{\eta^*}; i = \text{LS}, \text{CS}_{<20\%}, n = 500, 501, \dots, 2000\}$ and $j \in \{\text{LS}, \text{CS}_{<20\%}\}$, where $\hat{F}_{c,t}$ is the combination's forecast distribution for Y_t corresponding to the estimator $\hat{\vartheta}_n^{(1)}$ and past observations $y_{1:t-1}^{(1)}$. Given the large

number of observations used to compute $\hat{\mathcal{S}}_0(\hat{\vartheta}_n^{(1)})$, this estimate is viewed as an error-free representation of $\mathcal{S}_0(\hat{\vartheta}_n^{(1)})$.

7. Repeat Steps 4-6 1000 times, to obtain 1000 independent draws from the sampling distribution of the one-stage forecast performances $\mathcal{S}_0(i\hat{\theta}_n^{(1)})$, $\mathcal{S}_0(i\hat{\theta}_n^{(2)})$, \dots , $\mathcal{S}_0(i\hat{\theta}_n^{(1000)})$, the two-stage forecast performances $\mathcal{S}_0(i\tilde{\theta}_n^{(1)})$, \dots , $\mathcal{S}_0(i\tilde{\theta}_n^{(1000)})$ and the two-stage forecast performances with fixed combination function parameters $\mathcal{S}_0(i\tilde{\theta}_n^{(1)}|\eta^*)$, \dots , $\mathcal{S}_0(i\tilde{\theta}_n^{(1000)}|\eta^*)$, for each score i used to estimate the parameters, each score j used to measure performance, and each sample size n .

8. Using the draws from each generic sampling distribution, approximate the first two moments by their sample counterparts: $\overline{\hat{\mathcal{S}}_0(\hat{\vartheta}_n)} = \frac{1}{1000} \sum_{m=1}^{1000} \hat{\mathcal{S}}_0(\hat{\vartheta}_n^{(m)}) \approx \mathbb{E}[\mathcal{S}_0(\hat{\vartheta}_n)]$, and $\widehat{\text{Var}}(\hat{\mathcal{S}}_0(\hat{\vartheta}_n)) = \frac{1}{1000} \sum_{m=1}^{1000} \left(\hat{\mathcal{S}}_0(\hat{\vartheta}_n^{(m)}) - \overline{\hat{\mathcal{S}}_0(\hat{\vartheta}_n)} \right)^2 \approx \text{Var}(\mathcal{S}_0(\hat{\vartheta}_n))$, for all $\hat{\vartheta}_n \in \{i\hat{\theta}_n, i\tilde{\theta}_n, i\tilde{\theta}_n|\eta^*; i = \text{LS}, \text{CS}_{<20\%}, n = 500, 501, \dots, 2000\}$ where \mathcal{S}_0 is defined with respect to $S^j, j \in \{\text{LS}, \text{CS}_{<20\%}\}$.

9. Calculate the 95% confidence intervals for the population moments $\mathbb{E}[\mathcal{S}_0(\hat{\vartheta}_n)]$ and $\text{Var}(\mathcal{S}_0(\hat{\vartheta}_n))$ using the sample moments in Step 8 and standard asymptotics for i.i.d. draws.

10. Using the approximation for \mathcal{S}_{DGP} in (11) and the sample moments from Step 8, produce point estimates and confidence intervals for the expected divergence $\mathcal{S}_{DGP} - \mathbb{E}[\mathcal{S}_0(\hat{\vartheta}_n)]$. Note that the score j used to measure performance is identical in the definition of \mathcal{S}_{DGP} and $\mathbb{E}[\mathcal{S}_0(\hat{\vartheta}_n)]$.

4.3 Results

Figures 1 and 2 plot the expected divergence, $\mathcal{S}_{DGP} - \mathbb{E}[\mathcal{S}_0(\hat{\vartheta}_n)]$, in log score and censored log score terms, for a forecast combination parameterized by $\hat{\vartheta}_n$ and produced according to either the log score or the censored log score, and in either a one- or two-stage fashion. Approximation of this quantity, and construction of the corresponding 95% confidence intervals, occurs as in Steps 8-10 of the simulation instructions above. A lower expected divergence, and thereby a lower value on the vertical axis of any of these plots, indicates a better expected forecast performance, for a given sample size n . That is, on average, $\mathcal{S}_0(\hat{\vartheta}_n)$ is closer to the maximum possible value, \mathcal{S}_{DGP} . For the moment, we will restrict our attention to the results for the one-stage (red) and two-stage (green) estimators, returning

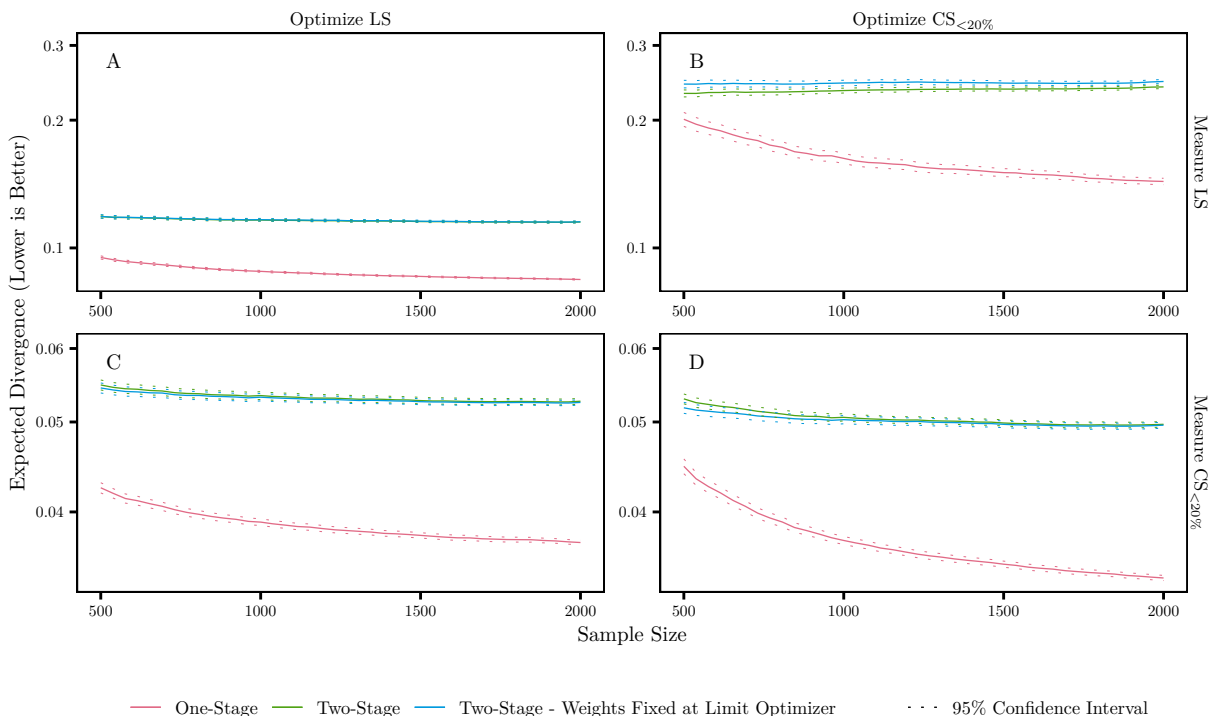


Figure 1: The expectation of the difference between the out-of-sample one-step-ahead forecast performance of a misspecified forecast combination and that of the true DGP, over a range of sample sizes. [Gneiting and Raftery \(2007\)](#) call this quantity the expected divergence. The forecast combination is optimized in a one-stage fashion (red), a two-stage fashion (green) or in a way that comprises the first stage of the two-stage combination, followed by a fixed combination at $\tilde{\eta}_n = \eta^*$ (blue). Parameters are optimized according to the log score (A and C, first column) or a censored log score that prioritizes accuracy in the lower 20% tail of the forecast distribution (B and D, second column). The divergence is measured on the vertical axes according to the log score (A and B, first row) or the censored log score (C and D, second row). The expectations and confidence intervals are constructed as per Steps 8-10 in the text, with the 95% confidence bounds appearing as small dashed lines.

to discuss the two-stage estimator with the weight, η (given in (10)), fixed at its limit optimizer (blue) at the end of the section.

In the diagonal Panels A and D of Figure 1, performance is measured and forecast combinations are produced according to the same score. In these plots, we see that the one-stage approach (red) has better expected forecast performance (a lower value on the vertical axis) than the two-stage approach (green) when measuring forecast performance and producing forecast combinations according to the same score. This is consistent with the Theorem 1, Part 1. result that the limiting forecast performance of the one-stage forecast combination is greater than that of the two-stage forecast combination when the same score is used to measure performance and produce forecast combinations.

In the off-diagonal Panels B and C, the scores used to measure performance and pro-

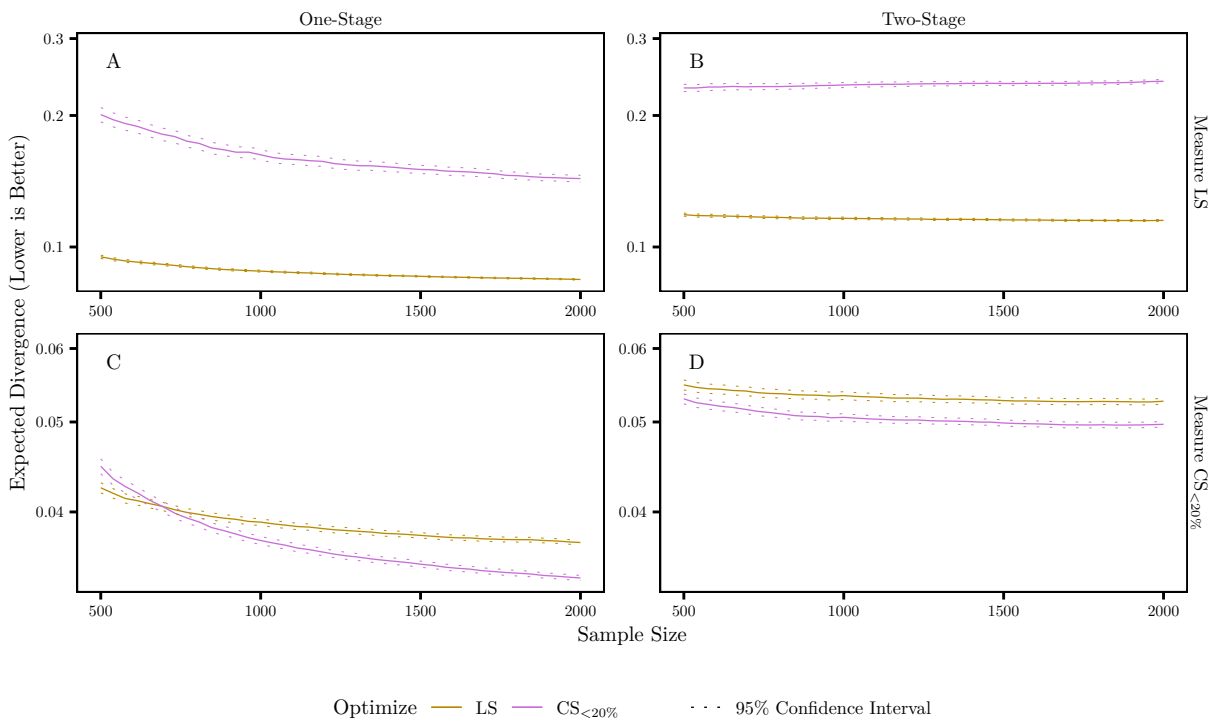


Figure 2: The expectation of the difference between the out-of-sample one-step-ahead forecast performance of a misspecified forecast combination and that of the true DGP, over a range of sample sizes. [Gneiting and Raftery \(2007\)](#) call this quantity the expected divergence. The forecast combination is optimized in a one-stage (A and C, first column) or two-stage (B and D, second column) manner according to the log score (gold) or a censored log score that prioritizes accuracy in the lower 20% tail of the forecast distribution (purple). The divergence is measured on the vertical axes according to the log score (A and B, first row) or the censored log score (C and D, second row). The expectations and confidence intervals are constructed as per Steps 8-10 in the text, with the 95% confidence bounds appearing as small dashed lines.

duce forecast combinations do not coincide and Theorem 1 does not apply, but uniform dominance of the one-stage forecast combination over the two-stage alternative nevertheless persists.

Figure 2 displays the same results as in Figure 1, but through a different lens, highlighting the advantage for either method (one-stage or two-stage) of producing forecast combinations according to the score used to measure forecast performance. Since our forecast combination is misspecified, optimizing for different scores (via either the one- or two-stage approach) leads, in general, to parameter estimates with different limit optimizers, and therefore to forecast combinations with different limiting expected average scores. Panels A and B show that producing forecast combinations according to the log score (gold) leads to a better expected forecast performance in terms of log score (a lower value on the vertical axis), across all sample sizes, relative to optimizing according to the censored log

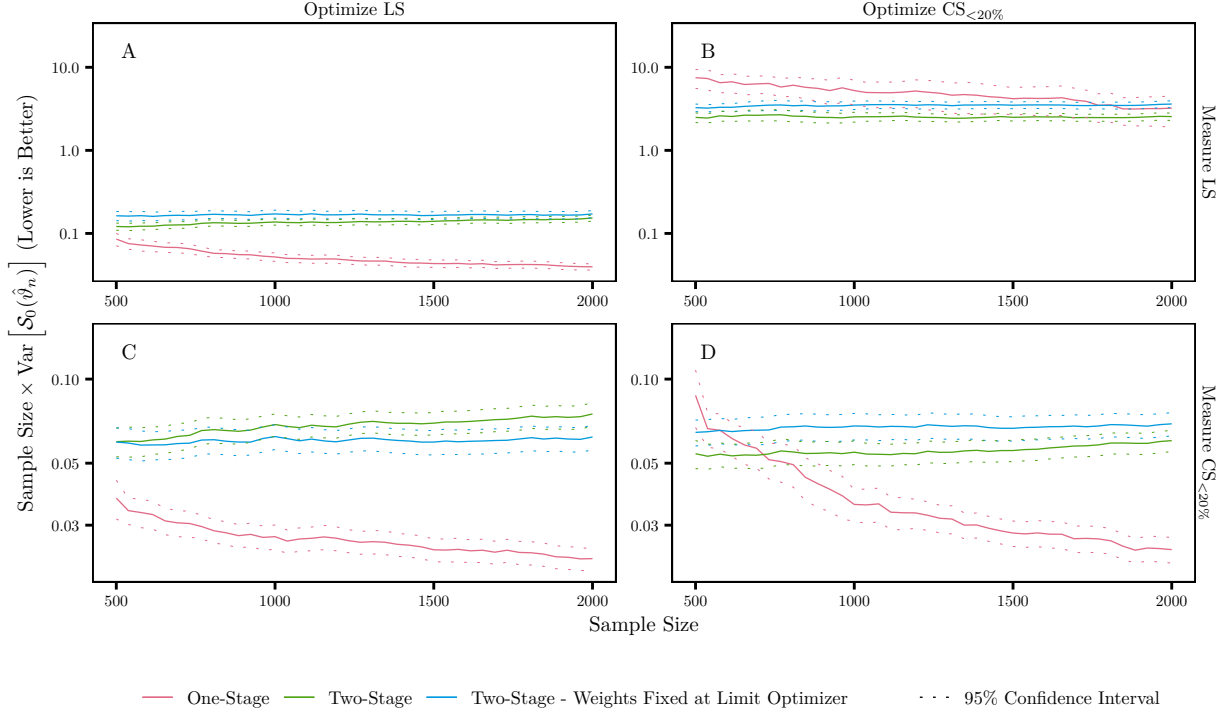


Figure 3: The sample size multiplied by the variance of the one-step-ahead forecast performance $\mathcal{S}_0(\hat{\vartheta}_n)$ of a misspecified forecast combination, over a range of sample sizes. Here, the forecast combination parameter estimator $\hat{\vartheta}_n$ optimizes the average log score (A and C, first column) or censored log score (B and D, second column) in a one-stage fashion (red), a two-stage fashion (green) or in a way that comprises the first stage of the two-stage combination, followed by a fixed combination at $\tilde{\eta}_n = \eta^*$, the optimal two-stage combination function parameter values (blue). The censored log score prioritizes accuracy in the lower 20% tail of the forecast distribution, and the variance of the forecast performance is measured on the vertical axes according to the log score (A and B, first row) or the censored log score (C and D, second row). The variances and confidence intervals are constructed as per Steps 8-10 in the text, with the 95% confidence bounds appearing as small dashed lines.

score (purple). The extent of the dominance is similar for both the one- and two-stage approaches. Likewise, Panels C (for sample sizes over 700) and D (for all sample sizes) show that producing forecast combinations according to the censored log score (purple) leads to a better expected forecast performance in terms of censored log score (a lower value on the vertical axis), relative to optimizing according to the log score (gold); and for both the one- and two-stage methods.

Figures 3 and 4 show the variance of the forecast performance of the combination, in log score and censored log score terms, for forecast combinations that optimize either the log score or the censored log score. Once again, all computational details are given in Steps 8-10 of the simulation instructions above. A lower value on the vertical axis indicates a forecast combination whose forecast performance has a lower sampling variability. The variance

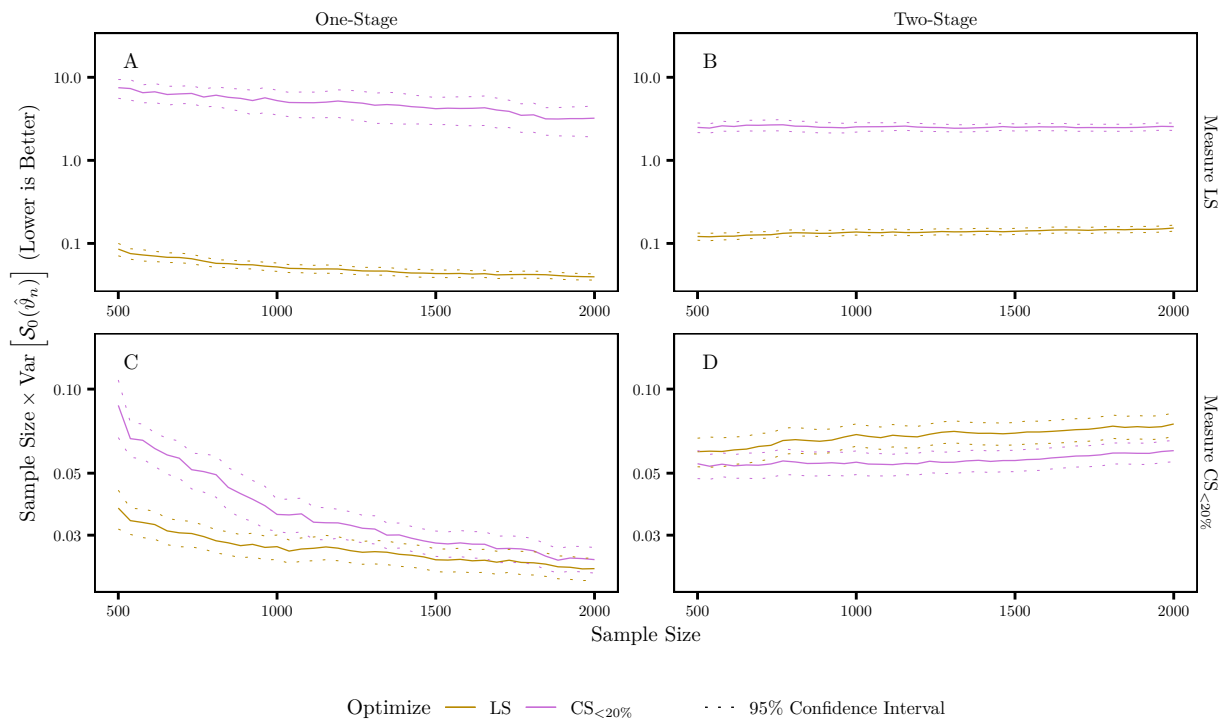


Figure 4: The sample size multiplied by the variance of the one-step-ahead forecast performance $\mathcal{S}_0(\hat{v}_n)$ of a misspecified forecast combination, over a range of sample sizes. Here, the forecast combination parameter estimator \hat{v}_n is optimized in a one-stage (A and C, first column) or two-stage (B and D, second column) fashion according to the log score (gold) or a censored log score that prioritizes accuracy in the lower 20% tail of the forecast distribution (purple). The variance of the forecast performance is measured on the vertical axes according to the log score (A and B, first row) or the censored log score (C and D, second row). The variances and confidence intervals are constructed as per Steps 8-10 in the text, with the 95% confidence bounds appearing as small dashed lines.

has been multiplied by the sample size to offset the $n^{-1/2}$ convergence of most estimators, the exception being the one-stage combination where the optimization criterion and the forecast performance measure coincide. This implies, of course, that the figures for this exceptional case (red in Panels A and D of Figure 3, and gold and purple in Panels A and C, respectively, of Figure 4) ought to decline towards $-\infty$ as the sample size increases, since the forecast performance of this forecast combination converges as n^{-1} (see Theorem 2, Part 1. and Remark 4). This is indeed visible in all cases, despite the slight plateauing effect arising due to Monte Carlo error. Once again, we defer discussion of the two-stage estimator with the weight fixed at its limit optimizer (blue) until the end of the section.

In Panels A and D of Figure 3, in which the optimization criterion and the forecast performance measure coincide, we see that the one-stage approach (red) has a forecast performance with a lower sampling variability for large sample sizes than the two-stage

approach (green). This is consistent with the Theorem 2 result that the forecast performance of the one-stage forecast combination converges faster than that of the two-stage forecast combination when the same score is used to measure performance and optimize the parameters (see Remark 4). In the off-diagonal Panels B and C, the optimization criterion and the forecast performance measure do not coincide. The uniform dominance of the one-stage forecast combination remains in evidence in Panel C. However, in Panel B, there is no such dominance on view.

Figure 4 mimics Figure 2, but with the (scaled) variance of the score being the focus. The advantage for both methods of producing forecast combinations according to the score used to measure forecast performance is on display, with Panel C illustrating the lone exception. In Panels A and B, we see that producing the forecast combination according to the log score (gold) leads to a smaller sampling variability in log-score forecast performance (a lower value on the vertical axis), across all sample sizes, relative to optimizing according to the censored log score (purple); and for both types of combinations. Panel D shows that, for the two-stage combination and for sample sizes over 1000 (at which point the two sets of confidence intervals do not overlap), optimizing according to the censored log score (purple) leads to a smaller sampling variability in censored-log-score forecast performance, relative to optimizing according to the log score (gold). For the one-stage combination in Panel C however, the sampling variability of the censored-log-score forecast performance is smaller for the combination optimizing the log score (gold) than for the combination optimizing the censored log score (purple), so long as sample sizes are small. For large sample sizes, optimizing either score leads to a similar censored-log-score sampling variability (with overlapping confidence intervals).

Finally, with reference to Panels A and D in both Figures 1 and 3 – in which the scores used to produce forecast combinations and measure performance coincide – we see that both the (expected) forecast performance and the forecast performance variance are indistinguishable (that is, have overlapping confidence intervals), or at least are very similar, for the two-stage forecast combination (green) and the two-stage forecast combination with the combination function parameter (weight) fixed at its limit optimizer (blue). This is consistent with Theorem 1, Part 2., which holds that the performance of these two forecast combinations converge to the same asymptotic distribution (given in Theorem 2, Part 2.) if

the score of the optimization criterion coincides with the score of the forecast performance measure. In this context, the limiting sampling variability of the forecast performance of the two-stage forecast combination derives entirely from sampling variability in the estimation of the constituent models, and the numerical results are simply highlighting this fact.

On the off-diagonal panels of these two figures (Panels B and C), where performance measurement and combination production are conducted according to different scores, results are mixed. Confidence intervals around the green and blue lines overlap in Panel C of Figures 1 and 3. For Panel B in Figures 1 and 3 on the other hand, confidence intervals around the green and blue lines do not overlap, indicating that both the (expected) forecast performance and the forecast performance variance are different between the two two-stage forecast combinations when the combinations are estimated according to the censored log score and performance is measured according to the log score, at least for the sample sizes considered.

The results of this simulation exercise have been obtained via repeatedly sampling from a known true DGP. In empirical settings, the true DGP is unknown, and we must rely only on the observations before us. In the next section, we illustrate how to approximate the sampling distribution of the forecast performance, without the expectation of correct specification. The results illustrated in the above simulation exercise will be shown to be robust to this added complexity.

5 Performance on S&P500 Returns

In this empirical exercise, we explore the extent to which selected theoretical results of Section 3 are reflected in estimated forecast combinations for S&P500 returns.

5.1 Data, Forecast Combination and Scoring Rules

Our dataset contains 8565 daily continuously-compounded returns y_1, \dots, y_{8565} extending from January 5th, 1988 to December 31st, 2021, and was obtained from the Global Financial Data database. The constituent forecasting models are as given in (8) and (9) in Section 4.1, which are combined via the linear pool to give the combination specified by (10). Whilst these constituent models are admittedly less sophisticated than those that would typically be used to model financial returns, and the simple linear combination of such

models less ambitious than other combination approaches adopted in the returns literature (e.g. Billio et al. 2013), these choices are sufficient for the purpose at hand. That purpose is simply to illustrate the role played by the different forms of sampling variability in the production of estimated forecast combinations; raw forecast accuracy, and the attainment of that accuracy via judicious model selection and combination, is not our goal.

Forecast combinations are produced in either a one- or two-stage fashion, and via optimization of a criterion function based on one of three scores: the log score (LS), and the censored log scores focusing on the lower 10% ($CS_{<10\%}$) and 20% ($CS_{<20\%}$) tails. The score $CS_{<20\%}$ is defined at the end of Section 4.1, as is $CS_{<10\%}$ on replacing “0.2” by “0.1”. We have selected the log score because it is ubiquitous, and chosen censored log scores with a focus on the lower tails for their relevance in financial settings where a primary goal is the accurate prediction of large or unlikely losses. To measure predictive performance, the average out-of-sample LS, $CS_{<10\%}$ and $CS_{<20\%}$ values are calculated for two distinct evaluation periods: the “overall” period and an “extreme” period. The overall period extends from January 3rd, 2017 to December 31st, 2021, and includes the first two years of the COVID-19 pandemic and the three years prior. For the extreme period we consider returns from the first six months of the pandemic, from January 2nd, 2020 to June 30th, 2020, during which the S&P500 experienced unusually extreme returns and high volatility. For both evaluation periods, forecast combinations are produced using all returns observed before the beginning of the period according to the steps outlined in Section 2.3. For forecast combinations evaluated over the overall period, the in-sample dataset has a size of $n = 7306$, and the size of the evaluation period itself is $\tau = 1259$. The corresponding sample sizes are $n = 8060$ and $\tau = 125$ when evaluating performance over the extreme period.

For any forecast combination parameterized generically by the vector θ , the expected average score $\mathcal{S}_0(\theta)$ is estimated according by $\hat{\mathcal{S}}_0^j(\theta) = \frac{1}{\tau} \sum_{t=n+1}^{n+\tau} S^j(F_{c,t}^\theta, y_t)$, where j denotes the score evaluated (one of LS, $CS_{<10\%}$, or $CS_{<20\%}$), y_t is the return on day t and $F_{c,t}^\theta$ is the corresponding predictive CDF of our forecast combination, defined immediately after (10) in Section 4.1. Unlike the simulation exercise of Section 4, we see that the expected average score must be estimated by an out-of-sample average score based on a *limited* sample size (either $\tau = 1259$ or $\tau = 125$). Our estimate of the expected average score is

therefore subject to three sources of sampling variability: 1) from in-sample estimation of the constituent models, 2) from in-sample estimation of the weight, η , and 3) from out-of-sample estimation of the expected average score via $\hat{\mathcal{S}}_0(\cdot)$. Throughout this exercise, we take the out-of-sample average score as given, and quantify and visualize sampling variability from parameter estimation only, leaving an accounting of the out-of-sample sampling variability as a topic for future research.

In detail, we perform the following steps:

1. Given returns $y_{1:n}$, produce the one- and two-stage forecast combinations parameterized by ${}_i\hat{\theta}_n$ and ${}_i\tilde{\theta}_n$, respectively, for the functional form given in (8)-(10), according to the scores $i \in \{\text{LS}, \text{CS}_{<10\%}, \text{CS}_{<20\%}\}$, and for sample size $n = 7306$ (for producing one-step-ahead forecasts for the overall period) and $n = 8060$ (when forecasting the extreme period). For both the overall period and the extreme period we use a fixed estimation window, such that the same parameter estimates (conditional on the same n observations) are used to produce all τ out-of-sample forecasts. Let ${}_i\theta^0$ and ${}_i\theta^*$ be the weak limits of the one- and two-stage forecast combinations parameterized by ${}_i\hat{\theta}_n$ and ${}_i\tilde{\theta}_n$, respectively, as $n \rightarrow \infty$.
2. For all $\hat{\vartheta}_n \in \{{}_i\hat{\theta}_n, {}_i\tilde{\theta}_n; i = \text{LS}, \text{CS}_{<10\%}, \text{CS}_{<20\%}, (n, \tau) = (7306, 1259), (8060, 125)\}$:
 - (a) Let $\vartheta = \text{plim}_n \hat{\vartheta}_n$ (either $\vartheta = {}_i\theta^0$ or $\vartheta = {}_i\theta^*$) and obtain a consistent estimate \hat{W}_n of the asymptotic covariance matrix W defined by $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \Rightarrow N(0, W)$ as $n \rightarrow \infty$.⁷
 - (b) Simulate i.i.d. draws $\hat{\vartheta}_n^{(1)}, \hat{\vartheta}_n^{(2)}, \dots, \hat{\vartheta}_n^{(20000)}$ from the distribution $N(\hat{\vartheta}_n, \hat{W}_n/n)$ (conditional on $\hat{\vartheta}_n$ and \hat{W}_n/n), and compute $\hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(1)}), \hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(2)}), \dots, \hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(20000)})$ for $j \in \{\text{LS}, \text{CS}_{<10\%}, \text{CS}_{<20\%}\}$.
 - (c) Produce a kernel density estimate based on the sample $\hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(1)}), \hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(2)}), \dots, \hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(20000)})$ and make note of the empirical estimate $\hat{\mathcal{S}}_0^j(\hat{\vartheta}_n)$ and the 95% confidence interval $\left(\hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(500)}), \hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(19500)})\right)$ for the out-of-sample score $\hat{\mathcal{S}}_0^j(\vartheta)$ of the limit optimizer ϑ .⁸

⁷See Appendix A.2 for a GMM representation of $\hat{\vartheta}_n$, to which many works on GMM can be applied to produce a consistent \hat{W}_n . We use Andrews (2002).

⁸We can show that the draws $\hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(1)}), \hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(2)}), \dots, \hat{\mathcal{S}}_0^j(\hat{\vartheta}_n^{(20000)})$ are from a consistent bootstrap distribution for $\hat{\mathcal{S}}_0^j(\hat{\vartheta}_n)$, conditional on the out-of-sample draws $y_{n+1:n+\tau}$ that define $\hat{\mathcal{S}}_0^j$, by applying Theorem 23.5 of van der Vaart (1998) with $\phi = \hat{\mathcal{S}}_0^j$, $\theta = \vartheta$, $\hat{\theta}_n = \hat{\vartheta}_n$, $T = N(0, W)$ and $\hat{\theta}_n^* \stackrel{i.i.d.}{\sim} N(\hat{\vartheta}_n, \hat{W}_n/n)$. Our confidence intervals are produced by Efron's percentile method, described in van der Vaart (1998, sec. 23.1).

5.2 Results

Table 1 contains the point estimates (labeled “Average”) and confidence intervals (labeled “95% CI”) for the average out-of-sample LS, $CS_{<10\%}$ and $CS_{<20\%}$ (columns) for the forecast combination described above optimized according to LS, $CS_{<10\%}$ or $CS_{<20\%}$ in either a one- or two-stage fashion (rows). Bolded figures in the “Average” rows indicate the highest performing method according to the out-of-sample measure in the column heading. Likewise, bolded figures in the “95% CI” rows indicate the method with the narrowest confidence interval around the out-of-sample score denoted in the column heading. Recall that the hold-out sample $y_{n+1:n+\tau}$ is taken as given when interpreting the confidence intervals, which thus reflect sampling variability in forecast combination production only. In short, the theoretical results in Section 3 indicate that the one-stage combination where the in-sample and out-of-sample scores coincide (appearing on the Table’s diagonals) ought to have both the highest average out-of-sample score and the narrowest 95% CI. We will discuss results for the overall period (Panel A) first, and results for the extreme period (Panel B) second. We supplement these tabulated results with a visualization of the overall (extreme) results in Figure 5 (6).

Panel A of the table contains results for the out-of-sample period extending from January 3rd, 2017 to December 31st, 2021. When measuring forecast performance in LS terms (first column), we find that our theoretical results are reflected in the dominance of the combination optimized by the LS in a one-stage fashion; it having both the highest average out-of-sample value (first row, bold) and the narrowest confidence interval (second row, bold). Also note that the average forecast performance of the LS-based one-stage combination is above the upper bounds of the confidence intervals for the other forecast combinations along that column, lending further support to its dominance. In the second column, where the out-of-sample score is $CS_{<10\%}$, we find that the LS-based one-stage combination once again has the highest average out-of-sample score (first row, bold) with the narrowest confidence interval (second row, bold), contrary to theory, which would have the $CS_{<10\%}$ -based one-stage combination claim those titles. Note, however, that the $CS_{<10\%}$ performance of all three one-stage forecast combinations are similar, with their average out-of-sample scores lying in each other’s confidence intervals. Results for the out-of-sample $CS_{<20\%}$ are displayed in the third column, where the one-stage forecast combination also

Panel A – Overall period: January 3rd, 2017 to December 31st, 2021						
Out-of-Sample Score						
			LS	CS _{<10%}	CS _{<20%}	
In-Sample Score	LS	One-Stage	Average	3.289	-0.003864	0.1268
			95% CI	(3.278, 3.295)	(-0.009395, -0.0006305)	(0.1203, 0.1313)
		Two-Stage	Average	3.186	-0.04076	0.07958
			95% CI	(3.172, 3.198)	(-0.04779, -0.03632)	(0.07239, 0.08438)
	CS _{<10%}	One-Stage	Average	3.206	-0.004871	0.1243
			95% CI	(3.134, 3.244)	(-0.06344, -0.0009392)	(0.07003, 0.1311)
		Two-Stage	Average	2.814	-0.009845	0.1309
			95% CI	(2.698, 2.924)	(-0.01797, -0.005612)	(0.1226, 0.1350)
	CS _{<20%}	One-Stage	Average	3.180	-0.003953	0.1263
			95% CI	(3.024, 3.241)	(-0.06838, 0.0002676)	(0.06320, 0.1360)
	Two-Stage	Average	2.971	-0.02151	0.1171	
		95% CI	(2.906, 3.035)	(-0.03008, -0.01702)	(0.1083, 0.1219)	
Panel B – Extreme period: January 2nd, 2020 to June 30th, 2020						
Out-of-Sample Score						
			LS	CS _{<10%}	CS _{<20%}	
In-Sample Score	LS	One-Stage	Average	2.085	-0.05693	0.1028
			95% CI	(1.972, 2.139)	(-0.1152, -0.02416)	(0.04506, 0.1348)
		Two-Stage	Average	1.739	-0.2501	-0.09458
			95% CI	(1.613, 1.827)	(-0.3159, -0.2043)	(-0.1605, -0.04892)
	CS _{<10%}	One-Stage	Average	2.263	-0.005679	0.1527
			95% CI	(1.898, 2.284)	(-0.2210, 0.01739)	(-0.06324, 0.1753)
		Two-Stage	Average	2.068	-0.06809	0.08888
			95% CI	(1.975, 2.121)	(-0.1438, -0.03125)	(0.01391, 0.1252)
	CS _{<20%}	One-Stage	Average	2.166	0.005156	0.1656
			95% CI	(1.198, 2.201)	(-0.4517, 0.02035)	(-0.2926, 0.1801)
	Two-Stage	Average	2.033	-0.1167	0.04320	
		95% CI	(1.922, 2.099)	(-0.1957, -0.07268)	(-0.03509, 0.08658)	

Table 1: Point estimates and confidence intervals for the average out-of-sample scores (columns) for one- and two-stage forecast combinations of S&P500 returns that optimize a variety of average in-sample scores (rows).

optimized according to the CS_{<20%} has neither the best performance (ninth row) nor the narrowest confidence interval (tenth row). Like the results for CS_{<10%}, we see that the CS_{<20%} out-of-sample scores are also similar for all one-stage combinations, with their average out-of-sample scores lying in each other's confidence intervals.

We now focus our attention on Panel B, which contains the results for the extreme period extending from January 2nd, 2020 to June 30th, 2020. Here, we find that the one-stage combination optimizing the CS_{<10%} (fifth row, bold) has the highest average out-of-sample LS (first column), which also lies outside the confidence intervals along that column for the out-of-sample LS of all other combinations. A similar observation can be made regarding the out-of-sample CS_{<10%} (column two) and CS_{<20%} (column three). Forecast performance in terms of these out-of-sample scores is dominated by the one-stage combination optimizing

Overall period: January 3rd, 2017 to December 31st, 2021

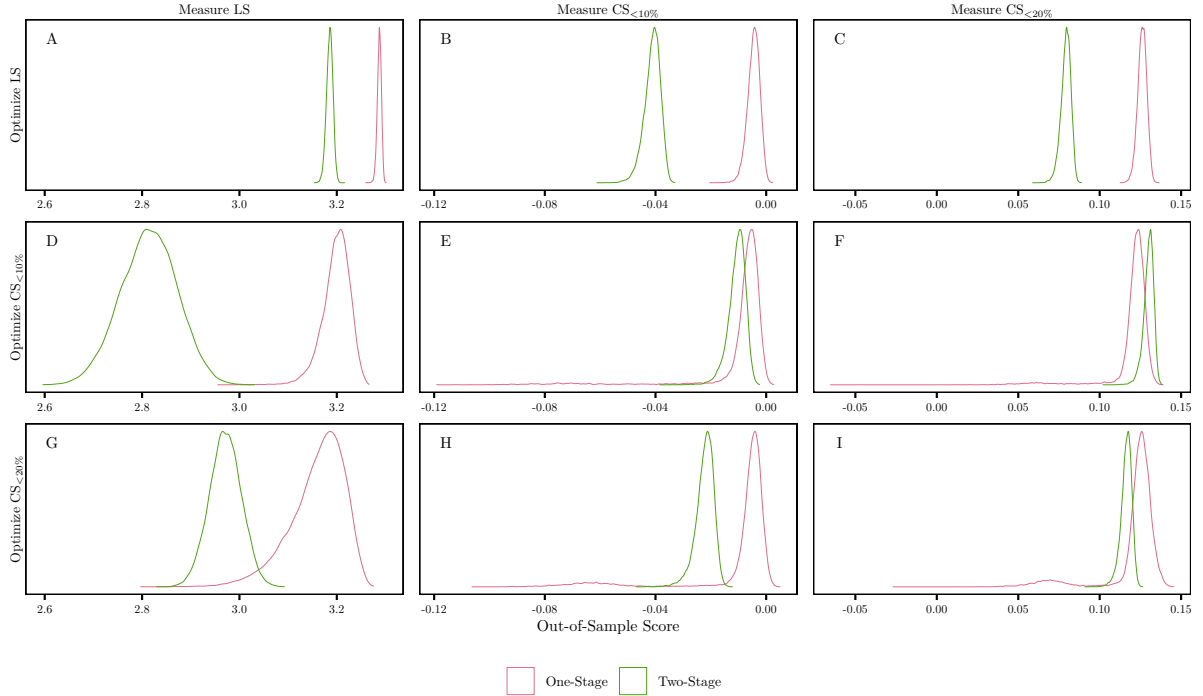


Figure 5: Sampling distributions capturing the variability of average overall-period scores (columns) due to the in-sample one-stage estimation (red) and two-stage estimation (green) of forecast combinations of S&P500 returns according to a variety of in-sample scores (rows).

the $CS_{<20\%}$ (ninth row, bold), whose average out-of-sample score lies outside the confidence intervals of all other forecast combinations for columns two and three, with the exception of the confidence intervals for the one-stage combination optimizing the $CS_{<10\%}$.

In Figures 5 to 8, we seek to shed a different light on the results in Table 1, with particular attention given to differences between results based on one- and two-stage estimation. Figure 5 displays, for the overall period, the kernel density estimates for the average out-of-sample score (columns), for one-stage and two-stage forecast combinations (colors) optimized according to the LS, $CS_{<10\%}$, and $CS_{<20\%}$ (rows). In the diagonal Panels A, E and I, where the in-sample and out-of-sample scores are the same, the sampling distributions for the out-of-sample forecast performance of the one-stage combination (red) is higher than (that is, to the right of) the sampling distribution for the two-stage combination (green), reflecting Theorem 1, Part 1. For all other panels with the exception of Panel F, the one-stage combination also has a higher forecast performance than its two-stage counterpart, despite the mismatch between the estimation and measurement criterion.

Figure 6 displays the same results as in Figure 5, but rearranged, for the purpose of

Overall period: January 3rd, 2017 to December 31st, 2021

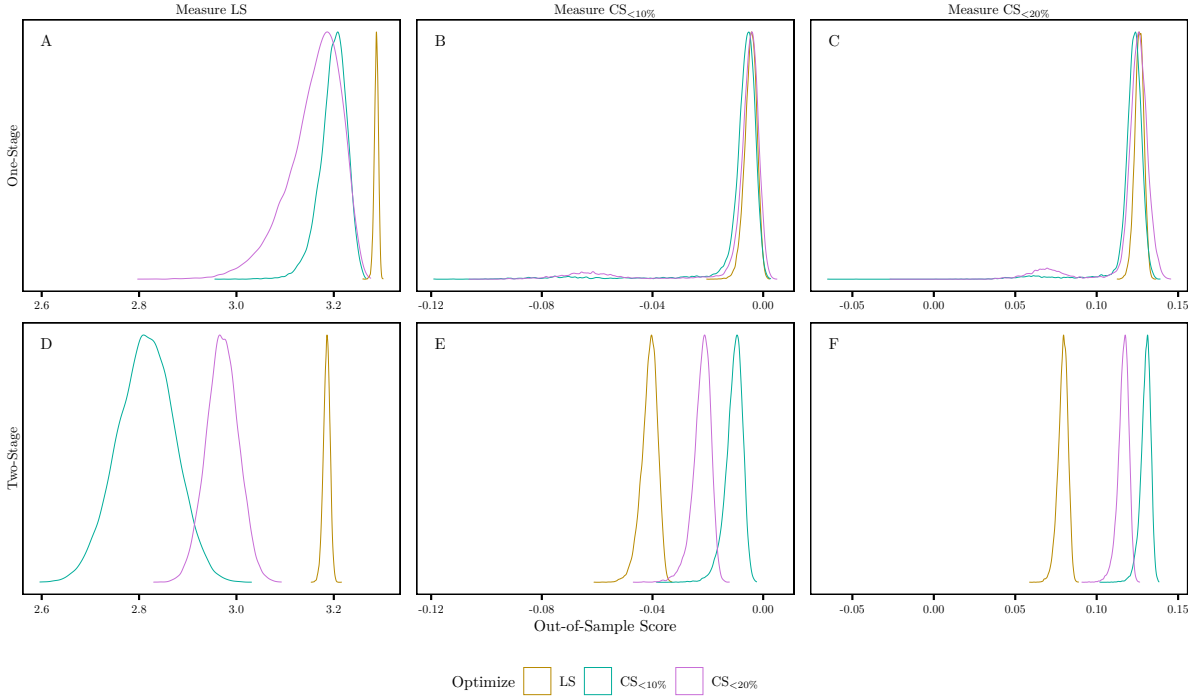


Figure 6: Sampling distributions capturing the variability of average overall-period scores (columns) due to the in-sample one-stage estimation (first row) and two-stage estimation (second row) of forecast combinations of S&P500 returns according to a variety of in-sample scores (colors).

addressing the extent to which our theory is reflected in a comparison between forecast combinations produced by different scores. The rearrangement is such that rows now delineate between one- and two-stage combination, and colors indicate the score used to produce the forecast combination. Theorem 1 is illustrated in Panel A, where we see that optimizing according to the LS (gold) leads to a one-stage combination with an out-of-sample LS being higher (to the right of), and with less sampling variability (a narrower kernel density), relative to optimizing either of the censored log scores (teal and purple). The same is observed for two-stage estimation in Panel D, only now the $CS_{<20\%}$ -based estimator outperforms its $CS_{<10\%}$ -based counterpart. On the other hand, sampling distributions for the out-of-sample scores of the one-stage forecast combinations in Panels B and C are similar in both location and spread, with substantial overlap. Looking again at the bottom row, the two-stage combination that optimizes the $CS_{<10\%}$ dominates out-of-sample forecast performance in terms of both the $CS_{<10\%}$ (Panel E) and the $CS_{<20\%}$ (Panel F), when compared to other two-stage combinations. Also note that the performances of combinations optimized according to the three different scores are more distinct for two-stage combinations

Extreme period: January 2nd, 2020 to June 30th, 2020

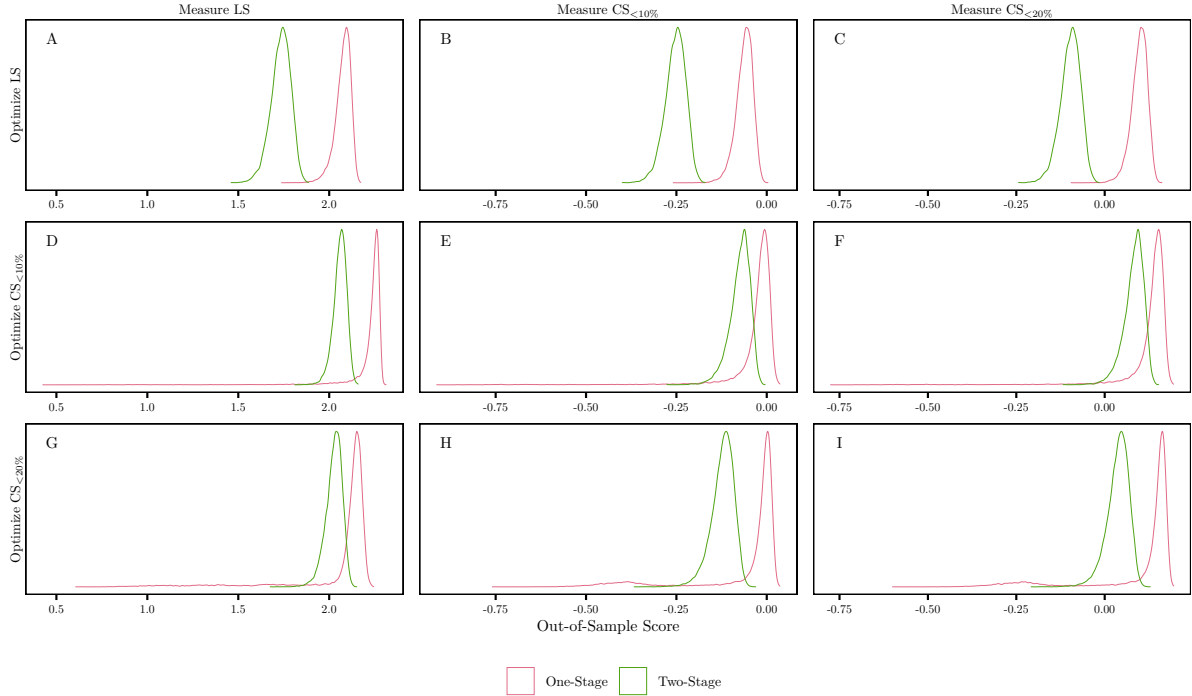


Figure 7: Sampling distributions capturing the variability of average extreme-period scores (columns) due to the in-sample one-stage estimation (red) and two-stage estimation (green) of forecast combinations of S&P500 returns according to a variety of in-sample scores (rows).

in the bottom row, than for the one-stage combinations on the top row. Further, with the exception of the small LS sampling variability for the LS-optimizing combination in Panels A and D of the first column, the sampling variability of the out-of-sample score is similar (similar kernel density widths) for all three optimized scores (colors) within each panel. One caveat to this interpretation is the bi-modality and long tail of the out-of-sample score for the one-stage combination (first row) optimizing the $CS_{<10\%}$ (teal) and $CS_{<20\%}$ (purple), especially in Panels B and C.

Remarkably, the dominance of the one-stage forecast combinations over their two-stage counterparts persists as we move from the overall period to the extreme period, and is now uniform. This is reflected in Figure 7, which shows a higher sampling distribution for the average out-of-sample scores of one-stage combinations (red) relative to their two-stage counterparts (green), regardless of how out-of-sample performance is measured (columns) or which in-sample score is optimized (rows). The asymptotic superiority of the one-stage approach implied by Theorem 1 is therefore reflected in our finite-sample results even as volatility increases and the size of the out-of-sample period declines.

Extreme period: January 2nd, 2020 to June 30th, 2020

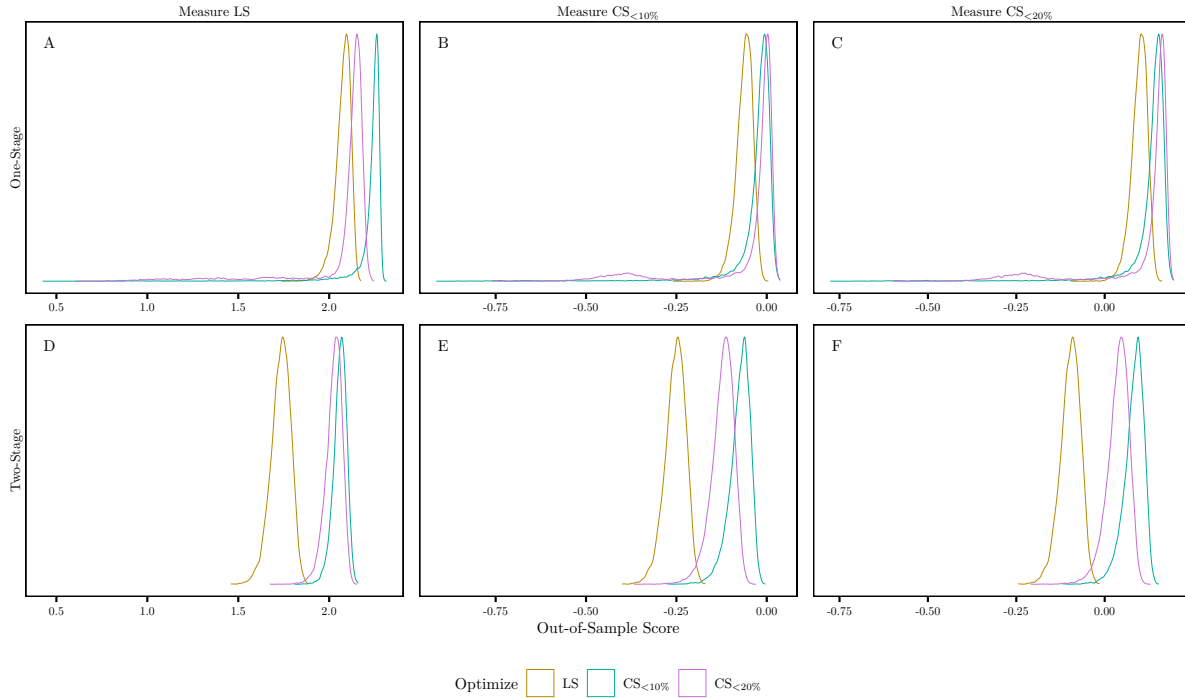


Figure 8: Sampling distributions capturing the variability of average extreme-period scores (columns) due to the in-sample one-stage estimation (first row) and two-stage estimation (second row) of forecast combinations of S&P500 returns according to a variety of in-sample scores (colors).

Finally, Figure 8 addresses, for the extreme period, the effect of optimizing for the different scores (colors). There are two key conclusions to draw from a comparison of the results for the extreme period in Figure 8 and the results for the overall period in Figure 6. First, relative to the corresponding panel in Figure 6, the sampling distributions for the out-of-sample scores for the extreme period in Figure 8 are both lower (that is, further to the left) and broader than those for the overall period, which is reflected primarily in the different scales on the x -axes of corresponding panels. Second, the LS-optimizing forecast combinations (gold) now underperforms in terms of all out-of-sample measures (columns), relative to the $CS_{<10\%}$ - and $CS_{<20\%}$ -optimizing forecast combinations (teal and purple, respectively), whether optimizing in a one-stage (first row) or two-stage (second row) fashion. In the extreme period, optimizing according to a censored log score *always* leads to a better out-of-sample performance, whichever measure of performance is used.

6 Conclusion

In this paper, we have compared the forecast performance of a variety of methods for estimating the parameters of a forecast combination. Forecast performance is measured according to an expected out-of-sample scoring rule, and we take into account the sampling variability of this measure of forecast performance that arises due to sampling variability in the estimation of the forecast combination parameters – parameters that underlie the constituent models as well as the parameters of the combination function. In this context, we analyze and compare the forecast performance of one- and two-stage forecast combinations produced according to either the score used to measure forecast performance, or some other score.

We have found via standard asymptotic arguments that for traditional two-stage forecast combinations, uncertainty in forecast performance is dominated by uncertainty in the estimation of the constituent models. Regarding the forecast combination puzzle, this lends support to optimal two-stage combinations over their equally weighted counterparts, since the limiting performance of the former is necessarily at least as high as the latter, with no added limiting sampling variability. Unfortunately, the standard practice in the forecast combinations literature is to hold the constituent models fixed when studying the forecast performance consequences of different estimation methods. Hence, this dominant source of uncertainty is often neglected in practice, including in studies of the forecast combination puzzle. Additionally, we show that producing one-stage forecast combinations according to a measure of forecast performance typically results in this measure having a higher limiting value and a lower limiting sampling variability, relative to two-stage forecast combinations. Alongside this result, we also provide formulae for producing estimates for the sampling distribution of the forecast performance of one- and two-stage forecast combinations.

A simulation exercise confirms the theoretical results on all fronts, across multiple measures of performance, and for sample sizes typically observed in practice. In particular, knowing the limiting combination weights before observing the data does not appreciably change the forecast performance of the two-stage combination, provided that the performance measure and the optimization criterion coincide. Further, one-stage forecast combinations always beats two-stage forecast combinations as long as the sample size is large,

even when assessed according to a scoring rule that differs from that used to produce the forecast combination.

In our empirical exercise, forecast combinations of S&P500 returns continue to provide strong support for the benefit of one-stage forecast combinations over two-stage forecast combinations, most notably in the period of pandemic-induced high volatility in the first half of 2020. Moreover, in this latter period, both one-stage *and* two-stage combinations optimized according to a tail-based scoring rule outperform the corresponding combinations optimized according to the log score, no matter what measure of out-of-sample performance is used.

We see several opportunities for further investigation. First, there is no doubt that this work has implications for other contexts where two-stage estimation is commonly employed, other than (distributional) forecast combinations. We expect that the advice given above on the source of forecast performance sampling variability in the two-stage cases, and the superior performance of one-stage forecast combinations, will apply equally to those contexts. Secondly, we are interested to discover the implications of these results for tests of predictive ability (e.g. [Diebold and Mariano 1995](#); [Hansen 2005](#); [Giacomini and White 2006](#)), especially within the context of the forecast combination puzzle. Thirdly, we would like to explore whether our estimates for the sampling distribution of forecast performance derived in Sections 3 and 5.1 can be improved upon by a more sophisticated bootstrap-based methodology.

Finally, we leave forecasters with the following *guidelines*:

- Where possible, consider using one-stage forecast combinations instead of the standard two-stage approach. We also advocate for producing forecast combinations by optimizing the measure of forecast performance most important for the problem at hand, and for the horizon to be forecast by the combination.
- When comparing the performance of competing forecast combinations, consider the impact that sampling variability in the parameter estimates has on the sampling variability of the forecast performance measure. In particular, it is important to consider the impact of sampling variability in *all* estimated parameters of a combination, including those that govern the constituent models. For combinations produced in two-stages, sampling variability in the estimation of the constituent model parameters can dominate overall forecast

performance; hence, ignoring this component of sampling variability may lead to misleading conclusions about the benefits, or otherwise, of optimizing the combination weights.

- When forecasting during times of high volatility, consider using a scoring rule that focuses on forecast accuracy in the tails, rather than relying only on the log score.

Acknowledgments

We would like to thank Eric Eisenstat, Rob Hyndman, Mervyn Silvapulle and Farshid Vahid-Araghi for their thoughtful comments, which have been of great benefit to the quality of our article.

Disclosure Statement

The authors report there are no competing interests to declare.

References

- Aastveit, K. A., Mitchell, J., Ravazzolo, F., and van Dijk, H. K. (2019), “The Evolution of Forecast Density Combinations in Economics,” in *Oxford Research Encyclopedia of Economics and Finance*, Great Clarendon Street, Oxford, OX2 6DP, United Kingdom: Oxford University Press.
- Aastveit, K. A., Ravazzolo, F., and van Dijk, H. K. (2018), “Combined Density Nowcasting in an Uncertain Economic Environment,” *Journal of Business & Economic Statistics*, 36(1), 131–145.
- Andrews, D. W. (2002), “Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators,” *Econometrica*, 70(1), 119–162.
- Baran, S., and Lerch, S. (2018), “Combining Predictive Distributions for the Statistical Post-Processing of Ensemble Forecasts,” *International Journal of Forecasting*, 34(3), 477–496.
- Bassetti, F., Casarin, R., and Ravazzolo, F. (2018), “Bayesian Nonparametric Calibration and Combination of Predictive Distributions,” *Journal of the American Statistical Association*, 113(522), 675–685.
- Baştürk, N., Borowska, A., Grassi, S., Hoogerheide, L., and van Dijk, H. K. (2019), “Forecast Density Combinations of Dynamic Models and Data Driven Portfolio Strategies,” *Journal of Econometrics*, 210(1), 170–186.
- Bates, J. M., and Granger, C. W. (1969), “The Combination of Forecasts,” *Journal of the Operational Research Society*, 20(4), 451–468.

- Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013), “Time-Varying Combinations of Predictive Densities Using Nonlinear Filtering,” *Journal of Econometrics*, 177(2), 213–232.
- Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. K. (2019), “Forecast Density Combinations With Dynamic Learning for Large Data Sets in Economics and Finance,” Tinbergen Institute Discussion Paper 2019-025/III.
- Casarin, R., Leisen, F., Molina, G., and ter Horst, E. (2015), “A Bayesian Beta Markov Random Field Calibration of the Term Structure of Implied Risk Neutral Densities,” *Bayesian Analysis*, 10(4), 791–819.
- Casarin, R., Mantoan, G., and Ravazzolo, F. (2016), “Bayesian Calibration of Generalized Pools of Predictive Distributions,” *Econometrics*, 4(1), 17.
- Chan, F., and Pauwels, L. L. (2018), “Some Theoretical Results on Forecast Combinations,” *International Journal of Forecasting*, 34(1), 64–74.
- Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016), “The Forecast Combination Puzzle: A Simple Theoretical Explanation,” *International Journal of Forecasting*, 32(3), 754–762.
- Clemen, R. T. (1989), “Combining Forecasts: A Review and Annotated Bibliography,” *International Journal of Forecasting*, 5(4), 559–583.
- Diebold, F. X., and Mariano, R. S. (1995), “Comparing Predictive Accuracy,” *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Diks, C., Panchenko, V., and van Dijk, D. (2011), “Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails,” *Journal of Econometrics*, 163(2), 215 – 230.
- Elliott, G. (2017), “Forecast Combination When Outcomes are Difficult to Predict,” *Empirical Economics*, 53(1), 7–20.
- Frazier, D. T., and Renault, E. (2017), “Efficient Two-Step Estimation via Targeting,” *Journal of Econometrics*, 201(2), 212–227.
- 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.
- Geweke, J., and Amisano, G. (2011), “Optimal Prediction Pools,” *Journal of Econometrics*, 164(1), 130–141.
- Giacomini, R., and White, H. (2006), “Tests of Conditional Predictive Ability,” *Econometrica*, 74(6), 1545–1578.
- Gneiting, T., and Raftery, A. E. (2007), “Strictly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association*, 102(477), 359–378.

- Gneiting, T., and Ranjan, R. (2013), “Combining Predictive Distributions,” *Electronic Journal of Statistics*, 7, 1747–1782.
- Hall, S. G., and Mitchell, J. (2007), “Combining Density Forecasts,” *International Journal of Forecasting*, 23(1), 1–13.
- Hansen, P. R. (2005), “A Test for Superior Predictive Ability,” *Journal of Business & Economic Statistics*, 23(4), 365–380.
- Hsiao, C., and Wan, S. K. (2014), “Is There an Optimal Forecast Combination?,” *Journal of Econometrics*, 178, 294–309.
- Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021), “Focused Bayesian Prediction,” *Journal of Applied Econometrics*, 36(5), 517–543.
- Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018), “The M4 Competition: Results, Findings, Conclusion and Way Forward,” *International Journal of Forecasting*, 34(4), 802–808.
- Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020), “The M4 Competition: 100,000 Time Series and 61 Forecasting Methods,” *International Journal of Forecasting*, 36(1), 54–74.
- Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., and Ramírez-Hassan, A. (2022), “Optimal Probabilistic Forecasts: When do They Work?,” *International Journal of Forecasting*, 38(1), 384–406.
- McAlinn, K., and West, M. (2019), “Dynamic Bayesian Predictive Synthesis in Time Series Forecasting,” *Journal of Econometrics*, 210(1), 155–169.
- Newey, W. K., and McFadden, D. (1994), “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, Vol. 4, Radarweg 29, Amsterdam 1043 NX, Netherlands: Elsevier, chapter 36, pp. 2111–2245.
- Opschoor, A., van Dijk, D., and van der Wel, M. (2017), “Combining Density Forecasts Using Focused Scoring Rules,” *Journal of Applied Econometrics*, 32(7), 1298–1313.
- Pagan, A. (1986), “Two Stage and Related Estimators and Their Applications,” *The Review of Economic Studies*, 53(4), 517–538.
- Pettenuzzo, D., and Ravazzolo, F. (2016), “Optimal Portfolio Choice Under Decision-Based Model Combinations,” *Journal of Applied Econometrics*, 31(7), 1312–1332.
- Ranjan, R., and Gneiting, T. (2010), “Combining Probability Forecasts,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 71–91.

- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014), “Combining Multiple Probability Predictions Using a Simple Logit Model,” *International Journal of Forecasting*, 30(2), 344–356.
- 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. (2004), “Combination Forecasts of Output Growth in a Seven-Country Data Set,” *Journal of Forecasting*, 23(6), 405–430.
- Stone, M. (1961), “The Linear Opinion Pool,” *The Annals of Mathematical Statistics*, 32, 1339–1342.
- Taylor, J. W. (2020), “Forecast Combinations for Value at Risk and Expected Shortfall,” *International Journal of Forecasting*, 36(2), 428–441.
- Thorey, J., Chaussin, C., and Mallet, V. (2018), “Ensemble Forecast of Photovoltaic Power with Online CRPS Learning,” *International Journal of Forecasting*, 34(4), 762–773.
- Timmermann, A. (2006), “Forecast Combinations,” in *Handbook of Economic Forecasting*, eds. G. Elliott, C. Granger, and A. Timmermann, Vol. 1, Radarweg 29, Amsterdam 1043 NX, Netherlands: Elsevier, chapter 4, pp. 135–196.
- van der Vaart, A. (1998), *Asymptotic Statistics*, 32 Avenue of the Americas, New York, NY 10013-2473, USA: Cambridge University Press.
- Wang, L., Wang, Z., Qu, H., and Liu, S. (2018), “Optimal Forecast Combination Based on Neural Networks for Time Series Forecasting,” *Applied Soft Computing*, 66, 1–17.
- Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2022), “Forecast Combinations: an Over 50-Year Review,” arXiv:2205.04216.
- West, K. D. (1996), “Asymptotic Inference About Predictive Ability,” *Econometrica*, pp. 1067–1084.

The Impact of Sampling Variability on Estimated Combinations of Distributional Forecasts*

Ryan Zischke^{†‡§}, Gael M. Martin[†], David T. Frazier[†], D. S. Poskitt[†]

June 6, 2022

Appendix: Theoretical Results

The appendix contains technical details supporting the theory, simulations and empirical exercises developed in the main paper. Appendix [A.1](#) provides standard regularity conditions assumed in the theorems of Section [3](#). In Appendix [A.2](#), we provide a GMM representation of two-stage forecast combinations and provide expressions for the asymptotic sampling distributions of the one- and two-stage parameter estimates. The structure of the asymptotic covariance matrix for the two-stage parameters are further explored in Appendix [A.3](#). The proofs for Theorems [1](#) and [2](#) can be found in Appendix [A.4](#).

*This research has been supported by Australian Research Council (ARC) Discovery Grant DP200101414. Frazier was also supported by ARC Early Career Researcher Award DE200101070; and Martin and Frazier were provided support by the ARC Centre of Excellence in Mathematics and Statistics.

[†]Department of Econometrics and Business Statistics, Monash University.

[‡]Methodology Division, Australian Bureau of Statistics. The views expressed in this paper are those of the authors alone, and do not in any way represent the Methodology Division, the Australian Bureau of Statistics, the Australian Public Service, or the Australian Government.

[§]Corresponding author

E-mail address: ryan.zischke@monash.edu.

A.1 Regularity Conditions

Assumption 1. *The parameter space Θ is compact. There exists a real-valued deterministic function $\mathcal{S}_0(\theta) \equiv \mathcal{S}_0(\eta, \gamma)$, continuous on Θ , such that the following are satisfied.*

1. $\sup_{\theta \in \Theta} \|\mathcal{S}_n(\theta) - \mathcal{S}_0(\theta)\| = o_p(1)$.
2. *There exists a unique vector $\theta^0 := \operatorname{argmax}_{\theta \in \Theta} \mathcal{S}_0(\theta)$.*

Assumption 2. *For each $j = 1, \dots, K$, there exists a real-valued deterministic function $\mathcal{S}_0(\gamma_j)$, continuous on Γ_j , such that the following are satisfied.*

1. $\sup_{\gamma_j \in \Gamma_j} \|\mathcal{S}_n(\gamma_j) - \mathcal{S}_0(\gamma_j)\| = o_p(1)$.
2. *There exists a unique vector $\gamma_j^* := \operatorname{argmax}_{\gamma_j \in \Gamma_j} \mathcal{S}_0(\gamma_j)$.*
3. *There exists a unique vector $\eta^* := \operatorname{argmax}_{\eta \in \mathcal{E}} \mathcal{S}_0(\eta, \gamma^*)$, where $\gamma^* := [\gamma_1^* \ \dots \ \gamma_K^*]'$.*

Remark 7. Assumptions 1 and 2 give the requisite regularity conditions to ensure consistency of $\hat{\theta}_n$ and $\tilde{\theta}_n$; the proofs of which are standard and hence omitted for brevity.

Furthermore, we assume the following, which, when satisfied, will allow us to deduce the asymptotic distributions of $\hat{\theta}_n$ and $\tilde{\theta}_n$.

Assumption 3. *The following are satisfied.*

1. $\theta^0, \theta^* \in \operatorname{Int}(\Theta)$.
2. *The functions $\mathcal{S}_0(\theta)$ and $\mathcal{S}_n(\theta)$ are twice continuously differentiable on $\operatorname{Int}(\Theta)$.*
3. *The functions $q_n(\eta, \gamma) := \partial \mathcal{S}_n(\eta, \gamma) / \partial \theta$ and $\tilde{g}_n(\eta, \gamma)$, defined in Appendix A.2, exist and satisfy the following.*

(a) $\sqrt{n}q_n(\eta^0, \gamma^0) \Rightarrow N(0, V^0)$ and $\sqrt{n}\tilde{g}_n(\eta^*, \gamma^*) \Rightarrow N(0, V^*)$.

(b) *There exist matrices $M^0(\theta) := \partial^2 \mathcal{S}_0(\theta) / \partial \theta \partial \theta'$ and $M^*(\theta)$, nonsingular at θ^0 and θ^* , respectively, such that for any nonnegative sequence $\delta_n \rightarrow 0$,*

$$\sup_{\|\theta - \theta^*\| \leq \delta_n} \|\partial \tilde{g}_n(\theta) / \partial \theta' - M^*(\theta)\| = o_p(1),$$

$$\sup_{\|\theta - \theta^0\| \leq \delta_n} \|\partial q_n(\theta) / \partial \theta' - M^0(\theta)\| = o_p(1).$$

A.2 GMM Representation of One- and Two-Stage Forecast Combinations

Consider the derivatives $g_n(\eta, \gamma) := \partial \mathcal{S}_n(\eta, \gamma) / \partial \eta$ and $m_{j,n}(\gamma_j) := \frac{1}{n} \sum_{t=1}^n \partial S(F_{j,t}^{\gamma_j}, Y_t) / \partial \gamma_j$, and stack $m_{j,n}(\gamma_j)$ to obtain $m_n(\gamma) := [m_{1,n}(\gamma_1)' \cdots m_{K,n}(\gamma_K)']'$. Since $\tilde{\eta}_n$ maximizes $\mathcal{S}_n(\eta, \tilde{\gamma}_n)$ and $\tilde{\gamma}_{j,n}$ maximizes $\frac{1}{n} \sum_{t=1}^n S(F_{j,t}^{\gamma_j}, Y_t)$, we have that, under standard regularity conditions, $\eta = \tilde{\eta}_n$ solves

$$g_n(\eta, \tilde{\gamma}_n) = 0, \tag{A.1}$$

with probability approaching one, and $\gamma = \tilde{\gamma}_n$ solves

$$m_n(\gamma) = \begin{bmatrix} m_{1,n}(\gamma) \\ \vdots \\ m_{K,n}(\gamma) \end{bmatrix} = 0, \tag{A.2}$$

with probability approaching one. Stacking g_n and m_n as \tilde{g}_n , we see that (A.1) and (A.2) form the joint moment equation

$$\tilde{g}_n(\tilde{\eta}_n, \tilde{\gamma}_n) := \begin{bmatrix} g_n(\eta, \gamma) \\ m_n(\gamma) \end{bmatrix} \Big|_{\eta=\tilde{\eta}_n, \gamma=\tilde{\gamma}_n}.$$

The two-stage estimator in (5) is therefore a GMM estimator based on the moment function \tilde{g}_n . See [Newey and McFadden \(1994\)](#) for further details regarding GMM.

Now let V^0 and V^* be the asymptotic covariance matrices of the scaled gradient vectors $\sqrt{n} \partial \mathcal{S}_n(\theta^0) / \partial \theta$ and $\sqrt{n} \tilde{g}_n(\eta^*, \gamma^*)$ of the one- and two-stage estimators, respectively. Given the assumptions in [Appendix A.1](#), we have, via the usual arguments,

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta^0) &\Rightarrow N(0, W^0), \\ W^0 &:= [M^0]^{-1} V^0 [M^0]^{-1'}, \end{aligned} \tag{A.3}$$

and

$$\begin{aligned}\sqrt{n}(\tilde{\theta}_n - \theta^*) &\Rightarrow N(0, W^*), \\ W^* &:= [M^*]^{-1}V^*[M^*]^{-1'},\end{aligned}$$

where $M^0 := \text{plim}_n \partial q_n(\theta^0)/\partial \theta'$ and $W^* := \text{plim}_n \partial \tilde{g}_n(\theta^*)/\partial \theta'$. The asymptotic covariance matrix W^* of the two-stage estimator $\tilde{\theta}_n$ has a specific structure due to its two-stage nature. See Appendix A.3 for details.

A.3 Structure of the Asymptotic Covariance Matrix of Two-Stage Forecast Combination Parameter Estimates

Recall that the asymptotic sampling distribution of the two-stage forecast combination parameters estimates is

$$\begin{aligned}\sqrt{n}(\tilde{\theta}_n - \theta^*) &\Rightarrow N(0, W^*), \\ W^* &:= [M^*]^{-1}V^*[M^*]^{-1'},\end{aligned}$$

where V^* is the asymptotic covariance matrix of the normalized gradient $\sqrt{n}\tilde{g}_n(\eta^*, \gamma^*)$, and $M^* := \text{plim}_n \partial \tilde{g}_n(\theta^*)/\partial \theta$. See Appendix A.2 for the definition of \tilde{g}_n , and for the definitions of g_n and m_n , used below.

The matrix M^* has a particular structure, namely, denoting $G_\eta := \text{plim}_n \partial g_n(\theta^*)/\partial \eta$, $G_\gamma := \text{plim}_n \partial g_n(\theta^*)/\partial \gamma$, and $M_\gamma := \text{plim}_n \partial m_n(\gamma^*)/\partial \gamma'$, we have

$$M^* := \begin{bmatrix} G_\eta & G_\gamma \\ 0 & M_\gamma \end{bmatrix},$$

so that

$$[M^*]^{-1} := \begin{bmatrix} G_\eta^{-1} & -G_\eta^{-1}G_\gamma M_\gamma^{-1} \\ 0 & M_\gamma^{-1} \end{bmatrix}.$$

Using this, we can conclude that (see, e.g., Theorem 6.1 in [Newey and McFadden 1994](#) for

details)

$$\begin{aligned}
\sqrt{n}(\tilde{\eta}_n - \eta^*) &\Rightarrow N(0, W_\eta^*), \\
W_\eta^* &:= G_\eta^{-1} V_\eta^* G_\eta^{-1'}, \\
V_\eta^* &:= \lim_{n \rightarrow \infty} \text{Var} [\sqrt{n} \{g_n(\theta^*) + G_\gamma M_\gamma^{-1} m_n(\gamma^*)\}].
\end{aligned} \tag{A.4}$$

Moreover, comparing (A.3) and (A.4), we see that ignoring the first stage in estimating the asymptotic distribution of the second-stage estimator, $\tilde{\eta}_n$, is equivalent to assuming $\sqrt{n}G_\gamma M_\gamma^{-1} m_n(\gamma^*) = o_p(1)$. Thus, ignoring the first stage leads to valid second-stage inference if $G_\gamma = 0$. Assuming we may differentiate inside the expectation, the implicit function theorem ensures the existence of a unique function $\eta^0(\gamma) := \operatorname{argmax}_{\eta \in \mathcal{E}} \mathcal{S}_0(\eta, \gamma)$, defined in a neighborhood of γ^* , such that

$$\frac{\partial \eta^0(\gamma^*)}{\partial \gamma'} = -G_\eta^{-1} G_\gamma.$$

Now consider the condition $\partial \eta^0(\gamma^*) / \partial \gamma' = 0$. Under our regularity conditions, this is equivalent to the expected-score-maximizing second-stage parameter vector $\eta^* = \eta^0(\gamma)$ being constant in a neighborhood of γ^* (Theorem 6.2, Newey and McFadden 1994). In this case, we would have $G_\gamma = 0$, and ignoring the first stage would lead to valid second stage inference.

In general, we see that ignoring first-stage estimation will deliver invalid inference on second-stage parameters, excepting cases where the optimal value of the first-stage parameters, γ^* , has no impact on the optimal value of the combination weights, η^* .

A.4 Proofs

Proof of Theorem 1. The first implication follows directly from Assumptions 1 and 2, and the definition of $\tilde{\theta}_n$ and $\hat{\theta}_n$. The second result is a consequence of the two-stage nature of

$\tilde{\theta}_n$ and a first-order Taylor series expansion. In particular,

$$\begin{aligned}
& \sqrt{n}\{\mathcal{S}_0(\tilde{\eta}_n, \tilde{\gamma}_n) - \mathcal{S}_0(\eta^*, \gamma^*)\} \\
&= [\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\eta]' \sqrt{n}(\tilde{\eta}_n - \eta^*) \\
&\quad + [\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\gamma]' \sqrt{n}(\tilde{\gamma}_n - \gamma^*) \\
&\quad + [\partial\mathcal{S}_0(\tilde{\theta}_n)/\partial\theta - \partial\mathcal{S}_0(\theta^*)/\partial\theta]' \sqrt{n}(\tilde{\theta}_n - \theta^*) \\
&= [\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\gamma]' \sqrt{n}(\tilde{\gamma}_n - \gamma^*) + o_p(1)
\end{aligned}$$

for $\|\tilde{\theta}_n - \theta^*\| \leq \|\tilde{\theta}_n - \theta^*\| = \mathcal{O}_p(n^{-1/2})$. By Assumptions 2.3, 3.1 and 3.2, $\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\eta = 0$ and the term on the second line disappears. By Assumption 3.2, the term on the fourth line converges to zero in probability. Similarly,

$$\begin{aligned}
& \sqrt{n}\{\mathcal{S}_0(\eta^*, \tilde{\gamma}_n) - \mathcal{S}_0(\eta^*, \gamma^*)\} \\
&= [\partial\mathcal{S}_0(\eta^*, \tilde{\gamma}_n)/\partial\gamma]' \sqrt{n}(\tilde{\gamma}_n - \gamma^*) \\
&= [\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\gamma]' \sqrt{n}(\tilde{\gamma}_n - \gamma^*) + o_p(1)
\end{aligned}$$

for $\|\tilde{\gamma}_n - \gamma^*\| \leq \|\tilde{\gamma}_n - \gamma^*\| = \mathcal{O}_p(n^{-1/2})$. Subtracting the two expansions yields the result. \square

Proof of Theorem 2, Part 1. Recall that $\sqrt{n}(\tilde{\theta}_n - \theta^*) \Rightarrow N(0, W^*)$, and apply the first-order delta method, using Assumption 3.2 to control the error. The asymptotic covariance matrix follows from $\partial\mathcal{S}_0(\eta^*, \gamma^*)/\partial\eta = 0$ and some block matrix manipulation. \square

Proof of Theorem 2, Part 2. By Assumptions 1.2, 3.1 and 3.2, we have $\partial\mathcal{S}_0(\theta^0)/\partial\theta = 0$. Now recall that $\sqrt{n}(\hat{\theta}_n - \theta^0) \Rightarrow X$, so we apply the second-order delta method using a second-order Taylor series expansion of $\mathcal{S}_0(\hat{\theta}_n)$ around $\theta = \theta^0$. Use Assumption 3.3 (b) to control the error. \square

Supplementary References

Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, Radarweg 29, Amsterdam 1043 NX, Netherlands: Elsevier, chapter 36, pp. 2111–2245.