Robust Forecast Superiority Testing with an Application to Assessing Pools of Expert Forecasters*

Valentina Corradi[1], Sainan Jin[2], and Norman R. Swanson[3]

[1]University of Surrey, [2]Singapore Management University, and [3]Rutgers University

August 2020

Abstract

We develop a forecast superiority testing methodology which is robust to the choice of loss function. Following Jin, Corradi and Swanson (JCS: 2017), we rely on a mapping between generic loss forecast evaluation and stochastic dominance principles. However, unlike JCS tests, which are not uniformly valid, and have correct asymptotic size only under the least favorable case, our tests are uniformly asymptotically valid and non-conservative. These properties are derived by first establishing uniform convergence (over error support) of HAC variance estimators and of their bootstrap counterparts, and by extending the asymptotic validity of generalized moment selection tests to the case of non-vanishing recursive parameter estimation error. Monte Carlo experiments indicate good finite sample performance of the new tests, and an empirical illustration suggests that prior forecast accuracy matters in the Survey of Professional Forecasters. Namely, for our longest forecast horizons (4 quarters ahead), selecting pools of expert forecasters based on prior accuracy results in ensemble forecasts that are superior to those based on forming simple averages and medians from the entire panel of experts.

Keywords: Robust Forecast Evaluation, Many Moment Inequalities, Bootstrap, Estimation Error, Combination Forecasts, Survey of Professional Forecasters.

---

*Valentina Corradi, School of Economics, University of Surrey, Guildford, Surrey, GU2 7XH, UK, v.corradi@surrey.ac.uk; Sainan Jin, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, snjin@smu.edu.sg; and Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, nswanson@econ.rutgers.edu. We are grateful to Kevin Lee, Patrick Marsh, Luis Martins, Jams Mitchell, Alessia Paccagini, Paulo Parente, Ivan Petrella, Valerio Poti, Barbara Rossi, Simon Van Norden, Claudio Zoli, and to the participants at the 2018 NBER-NSF Times Series Conference, the 2016 European Meeting of the Econometric Society, Conference for 50 years of Keynes College at Kent University, and seminars at Mannheim University, the University of Nottingham, University College Dublin, Instituto Universitário de Lisboa, Universita' di Verona and the Warwick Business School for useful comments and suggestions. Additionally, many thanks are owed to Mingmian Cheng for excellent research assistance.

1 Introduction

Forecast accuracy is typically measured in terms of a given loss function, with quadratic and absolute loss being the most common choices. In recent years, there has been a growing discussion about the choice of the "right" loss function. Gneiting (2011) stresses the importance of matching the quantity to be forecasted and the choice of loss function (or scoring rule). The latter is said to be consistent for a given statistical functional (e.g. the mean or the median), if expected loss is minimized when such a functional is used. In a recent paper, Patton (2019) shows that if forecasts are based on nested information sets and on correctly specified models, then in the absence of estimation error, forecast ranking is robust to the choice of loss function within the class of consistent functions. On the other hand, if any of the above conditions fail, then model ranking is dependent on the specific loss function used. This is an important finding, given that it is natural for researchers to focus on the comparison of multiple misspecified models; immediately implying that model rankings are loss function dependent.

In summary, given the importance of loss function dependence when comparing forecast accuracy, an issue of key concern to empirical economists is the construction of loss function robust forecast accuracy tests. A loss function free forecast evaluation criterion of interest should be based on the distribution of raw forecast errors. Heuristically, one can define the best forecasting model as that producing errors having a step cumulative distribution function that is equal to zero on the negative real line and equal to one on the positive real line. Diebold and Shin (2015, 2017) build on this idea, and suggest choosing the model for which the cumulative distribution of the forecast errors is closest to a step function. This idea is also discussed in Corradi and Swanson (2013). Jin, Corradi and Swanson (JCS: 2017) establish a one to one mapping between generalized loss (GL) forecast superiority and first order stochastic dominance, as well as a one to one mapping between convex loss function (CL) and second order stochastic dominance.[1] In particular, they show that the "best" model (regardless of loss function) according to a GL (CL) function is the one which is first (second) order stochastically dominated on the negative real line and first (second) order stochastically dominant on the positive real line, when comparing forecast errors. In this sense, JCS (2017) establish that loss function free tests for forecast superiority can be framed in terms of tests for stochastic dominance. In this paper, we note that tests for stochastic dominance can be seen as tests for infinitely many moment inequalities. This allows us to utilize tools recently developed by Andrews and Shi (2013, 2017) to derive asymptotically uniformly valid and non conservative forecast superiority tests. Importantly, these tests improve over those introduced in JCS (2017), as the latter were asymptotically non conservative only in the least favorable case under the null (i.e., when all moment weak inequalities hold with equality). Needless to say, controlling for slack inequalities is crucial when there are infinitely many of them.

[1] A loss function is a GL function if it is monotonically non-decreasing as the error moves away from zero. Additionally, CL functions are the subset of convex GL functions.

The implementation of our tests require that sample moments are standardized by an estimator of the standard deviation. Now, forecast errors are typically non martingale difference sequences, either because they are based on dynamically misspecified models or because forecasters do not efficiently use all the available information, in the case of subjective predictions. Hence, we require heteroskedasticity and autocorrelation (HAC) robust variance estimators. In our set-up, each variance estimator depends on a specific point in the forecasting error support. Thus, in order to introduce our new tests for forecast superiority, we must establish the consistency of HAC variance estimators uniformly over the error support. Moreover, in order to carry out inference, using our tests, we also establish uniform convergence of the HAC variance estimator bootstrap counterparts. Because of the presence of the lag truncation parameter, uniform convergence of HAC estimators and of their bootstrap analogs does not follow straightforwardly from uniform convergence of (kernel) nonparametric estimators. To the best of our knowledge this contribution is a novel addition to the vast literature on HAC covariance matrix estimation. In the sequel, we focus on the case of judgmental forecasts, in which there is no parameter estimation error. In a supplemental online appendix, we consider the case of predictions based on estimated models, and extend all of our results to the case of non vanishing estimation error. This is accomplished under a recursive estimation scheme, by extending the recursive block bootstrap introduced in Corradi and Swanson (2007).

Linton, Song and Whang (2010) also develop tests for stochastic dominance which are correctly asymptotically sized over the boundary of the null, for the pairwise comparison case. A key role in their asymptotic analysis is played by the contact set (i.e., the set of x over which the two CDFs are equal). However, the notion of contact set does not extend straightforwardly to the multiple comparison case considered in this paper. It should also be noted that other papers have addressed the problem of forecast evaluation in the absence of full specification of the loss function. For example, Patton and Timmermann (2007) have studied forecast optimality under only generic assumptions on the loss function. However, they do not address the issue of forecast ranking under (partially) unknown loss. More recently, Barendse and Patton (2019) introduce forecast multiple comparison under loss functions which are specified only up to a shape parameter.

We assess the forecast superiority testing methodology discussed in this paper via a series of Monte Carlo experiments. Simulation results show that our new tests are in some key cases much more accurately sized and have much higher power than JCS tests. For example, in size experiments where DGPs contain some models which are worse than the benchmark model, our new tests are substantially better sized than the tests of JCS (2017). Additionally, our new tests exhibit notable power gains, relative to JCS tests, in power experiments where DGPs contain some alternative models that dominate the benchmark, while others are strictly dominated. These findings are as expected, given that JCS tests are undersized, while our new tests are asymptotically non conservative.

In an empirical illustration, we apply our testing procedure to the Survey of Professional Forecasters (SPF) dataset. In the SPF, participants are told which variables to forecast and whether they should provide a point forecast or instead a probability interval, but they are not given a loss function (see Crushore (1993) for a detailed description of the SPF). In the context of analyzing the predictive content of the SPF, many papers find evidence of the usefulness of forecast combinations constructed using individual SPF predictions, under quadratic or absolute loss. For example, Zarnowitz and Braun (1993) find that using the mean or median provides a consensus forecast with lower average errors than most individual forecasts. Aiolfi, Capistrán, and Timmermann (2011) and Genre, Kenny, Meyler, and Timmermann (2013) find that equal weighted averages of SPF and ECB *(European Central Bank)* SPF forecasts often outperform model based forecasts. In our illustration, we depart from these papers by noting that the SPF naturally lends itself to loss function free forecast superiority testing, since participants are not given loss functions. In light of this, we apply our new tests, and show that forecast averages (and medians) from small pools of survey participants ranked according to recent forecast performance are preferred to forecast averages based on the entire pool of experts, for our longest forecast horizon (1-year ahead). We thus conclude that simple average and median forecasts can in some cases be "beaten", regardless of loss function.

The rest of the paper is organized as follows. Section 2 outlines the set-up and introduces our new tests. Section 3 establishes the asymptotic properties of the tests in the context of generalized moment selection. Section 4 contains the results of our Monte Carlo experiments, and Section 5 contains the results of our analysis of GDP growth forecasts from the SPF. Finally, Section 6 provides a number of concluding remarks. Proofs are gathered in an appendix. In a supplemental appendix, we establish the asymptotic properties of our new tests in the context of non-vanishing parameter estimation error, for the recursive estimation schemes.

2 Forecast Superiority Tests

Assume that we have a time series of forecast errors for each model/forecaster. Namely, we observe $e_{j,t}$, for $j = 1, ..., k$ and $t = 1, ...n$, where k denotes the number of models/forecasters, and n denotes the number of observations. As stated earlier, we focus on the case in which we can ignore estimation error, such as when forecasts are judgmental or subjective. Surveys including the SPF are leading examples of judgmental forecasts. The case of non-vanishing recursive estimation error is analyzed in the supplemental appendix. Hereafter, the sequence $e_{1,t}$, $t = 1, ..., n$ is called the "benchmark". In the context of the SPF, an example of a relevant benchmark against which to compare all other sequences is the consensus forecast constructed as the simple arithmetic average of individual forecasts in the survey. Our goal is to test whether there exists some competing forecast that is superior to the benchmark for any loss function, L, satisfying Assumption A0.

Assumption A0 (i) $L \in \mathcal{L}_G$ if $L : \mathbb{R} \to \mathbb{R}^+$ is continuously differentiable, except for finitely many points, with derivative L', such that $L'(z) \leq 0$, for all $z \leq 0$, and $L'(z) \geq 0$, for all $z \geq 0$. (ii) $L \in \mathcal{L}_C$ is a convex function belonging to $\mathcal{L}_G$.

Note that $\mathcal{L}_G$ includes most of the loss functions commonly used by practitioners, including asymmetric loss, and it basically coincides with notion of generalized loss in Granger (1999). The only restriction is that the loss depends solely on the forecast errors. This rules out the class of loss function considered in e.g. Section 3 of Patton and Timmermann (2007).

Hereafter, let $F_j(x)$ denote the cumulative distribution function (CDF) of forecast error e_j. Also, define $sgn(x) = 1 \; if \; x \geq 0$ *and* $sgn(x) = -1 \; if \; x < 0$. Propositions 2.2 and 2.3 in JCS (2017) establish the following results.

1. *For any* $L \in L_G$, $E(L(e_1)) \leq E(L(e_2))$, *if and only if* $(F_2(x) - F_1(x))sgn(x) \leq 0$, *for all* $x \in \mathcal{X}$.

2. *For any* $L \in L_C$, $E(L(e_1)) \leq E(L(e_2))$, *if and only if* $\left(\int_{-\infty}^{x}(F_1(t) - F_2(t))dt1(x < 0) + \int_{x}^{\infty}(F_2(t) - F_1(t))dt1(x \geq 0)\right) \leq 0$, *for all* $x \in \mathcal{X}$.

The first statement establishes a mapping between GL forecast superiority and first order stochastic dominance (FOSD). In particular, e_1 is not GL dominated by e_2 if $F_1(x)$ lies below $F_2(x)$ on the negative real line, and lies above $F_2(x)$ on the positive real line. Indeed, this ensures that we choose the forecast whose CDF has larger mass around zero. Likewise, the second statement establishes a mapping between CL superiority and second order stochastic dominance.

In this framework, it follows that testing for loss function robust forecast superiority involves testing:

$$H_0^G : \max_{j=2,...,k} (E(L(e_1)) - E(L(e_k))) \leq 0 \text{ for all } L \in L_G \tag{2.1}$$

versus

$$H_A^G : \max_{j=2,...,k} (E(L(e_1)) - E(L(e_k))) > 0 \text{ for some } L \in L_G, \tag{2.2}$$

with H_0^C and H_A^C, defined analogously, replacing L_G with L_C.
Hereafter, let $\mathcal{X} = \mathcal{X}^- \cup \mathcal{X}^+$ be the union of the support of $(e_1, ..., e_k)$. Given the equivalence between GL (CL) forecast superiority and first (second) order stochastic dominance, we can restate H_0^G, H_A^G, H_0^C and H_A^C as

$$\begin{aligned} H_0^G &= H_0^{G-} \cap H_0^{G+} \\ &: \left(F_1(x) - F_j(x) \leq 0, \text{ for } j = 2, ..., k, \text{ and for all } x \in \mathcal{X}^-\right) \\ &\quad \cap \left(F_j(x) - F_1(x) \leq 0, \text{ for } j = 2, ..., k, \text{ and for all } x \in \mathcal{X}^+\right) \end{aligned}$$

versus

$$\begin{aligned} H_A^G &= H_A^{G-} \cup H_A^{G+} \\ &: \quad \left(F_1(x) - F_j(x) > 0, \text{ for some } j = 2, ..., k, \text{ and for some } x \in \mathcal{X}^-\right) \\ &\quad \cup \left(F_j(x) - F_1(x) > 0, \text{ for some } j = 2, ..., k, \text{ and for some } x \in \mathcal{X}^+\right). \end{aligned}$$

Analogously,

$$\begin{aligned} H_0^C &= H_0^{C-} \cap H_0^{C+} \\ &: \quad \left(\int_{-\infty}^{x} (F_1(t) - F_j(t))dt \leq 0, \text{ for } j = 2, ..., k, \text{ and for all } x \in \mathcal{X}^-\right) \\ &\quad \cap \left(\int_{x}^{\infty} (F_j(t) - F_1(t))dt \leq 0, \text{ for } j = 2, ..., k, \text{ and for all } x \in \mathcal{X}^+\right) \end{aligned}$$

versus

$$\begin{aligned} H_A^C &= H_A^{C-} \cup H_A^{C+} \\ &: \quad \left(\int_{-\infty}^{x} (F_1(t) - F_j(t))dt > 0, \text{ for some } j = 2, ..., k, \text{ and for some } x \in \mathcal{X}^-\right) \\ &\quad \cup \left(\max_{j=2,...,k} \int_{x}^{\infty} (F_j(t) - F_1(t))dt > 0, \text{ for some } j = 2, ..., k, \text{ and for some } x \in \mathcal{X}^+\right). \end{aligned}$$

It is immediate to see that H_0^G and H_0^C can be written as the intersection of $(k-1)$ moment inequalities, which have to hold uniformly over $\mathcal{X}$. This gives rise to an infinite number of moment conditions. Andrews and Shi (2013) develop tests for conditional moment inequalities, and as is well known in the literature on consistent specification testing (e.g., see Bierens (1982, 1990)) a finite number of conditional moments can be transformed into an infinite number of unconditional moments. The same is true in the case of weak inequalities. Andrews and Shi (2017) consider tests for conditional stochastic dominance, which are then characterized by an infinite number of conditional moment inequalities and so by a "twice" infinite number of unconditional inequalities. Recalling that our interest is on testing GL or CL forecast superiority as in (2.1) and (2.2), we confine our attention to unconditional testing of stochastic dominance.

Because of the discontinuity at zero in the tests, $H_0^{G+}\left(H_0^{C^+}\right)$ and $H_0^{G-}(H_0^{C-})$ should be tested separately, and then one can use Holm (1979) bounds to control the two resulting p-values (see Rules TG and TC in JCS (2017)). In the sequel, for the sake of brevity, but without loss of generality, we focus our discussion on testing H_0^{G+} versus H_A^{G+} and H_0^{C+} versus H_A^{C+}. However, when defining statistics, some discussion of the statistics associated with the case where $x \in \mathcal{X}^-$ is also given, when needed for clarity of exposition.

We begin by testing GL forecast superiority. Let $G^+(x) = \left(G_2^+(x), ..., G_k^+(x)\right)$, with $G_j^+(x) = F_j(x) - F_1(x)$, for $x \geq 0$. Define the empirical analog of $G^+(x)$ as $G_n^+(x) = \left(G_{2,n}^+(x), ..., G_{k,n}^+(x)\right)$, and for $x \geq 0$, let

$$G_{j,n}^+(x) = \widehat{F}_{j,n}(x) - \widehat{F}_{1,n}(x), \tag{2.3}$$

where $\widehat{F}_{j,n}(x)$ denotes the empirical CDF of e_j. Similarly, let $C^+(x) = \left(C_2^+(x), ..., C_k^+(x)\right)$, with $C_j^+(x) = \int_x^\infty \left(F_j(t) - F_1(t)\right) dt 1(x \geq 0)$. Define the empirical analog of $C^+(x)$ as $C_n^+(x) = \left(C_{2,n}^+(x), ..., C_{k,n}^+(x)\right)$, and let

$$\begin{aligned} C_{j,n}^+(x) &= \int_x^\infty \left(\widehat{F}_{j,n}(t) - \widehat{F}_{1,n}(t)\right) dt 1(x \geq 0) \\ &= \frac{1}{n}\sum_{t=1}^n \left([(e_{1,t} - x)]_+ - [(e_{j,t} - x)]_+ \right), \end{aligned} \tag{2.4}$$

where $[z]_+ = \max\{0, z\}$. Further, define

$$\Sigma^{G+}(x, x') = \text{acov}\left(\sqrt{n}G^+(x), \sqrt{n}G^+(x')\right) \tag{2.5}$$

and

$$\overline{\Sigma}_n^{G+}(x, x') = \widehat{\Sigma}_n^{G+}(x, x') + \varepsilon I_{k-1}, \tag{2.6}$$

where $\varepsilon \geq 0$, and where $\widehat{\Sigma}_n^{G+}(x, x')$ is the sample analog of $\Sigma^{G+}(x, x')$. In (2.6), the role of the additional εI_{k-1} term is to correct for the possible singularity of the covariance estimator, for certain values of x. This is the case when we compare forecast errors from nested models. Let $\widehat{u}_{j,t}(x) = 1\{e_{j,t} \leq x\} - \frac{1}{n}\sum_{t=1}^n 1\{e_{j,t} \leq x\}$, so that the $jj-$th element of $\widehat{\Sigma}_n^{G+}(x, x)$ is given by

$$\begin{aligned} \widehat{\sigma}_{jj,n}^{2,G+}(x) &= \frac{1}{n}\sum_{t=1}^n (\widehat{u}_{j,t}(x) - \widehat{u}_{1,t}(x))^2 \\ &\quad + 2\frac{1}{n}\sum_{\tau=1}^{l_n}\sum_{t=\tau+1}^n w_\tau (\widehat{u}_{j,t}(x) - \widehat{u}_{1,t}(x))(\widehat{u}_{j,t-\tau}(x) - \widehat{u}_{1,t-\tau}(x)), \end{aligned} \tag{2.7}$$

where $w_\tau = 1 - \frac{\tau}{1+l_n}$, with $l_n \to \infty$ as $n \to \infty$. Also, let $\overline{\sigma}_{jj,n}^{2,G^+}(x, x')$ be the $jj-$th element of $\overline{\Sigma}_n^{G+}(x, x')$, and let $\overline{\sigma}_{jj,n}^{2,C+}(x, x')$ be the $jj-$th element of $\overline{\Sigma}_n^{C+}(x, x')$. Analogously,

$$\Sigma^{C+}(x, x') = \text{acov}\left(\sqrt{n}C^+(x), \sqrt{n}C^+(x')\right)$$

and

$$\overline{\Sigma}_n^{C+}(x, x') = \widehat{\Sigma}_n^{C+}(x, x') + \varepsilon I_{k-1},$$

where$\widehat{\Sigma}_n^{C+}(x, x')$ is the sample analog of $\Sigma^{C+}(x, x')$

Furthermore, $\widehat{\sigma}_{jj,n}^{2,C+}(x)$ is constructed by replacing $\widehat{u}_{1,t}(x)$ and $\widehat{u}_{j,t}(x)$ in the above expression with

$$\widehat{\eta}_{1,t}(x) = [(e_{1,t} - x)]_+ - \frac{1}{n}\sum_{t=1}^n [(e_{1,t} - x)]_+$$

and

$$\widehat{\eta}_{j,t}(x) = [(e_{j,t} - x)]_+ - \frac{1}{n}\sum_{t=1}^n [(e_{j,t} - x)]_+ .$$

Note that $G^{-}_{j,n}(x), C^{-}_{j,n}(x), \widehat{\sigma}^{2,G-}_{jj,n}(x)$, and $\widehat{\sigma}^{2,C-}_{jj,n}(x)$ can be defined by utilizing the sgn function. Namely, regardless of whether $x \geq 0$ or $x < 0$, one can construct $G_{j,n}(x) = \left(\widehat{F}_{j,n}(x) - \widehat{F}_{1,n}(x)\right) sgn(x)$ and

$$\begin{aligned} C_{j,n}(x) &= \int_{-\infty}^{x} \left(\widehat{F}_{1,n}(t) - \widehat{F}_{j,n}(t)\right) dt 1(x<0) - \int_{x}^{\infty} \left(\widehat{F}_{j,n}(t) - \widehat{F}_{1,n}(t)\right) dt 1(x \geq 0) \\ &= \frac{1}{n}\sum_{t=1}^{n} \left([(e_{1,t} - x)\, sgn(x)]_{+} - [(e_{j,t} - x)\, sgn(x)]_{+} \right), \end{aligned}$$

$$\begin{aligned} \widehat{\sigma}^{2,G}_{jj,n}(x) &= \frac{1}{n}\sum_{t=1}^{n} (\widehat{u}_{j,t}(x) - \widehat{u}_{1,t}(x))^2 \\ &\quad + 2\frac{1}{n}\sum_{\tau=1}^{l_n}\sum_{t=\tau+1}^{n} w_\tau (\widehat{u}_{j,t}(x) - \widehat{u}_{1,t}(x)) sgn(x) (\widehat{u}_{j,t-\tau}(x) - \widehat{u}_{1,t-\tau}(x)) sgn(x), \end{aligned}$$

and $\widehat{\sigma}^{2,C}_{jj,n}(x)$ by replacing $\widehat{u}_{1,t}(x)$ and $\widehat{u}_{j,t}(x)$ in the above expression with

$$\widehat{\eta}_{1,t}(x) = [(e_{1,t} - x)\, sgn(x)]_{+} - \frac{1}{n}\sum_{t=1}^{n} [(e_{1,t} - x)\, sgn(x)]_{+}$$

and

$$\widehat{\eta}_{j,t}(x) = [(e_{j,t} - x)\, sgn(x)]_{+} - \frac{1}{n}\sum_{t=1}^{n} [(e_{j,t} - x)\, sgn(x)]_{+}.$$

Given the above framework, our new robust forecast superiority test statistics are:

$$S_n^{G+} = \int_{x\in\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{0, \sqrt{n}\frac{G_{j,n}(x)}{\overline{\sigma}^{G}_{jj,n}(x)}\right\}\right)^2 dQ(x) \text{ and } S_n^{G-} = \int_{x\in\mathcal{X}^-} \sum_{j=2}^{k} \left(\max\left\{0, \sqrt{n}\frac{G_{j,n}(x)}{\overline{\sigma}^{G}_{jj,n}(x)}\right\}\right)^2 dQ(x) \tag{2.8}$$

and

$$S_n^{C+} = \int_{x\in\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{0, \sqrt{n}\frac{C_{j,n}(x)}{\overline{\sigma}^{C}_{jj,n}(x)}\right\}\right)^2 dQ(x) \text{ and } S_n^{C-} = \int_{x\in\mathcal{X}^-} \sum_{j=2}^{k} \left(\max\left\{0, \sqrt{n}\frac{C_{j,n}(x)}{\overline{\sigma}^{C}_{jj,n}(x)}\right\}\right)^2 dQ(x), \tag{2.9}$$

where Q is a weighting function defined below; $G^{+}_{j,n}(x)$ and $C^{+}_{j,n}(x)$ are the $j-$th components of $G^{+}_{n}(x)$ and $C^{+}_{n}(x)$, as defined in (2.3) and (2.4), respectively. Here, S_n^{G+} and S_n^{C+} are "sum" functions, as in equation (3.8) in Andrews and Shi (2013), and satisfy their Assumptions S1-S4, which are required to guarantee that convergence is uniform over the null DGPs.[2],[3] If $k = 2$ and $\overline{\sigma}_{jj,n}(x) = 1$, for all j and x (i.e. no standardization), then S_n^{G+} is the statistic used in Linton, Song and Whang (2010) for testing FOSD.

[2] Note that we could have constructed a different "sum" function, using the statistic in (3.9) of Andrews and Shi (2013).

[3] Recall that one main drawback of the $\max_{j=2,\ldots,k} \sup_{x\in\mathcal{X}^+} \sqrt{n}G^{+}_{n}$ statistic in JCS (2017) is that it diverges to $-\infty$ under some sequence of probability measures under the null, thus ruling out uniformity.

Of note is that in our context, potential slackness causes a discontinuity in the pointwise asymptotic distribution of the statistic.[4] This is because the pointwise asymptotic distribution is discontinuous, unless all moment conditions hold with equality. On the other hand, the finite sample distribution is not necessarily discontinuous. Thus, in the presence of slackness, the pointwise limiting distribution is not a good approximation of the finite sample distribution, and critical values based on pointwise asymptotics may be invalid. This is why we construct tests that are uniformly asymptotically valid (i.e., this is why we study the limiting distribution of our tests under drifting sequences of probability measures belonging to the null hypothesis). Moreover, in the infinite dimensional case, there is an additional source of discontinuity. In particular, the number of moment inequalities which contributes to the statistic varies across the different values of x. For example, the key difference between the case of $k = 2$ and $k > 2$ is that in the former case, for each value of x there is only one moment inequality which can be binding (or not). On the other hand, if $k = 3$, say, then for each value of x there can be either one or two moment inequalities which may be binding (or not), and whether or not a particular inequality is binding (or not) varies over x. Under this setup, we require the following assumptions in order to analyze the asymptotic behavior of our test statistics.

Assumption A1: For $j = 1, ..., k$, $e_{j,t}$ is strictly stationary and $\beta-$mixing, with mixing coefficients, $a_m = m^{-\beta}$, where $\beta > \frac{6\delta}{1-2\delta}$, $0 < \delta < 1/2$ and $\beta\delta > 1$.

Assumption A2: The union of the supports of $e_1, .., e_k$ is the compact set, $\mathcal{X} = \mathcal{X}^- \cup \mathcal{X}^+$.

Assumption A3: $F_j(x)$ has a continuos bounded density.

Assumption A4: The weighting function Q has full support $\mathcal{X}^+$ (or $\mathcal{X}^-$).

We use Assumption A2 in the proof of Lemma 1, where we require $\mathcal{X}^+$ in (2.8) and (2.9) to be a compact set. However, for the case of generalized loss superiority, the union of the supports of $e_1, .., e_k$ can be unbounded. This is because S_n^{G+} is bounded, regardless of the boundedness of the support. On the other hand, $S_n^{C^+}$ is bounded only when the union of the support of the forecasting error is bounded.

3 Asymptotic Properties

3.1 Uniform Convergence of the HAC Estimator

We now turn to a discussion of the estimation of the variance in our forecast superiority test statistics. If $e_{1,t}, ..., e_{k,t}$ were martingale difference sequences, then we can still use the sample second moment as a variance estimator, and uniform consistency will follow by application of an appropriate uniform law of large numbers. In our set-up we can assume that $e_1, ..., e_k$ are martingale difference sequences if either: (i) they are judgmental forecasts from professional forecasters, say, who efficiently use all available information at time t (a strong assumption, which is tested in the forecast rationality literature); or (ii)

[4] By pointwise asymptotic distribution we mean the limiting distribution under a fixed probability measure.

they are prediction errors from one-step ahead forecasts based on dynamically correctly specified models. With respect to (i), it is worth noting that professional forecasters may be rational, ex-post, according to some loss function (see Elliott, Komunjer and Timmermann (2005,2008), although it is not as likely that they are rational according to a generalized loss function. With respect to (ii), it should be noted that at most one model can be dynamically correctly specified for a given information set, and thus e_j cannot be a martingale difference sequence, for all $j = 1, ..., k$. In light of these facts, we allow for time dependence in the forecast error sequences used in our statistics, and use a HAC variance estimator in (2.8) and (2.9). In order to ensure that the HAC estimators converge uniformly over $\mathcal{X}^+$, it suffices to establish the counterpart of Lemma A1 of Supplement A of Andrews and Shi (2013) to the case of mixing sequences. This is done below.

Lemma 1*: Let Assumptions A1-A3 hold. Then, if* $l_n \approx n^\delta$ $0 < \delta < \frac{1}{2}$, *with* δ *defined as in Assumption A1:*

(i)

$$\sup_{x \in \mathcal{X}^+} \left| \widehat{\sigma}^{2,G+}_{jj,n}(x) - \sigma^{2,G+}_{jj}(x) \right| = o_p(1),$$

with $\sigma^{2,G+}_{jj}(x) = avar\left(\sqrt{n} G^+_{j,n}(x)\right)$; *and*

(ii)

$$\sup_{x \in \mathcal{X}^+} \left| \widehat{\sigma}^{2,C+}_{jj,n}(x) - \sigma^{2,C+}_{jj}(x) \right| = o_p(1),$$

with $\sigma^{2C+}_{jj}(x) = avar\left(\sqrt{n} C^+_{j,n}(x)\right)$.

Lemma 1 establishes the uniform convergence over $\mathcal{X}^+$ of HAC estimators. It is the time series counterpart of Lemma A1 in Andrews and Shi (2013). Of note is that we require $\beta-$mixing. This differs from the stationary pointwise HAC variance estimator case studied by Andrews (1991), where $\alpha-$mixing suffices, and where the mixing coefficients decline to zero slightly slower than in our Assumption A1. This is because there is a trade-off between the degree of dependence and the rate of growth of the lag truncation parameter in the HAC estimator. Indeed, in the uniform case, the covering number (e.g., see Andrews and Pollard (1994)) grows with both l_n and the degree of dependence, thus leading to a trade-off between the two. For example, in the case of exponential mixing series, δ can be arbitrarily close to 1/2.

For carrying out inference on our forecast superiority tests, we require a bootstrap analog of the HAC variance estimator, which can be constructed as follows. Using the block bootstrap, make b_n draws of length l_n from $e_{j,1}, ..., e_{j,n}$, in order to obtain $\left(e^*_{j,1}, ..., e^*_{j,n}\right) = \left(e_{j,I_{1+1}}, ..., e_{j,I_{1+l_n}}, ..., e_{j,I_{b_n+l_n}}\right)$, with $b_n l_n = n$, where the block size, l_n, is equal to the lag truncation parameter in the HAC estimator described above.[5] Now, let $u^*_{1,t}(x) = 1\left\{e^*_{1,t} \leq x\right\} - \frac{1}{n}\sum_{t=1}^n 1\left\{e_{1,t} \leq x\right\}$, $u^*_{j,t}(x) = 1\left\{e^*_{j,t} \leq x\right\} -$

[5] We thus use the same notation, l_n, for both the lag truncation parameter and the block length.

$\frac{1}{n}\sum_{t=1}^{n} 1\{e_{j,t} \leq x\}$, and

$$\widehat{\sigma}^{2*G+}_{jj,n}(x) = \frac{1}{b_n}\sum_{k=1}^{b_n}\left(\frac{1}{l_n^{1/2}}\sum_{i=1}^{l_n}\left(u^*_{j,(k-1)l_n+i}(x) - u^*_{1,(k-1)l_n+i}(x)\right)\right)^2. \tag{3.1}$$

Define $\widehat{\sigma}^{*2C^+}_{jj,n}(x)$ analogously, replacing $u^*_{1,t}(x)$ with $\eta^*_{1,t}(x) = \left[e^*_{1,t} - x\right]_+ - \frac{1}{n}\sum_{t=1}^{n} 1\left[e_{1,t} - x\right]_+$ and $u^*_{j,t}(x)$ with $\eta^*_{j,t}(x) = \left[e^*_{j,t} - x\right]_+ - \frac{1}{n}\sum_{t=1}^{n} 1\left[e_{j,t} - x\right]_+$. Additionally, define

$$\widehat{\sigma}^{2*G+}_{jj',n}(x) = \frac{1}{b_n}\sum_{k=1}^{b_n}\left(\frac{1}{l_n^{1/2}}\sum_{i=1}^{l_n}\left(u^*_{j,(k-1)l_n+i}(x) - u^*_{1,(k-1)l_n+i}(x)\right)\left(u^*_{j',(k-1)l_n+i}(x) - u^*_{1,(k-1)l_n+i}(x)\right)\right).$$

The following result holds.

Lemma 2: *Let Assumptions A1-A3 hold. Then, if $l_n \approx n^\delta$ $0 < \delta < \frac{1}{2}$, with δ defined as in Assumption A1:*

(i)

$$\sup_{x\in\mathcal{X}^+}\left|\widehat{\sigma}^{*G+}_{jj,n}(x) - \mathrm{E}^*\left(\widehat{\sigma}^{*G+}_{jj,n}(x)\right)\right| = o^*_p(1),$$

and (ii)

$$\sup_{x\in\mathcal{X}^+}\left|\widehat{\sigma}^{*C+}_{jj,n}(x) - \mathrm{E}^*\left(\widehat{\sigma}^{*C+}_{jj,n}(x)\right)\right| = o^*_p(1),$$

where $o^*_p(1)$ denotes convergence to zero according to the bootstrap law, P^*, conditional on the sample.

As in our above discussion, when constructing bootstrap counterparts for the statistics defined in (2.8) and (2.9) on both the positive and negatives supports of X , it suffices to utilize the *sgn* function, and note that $sgn(x)^2 = 1$. For example, replace $\eta^*_{1,t}(x)$ with $\eta^*_{1,t}(x) = \left[(e^*_{1,t} - x)sgn(x)\right]_+ - \frac{1}{n}\sum_{t=1}^{n}\left[(e_{1,t} - x)sgn(x)\right]_+$, replace $\eta^*_{j,t}(x)$ with $\eta^*_{j,t}(x) = \left[(e^*_{j,t} - x)sgn(x)\right]_+ - \frac{1}{n}\sum_{t=1}^{n}\left[(e_{j,t} - x)sgn(x)\right]_+$, and define

$$\widehat{\sigma}^{2*G}_{jj',n}(x) = \frac{1}{b_n}\sum_{k=1}^{b_n}\left(\frac{1}{l_n^{1/2}}\sum_{i=1}^{l_n}\left(u^*_{j,(k-1)l_n+i}(x) - u^*_{1,(k-1)l_n+i}(x)\right)\left(u^*_{j',(k-1)l_n+i}(x) - u^*_{1,(k-1)l_n+i}(x)\right)\right)$$

and

$$\widehat{\sigma}^{2*C}_{jj',n}(x) = \frac{1}{b_n}\sum_{k=1}^{b_n}\left(\frac{1}{l_n^{1/2}}\sum_{i=1}^{l_n}\left(\eta^*_{j,(k-1)l_n+i}(x) - \eta^*_{1,(k-1)l_n+i}(x)\right)\left(\eta^*_{j',(k-1)l_n+i}(x) - \eta^*_{1,(k-1)l_n+i}(x)\right)\right).$$

3.2 Inference Using the Bootstrap and Bounding Limiting Distributions

The statistics S_n^{G+} and S_n^{C+} are highly discontinuous over x. Exactly which moment conditions, and how many of them are binding varies over x. Hence, S_n^{G+} and S_n^{C+} do not necessarily have a well defined limiting distribution; and the continuous mapping theorem cannot be applied. However, following the

generalized moment selection (GMS) test approach of Andrews and Shi (2013) we can establish lower and upper bound limiting distributions. Let

$$D^{G+}(x) = \text{diag}\Sigma^{G+}(x,x),$$

$$h_{A,n}^{G+}(x) = D^{G+}(x)^{-1/2}\left(\sqrt{n}G_2^+(x),...,\sqrt{n}G_k^+(x)\right)', \tag{3.2}$$

$$h_B^{G+}(x,x') = D^{G+}(x)^{-1/2}\left(\Sigma^{G+} + \varepsilon I_{k-1}\right)(x,x')\, D^{G+}(x')^{-1/2}, \tag{3.3}$$

and

$$v^{G+}(.) = (v_2^{G+}(.),...,v_k^{G+}(.))', \tag{3.4}$$

where $v^{G+}(.)$ is a $(k-1)-$dimensional zero mean Gaussian process with correlation $h_B^{G+}(x,x')$. Also, let $D^{C+}(x), h_{A,n}^{C+}(x), h_B^{C+}(x,x'), v^{C+}(.)$ be defined analogously, by replacing $\Sigma^{G+}(x,x), G_2^+(x),...,G_k^+(x)$ with $\Sigma^{C+}(x,x), C_2^+(x),...,C_k^+(x)$. Finally, define

$$S_n^{\dagger G+} = \int_{\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{0, \frac{v_j^{G+}(x) + h_{j,A,n}^{G+}(x)}{\sqrt{h_{jj,B}^{G+}(x)}}\right\}\right)^2 \mathrm{d}Q(x) \tag{3.5}$$

where $h_{jj,B}^{G+}(x)$ is the $jj-$th element of $h_B^{G+}(x,x)$, and let

$$S_\infty^{\dagger G+} = \int_{\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{0, \frac{v_j^{G+}(x) + h_{j,A,\infty}^{G+}(x)}{\sqrt{h_{jj,B}^{G+}(x)}}\right\}\right)^2 \mathrm{d}Q(x), \tag{3.6}$$

where $h_{j,A,\infty}^{G+}(x) = 0$, if $G_j(x) = 0$, and $h_{j,A,\infty}^{G+}(x) = -\infty$, if $G_j(x) < 0$. Also, define $S_n^{\dagger C+}$ and $S_\infty^{\dagger C+}$ analogously, by replacing $v_j^{G+}(x), h_{j,A,n}^{G+}(x), h_{j,A,\infty}^{G+}(x)$, and $h_{jj,B}^{G+}(x)$ with $v_j^{C+}(x), h_{j,A,n}^{C+}(x), h_{j,A,\infty}^{C+}(x)$, and $h_{jj,B}^{C+}(x)$. Hereafter let

$$\mathcal{P}_0^{G+} = \left\{P:\ H_0^{G+} \text{ holds}\right\}$$

so that $\mathcal{P}_0^{G+}$ is the collection of DGPs under which the null hypothesis holds. Let $\mathcal{P}_0^{C+}$ be defined analogously, with H_0^{G+} replaced by H_0^{C+}. The following result holds.

Theorem 1: *Let Assumptions A1-A4 hold. Then:*

(i) under H_0^{G+}, there exists a $\delta > 0$ such that

$$\lim\sup_{n\to\infty} \sup_{P\in\mathcal{P}_0^{G+}} \left[P\left(S_n^{G+} > a_{h_{A,n}}^{G+}\right) - P\left(S_n^{\dagger G+} + \delta > a_{h_{A,n}}^{G+}\right)\right] \leq 0$$

and

$$\lim\inf_{n\to\infty} \inf_{P\in\mathcal{P}_0^{G+}} \left[P\left(S_n^{G+} > a_{h_{A,n}}^{G+}\right) - P\left(S_n^{\dagger G+} - \delta > a_{h_{A,n}}^{G+}\right)\right] \geq 0;$$

and

(ii) under H_0^{C+}, there exists a $\delta > 0$ such that

$$\lim \sup_{n\to\infty} \sup_{P\in\mathcal{P}_0^{C+}} \left[P\left(S_n^{C+} > a_{h_{A,n}}^{C+}\right) - P\left(S_n^{\dagger C+} + \delta > a_{h_{A,n}}^{C+}\right)\right] \leq 0$$

and

$$\lim \inf_{n\to\infty} \inf_{P\in\mathcal{P}_0^{C+}} \left[P\left(S_n^{C+} > a_{h_{A,n}}^{C+}\right) - P\left(S_n^{\dagger C+} - \delta > a_{h_{A,n}}^{C+}\right)\right] \geq 0.$$

Theorem 1 provides upper and lower bounds for $P\left(S_n^{G+} > a_{h_{A,n}}^{G+}\right)$ and $P\left(S_n^{C+} > a_{h_{A,n}}^{C+}\right)$, uniformly, over the probabilities under H_0^{G+} and H_0^{C+}, respectively. Note that $h_{j,A,n}^{G+}(\cdot)$ and $h_{j,A,n}^{C+}(\cdot)$ depend on the degree of slackness, and do not need to converge. Indeed, S_n^{G+} and/or S_n^{C+} do not have to converge in distribution for this result to hold.

Following Andrews and Shi (2013), we can construct bootstrap critical values which properly mimic the critical values of $S_\infty^{\dagger G+}$ and $S_\infty^{\dagger C+}$. We rely on the block bootstrap to capture the dependence in the data when constructing our bootstrap statistics. Consider the case of $S_\infty^{\dagger G+}$. Let $\left(e_{j,1}^*, ..., e_{j,n}^*\right), b_n$, and l_n be defined as in the previous subsection, and let:

$$G_{j,n}^{*+}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(1\left\{e_{j,i}^* \leq x\right\} - 1\left\{e_{1,i}^* \leq x\right\}\right) \tag{3.7}$$

and

$$v_n^{*G+}(x) = \sqrt{n}\widehat{D}_n^{-1/2,G+}(x)\left(G_n^{*+}(x) - G_n^{+}(x)\right) \tag{3.8}$$

with $v_n^{*G+}(x) = \left(v_{2,n}^{*G+}(x), ..., v_{k,n}^{*G+}(x)\right)$ and $\widehat{D}_n^{G+}(x) = \text{diag}\widehat{\Sigma}_n^{G+}(x,x)$. Then, define:

$$\xi_{j,n}^{G+}(x) = \kappa_n^{-1} n^{1/2} \overline{D}_{jj,n}^{-1/2,G+}(x) G_{j,n}^{+}(x), \tag{3.9}$$

with $\kappa_n \to \infty$, as $n \to \infty$. Here, $\overline{D}_{jj,n}^{G+}(x)$ is the jj-th element of $\overline{D}_n^{G+}(x) = \text{diag}\left(\overline{\Sigma}_n^{G+}(x,x)\right)$, $\xi_n^{G+}(x) = \left(\xi_{2,n}^{G+}(x), ..., \xi_{k,n}^{G+}(x)\right)$, and

$$\phi_{j,n}^{G+}(x) = c_n 1\left\{\xi_{j,n}^{G+}(x) < -1\right\}, \tag{3.10}$$

with c_n a positive sequence, which is bounded away from zero. Thus, $\phi_{j,n}^{G+}(x) = c_n$, when $G_{j,n}^{+}(x) < -\kappa_n n^{-1/2}\overline{D}_{jj,n}^{1/2,G+}(x)$ (i.e., when the $j-$th inequality is slack at x), and is zero otherwise.

It it clear from the selection rule in (3.10), that we do need an estimator of the variance of the moment conditions, despite the fact we use bootstrap critical values. In fact, standardization does not play a crucial role in the statistics, as all positive sample moment conditions matter. On the other hand, without the scaling factor in (3.9), the number of non-slack moment conditions would depend on the scale, and hence our bootstrap critical values would no longer be scale invariant. Let

$$S_n^{*G+} = \int_{\mathcal{X}^+} \sum_{j=2}^{k} \max\left(\left\{0, \frac{v_{j,n}^{*G+}(x) - \phi_{j,n}^{G+}(x)}{\sqrt{\overline{h}_{B,jj}^{*G+}(x)}}\right\}\right)^2 \text{d}Q(x), \tag{3.11}$$

where $\overline{h}_{B,jj}^{*G+}(x)$ is the $jj-th$ element of $\overline{D}_n^{-1/2,G+}(x)\overline{\Sigma}_n^{*G+}(x,x)\,\overline{D}_n^{-1/2,G+}(x)$ and $\overline{\Sigma}_n^{*G+}(x,x)$ is the bootstrap analog of $\overline{\Sigma}_n^{G+}(x,x)$.[6] Note that if c_n grows with n, then all slack inequalities are discarded, asymptotically. It is immediate to see that S_n^{*G+} is the bootstrap counterpart of $S_n^{\dagger G+}$ in (3.5), with $\phi_{j,n}^{G+}(x)$ mimicking the contribution of the slackness of inequality j (i.e., of $j-$th element of $h_{A,n}^{G+}(x)$). However, $\phi_{j,n}^{G+}(x)$ is not a consistent estimator of $h_{A,n}^{G+}(x)$, since the latter cannot be consistently estimated. Now, consider the case of $S_\infty^{\dagger C+}$. Let:

$$C_{j,n}^{*+}(x)=\frac{1}{n}\sum_{i=1}^{n}\left(\left[e_{j,t}^{*}-x\right]_{+}-\left[e_{1,t}^{*}-x\right]_{+}\right),$$

and define $v_n^{*C+}(x),\widehat{D}_n^{C+}(x),\xi_{j,n}^{C+}(x)$, and $\phi_{j,n}^{C+}(x)$ analogously to $v_n^{*G+}(x),\widehat{D}_n^{G+}(x),\xi_{j,n}^{G+}(x)$, and $\phi_{j,n}^{G+}(x)$, by replacing $G_n^{*+}(x),G_n^{+}(x)$ and $\widehat{\Sigma}_n^{G+}(x,x)$ with $C_n^{*+}(x),C_n^{+}(x)$ and $\widehat{\Sigma}_n^{C+}(x,x)$. Then, construct:

$$S_n^{*C+}=\int_{\mathcal{X}^+}\sum_{j=2}^{k}\max\left(\left\{0,\frac{v_{j,n}^{*C+}(x)-\phi_{j,n}^{C+}(x)}{\sqrt{\overline{h}_{B,jj}^{*C+}(x)}}\right\}\right)^2 \mathrm{d}Q(x). \tag{3.12}$$

By comparing (2.8) and (2.9) with (3.11) and (3.12), it is immediate to see that $G_{j,n}^{+}(x)$ ($C_{j,n}^{+}(x)$) does not contribute to the test statistic when $G_{j,n}^{+}(x)<0$, ($C_{j,n}^{+}(x)<0$) while it does not contribute to the bootstrap statistic when $G_{j,n}^{+}(x)<-\kappa_n n^{-1/2}\overline{D}_{jj,n}^{1/2,G+}(x)$ ($C_{j,n}^{+}(x)<-\kappa_n n^{-1/2}\overline{D}_{jj,n}^{1/2,C+}(x)$) with $\kappa_n n^{-1/2}\to 0$. Heuristically, by letting κ_n grow with the sample size, we control the rejection rates in a uniform manner.

It remains to define the GMS bootstrap critical values. Let $c_{n,B,1-\alpha}^{*G+}\left(\phi_n^{G+},\overline{h}_{B,n}^{*G+}\right)$ be the $(1-\alpha)$-th critical value of S_n^{*G+}, based on B bootstrap replications, with ϕ_n^{G+} defined as in (3.10) and $\overline{h}_{B,n}^{*G+}(x)=\widehat{D}_n^{-1/2,G+}(x)\overline{\Sigma}_n^{*G+}(x,x)\,\widehat{D}_n^{-1/2,G+}(x)$. The $(1-\alpha)$-th GMS bootstrap critical value, $c_{0,n,1-\alpha}^{*G+}\left(\phi_n^{G+},\overline{h}_{B,n}^{*G+}\right)$, is defined as:

$$c_{0,n,1-\alpha}^{*G+}\left(\phi_n^{G+},\overline{h}_{B,n}^{*G+}\right)=\lim_{B\to\infty}c_{n,B,1-\alpha+\eta}^{*G+}\left(\phi_n^{G+},\overline{h}_{B,n}^{*G+}\right)+\eta,$$

for $\eta>0$, arbitrarily small. Further, $c_{n,B,1-\alpha}^{*C+}\left(\phi_n^{C+},\overline{h}_{B,n}^{*C+}\right)$ and $c_{0,n,1-\alpha}^{*C+}\left(\phi_n^{C+},\overline{h}_{B,n}^{C+}\right)$ are defined analogously.

Here, the constant η is used to guarantee uniformity over the infinite dimensional nuisance parameters, $h_{A,n}^{G+}(.),h_{A,n}^{C+}(.)$, uniformly on $x\in\mathcal{X}^+$, and is termed the infinitesimal uniformity factor by Andrews and Shi (2013). Heuristically, if all moment conditions are slack, then both the statistic and its bootstrap counterpart are zero, and by having $\eta>0$ though arbitrarily close to zero we control the asymptotic rejection rate.

Finally, let

$$\mathcal{B}^{G+}=\left\{x\in\mathcal{X}^+ \text{ s.t. } h_{A,j,\infty}^{G+}=0, \text{ for some } j=2,...,k\right\} \tag{3.13}$$

[6] Thus, the diagonal elements of $\widehat{\Sigma}_n^{*G+}(x,x)$ are the $\widehat{\sigma}_{jj,n}^{2*G+}(x)$ described in the previous subsection, while the off-diagonal elements of $\widehat{\Sigma}_n^{*G+}(x,x)$ are defined accordingly, as $\widehat{\sigma}_{jj',n}^{2*G+}(x)$, with $j\neq j'$.

and

$$\mathcal{B}^{C+} = \left\{ x \in \mathcal{X}^{+} \text{ s.t. } h_{A,j,\infty}^{C+} = 0, \text{ for some } j = 2, ..., k \right\}, \tag{3.14}$$

where $\mathcal{B}^{G+}$ and $\mathcal{B}^{C+}$ define the sets over which at least one moment condition holds with strict equality, and these sets represent the boundaries of H_0^{G+} and H_0^{C+}, respectively.

Although we require that the block length grows at the same rate as the lag truncation parameter, l_n, in Lemma 2 (i.e., we require that $l_n \approx n^{\delta}$ $0 < \delta < \frac{1}{2}$ with δ being the mixing coefficient in A1), for the asymptotic uniform validity of the bootstrap critical values, we require that the block length grows at a rate slower than $n^{1/3}$. This slower rate is required for the bootstrap empirical central limit theorem for a mixing process to hold (see Peligrad (1998)). Needless to say, even in the construction of $\widehat{\sigma}_{jj,n}^{2,G+}(x)$, we should thus use $l_n = o(n^{1/3})$. The following result holds.

Theorem 2: *Let Assumptions A1-A4 hold, and let* $l_n \to \infty$ *and* $l_n n^{\frac{1}{3}-\varepsilon} \to 0$ *as* $n \to \infty$. *Under* H_0^{G+} :

(i) if as $n \to \infty$, $\kappa_n \to \infty$ *and* $c_n/\kappa_n \to 0$, *then*

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_0^{G+}} P\left(S_n^{G+} \geq c_{0,n,1-\alpha}^{*G+}\left(\phi_n^{G+}, \overline{h}_{B,n}^{*G+} \right) \right) \leq \alpha;$$

and

(ii) if as $n \to \infty$, $\kappa_n \to \infty$, $c_n \to \infty$, $\sqrt{n}/\kappa_n \to \infty$, *and* $Q\left(\mathcal{B}^{G+}\right) > 0$, *then*

$$\lim_{\eta\to 0} \limsup_{n\to\infty} \sup_{P \in \mathcal{P}_0^{G+}} P\left(S_n^{G+} \geq c_{0,n,1-\alpha}^{*G+}\left(\phi_n^{G+}, \overline{h}_{B,n}^{*G+} \right) \right) = \alpha.$$

Also, under H_0^{C+} :

(iii) if as $n \to \infty$, $\kappa_n \to \infty$ *and* $c_n/\kappa_n \to 0$, *then*

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_0^{C+}} P\left(S_n^{C+} \geq c_{0,n,1-\alpha}^{*C+}\left(\phi_n^{C+}, \overline{h}_{B,n}^{*C+} \right) \right) \leq \alpha;$$

and (iv) if as $n \to \infty$, $\kappa_n \to \infty$, $c_n \to \infty$, $\sqrt{n}/\kappa_n \to \infty$, *and* $Q\left(\mathcal{B}^{C+}\right) > 0$, *then*

$$\lim_{\eta\to 0} \limsup_{n\to\infty} \sup_{P \in \mathcal{P}_0^{C+}} P\left(S_n^{C+} \geq c_{0,n,1-\alpha}^{*C+}\left(\phi_n^{C+}, \overline{h}_{B,n}^{*C+} \right) \right) = \alpha.$$

Statements (i) and (iii) of Theorem 2 establish that inference based on GMS bootstrap critical values are uniformly asymptotically valid. Statements (ii) and (iv) of the theorem establish that inference based on GMS bootstrap critical values is asymptotically non-conservative, whenever $Q\left(\mathcal{B}^{+}\right) > 0$ or $Q\left(\mathcal{B}^{C+}\right) > 0$ (i.e., whenever at least one moment condition holds with equality, over a set $x \in \mathcal{X}^{+}$ with non-zero $Q-$measure). Although the GMS based tests are not similar on the boundary, the degree of non similarity, which is

$$\begin{aligned}
&\lim_{\eta\to 0} \limsup_{n\to\infty} \sup_{P \in \mathcal{P}_0^{G+}} P\left(S_n^{G+} \geq c_{0,n,1-\alpha}^{*G+}\left(\phi_n^{G+}, \overline{h}_{B,n}^{*G+} \right) \right) \\
&- \lim_{\eta\to 0} \liminf_{P \in \mathcal{P}_0^{G+}} P\left(S_n^{G+} \geq c_{0,n,1-\alpha}^{*G+}\left(\phi_n^{G+}, \overline{h}_{B,n}^{*G+} \right) \right),
\end{aligned}$$

is much smaller than that associated with using the "usual" recentered bootstrap. In the case of pairwise comparison (i.e., $k = 2$), Theorem 2(ii) of Linton, Song and Whang (2010) establishes similarity of stochastic dominance tests on a subset of the boundary.

For implementation of the tests discussed in this paper, it thus follows that one can use Holm bounds as is done in JCS (2017), with modifications due to the presence of the constant η. Estimate bootstrap $p-$values $p^{G+}_{B,n,S_n^{G+}} = \frac{1}{B}\sum_{s=1}^{B} 1\left((S_n^{*G^+} + \eta) \geq S_n^{G^+}\right)$ and $p^{G-}_{B,n,S_n^{G-}} = \frac{1}{B}\sum_{s=1}^{B} 1\left((S_n^{*G^-} + \eta) \geq S_n^{G^-}\right)$. Estimate $p^{C+}_{B,n,S_n^{C+}}$ and $p^{C-}_{B,n,S_n^{C-}}$ in analogous fashion. Then, use the following rules (Holm (1979)):

Rule S_n^G: Reject H_0^G at level α, if $\min\left\{p^{G+}_{B,n,S_n^{G+}}, p^{G-}_{B,n,S_n^{G-}}\right\} \leq (\alpha - \eta)/2$.

Rule S_n^C: Reject H_0^C at level α, if $\min\left\{p^{C+}_{B,n,S_n^{C+}}, p^{C-}_{B,n,S_n^{C-}}\right\} \leq (\alpha - \eta)/2$.

3.3 Power against Fixed and Local Alternatives

As our statistics are weighted averages over $\mathcal{X}^+$, they have non-trivial power only if the null is violated over a subset of non zero $Q-$measure. This applies to both power against fixed alternative, as well as to power against $\sqrt{n}-$local alternatives. In particular, for power against fixed alternatives, we require the following assumption.

Assumption FA: (i) $Q(B_{FA}^{G+}) > 0$, where $B_{FA}^{G+} = \{x \in \mathcal{X}^+ : G_j(x) > 0 \text{ for some } j = 2, ..., k\}$.. (ii) $Q(B_{FA}^{C+}) > 0$ where $B_{FA}^{C+} = \{x \in \mathcal{X}^+ : C_j(x) > 0 \text{ for some } j = 2, ..., k\}$.

The following result holds.

Theorem 3: *Let Assumptions A0-A4 hold.*

(i) If Assumption FA(i) holds, then under H_A^{G+} :

$$\lim_{n\to\infty} P\left(S_n^{G+} \geq c^{*G+}_{0,n,1-\alpha}\left(\phi_n^{G+}, \overline{h}^{*G+}_{B,n}\right)\right) = 1.$$

(ii) If Assumption FA(ii) holds, then under H_A^{C+} :

$$\lim_{n\to\infty} P\left(S_n^{C+} \geq c^{*C+}_{0,n,1-\alpha}\left(\phi_n^{C+}, \overline{h}^{*C+}_{B,n}\right)\right) = 1.$$

It is immediate to see that we have unit power against fixed alternatives, provided that the null hypothesis is violated, for at least one $j = 2, ..., k$, over a subset of $\mathcal{X}^+$ of non-zero $Q-$measure. Now, if we instead used a Kolmogorov type statistic (i.e., replace the integral over $\mathcal{X}^+$ with the supremum over $\mathcal{X}^+$), then we would not need Assumption FA, and it would suffice to have violation for some x, with possibly zero $Q-$measure, or in general with zero Lebesgue measure.[7] However, as pointed out in Supplement B of

[7] The Kolmorogov versions of S_n^{G+} and S_n^{C+} are:

$$KS_n^{G+} = \max_{x\in\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{0, \frac{\sqrt{n}G^+_{j,n}(x)}{\overline{\sigma}^{G+}_{jj,n}(x)}\right\}\right)^2$$

$$KS_n^{C+} = \max_{x\in\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{0, \frac{\sqrt{n}C^+_{j,n}(x)}{\overline{\sigma}^{G+}_{jj,n}(x)}\right\}\right)^2$$

Andrews and Shi (2013) the statements in parts (ii) and (iv) of Theorem 2 do not apply to Kolmogorov tests, and hence asymptotic non-conservativeness does not necessarily hold. This is because the proof of those statements use the bounded convergence theorem, which applies to integrals but not to suprema.

We now consider the following sequences of local alternatives:

$$H_{L,n}^{G+}: G_{Lj}^{+}(x) = G_j^{+}(x) + \frac{\delta_{1,j}(x)}{\sqrt{n}} + o\left(n^{-1/2}\right), \text{ for } j = 2, ..., k, \ x \in \mathcal{X}^{+}$$

and

$$H_{L,n}^{C+}: C_{Lj}^{+}(x) = C_j^{+}(x) + \frac{\delta_{2,j}(x)}{\sqrt{n}} + o\left(n^{-1/2}\right), \text{ for } j = 2, ..., k, \ x \in \mathcal{X}^{+}.$$

We have $\lim_{n\to\infty} \sqrt{n} D^{G+}(x)^{-1/2} G_{Lj}^{+}(x) \to h_{j,A,\infty}^{G+}(x) + \delta_{1,j}(x)$, and $\lim_{n\to\infty} \sqrt{n} D^{C+}(x)^{-1/2} C_{Lj}^{+}(x) \to h_{j,A,\infty}^{C+}(x) + \delta_{2,j}(x)$. Define,

$$S_{\infty,\delta_1,LG}^{\dagger,G+} = \int_{\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{ 0, \frac{v_j^{G+}(x) + h_{j,A,\infty}^{G+}(x) + \delta_{1,j}(x)}{\sqrt{h_{jj,B}^{G+}(x)}} \right\} \right)^2 \mathrm{d}Q(x)$$

and

$$S_{\infty,\delta_2,LC}^{\dagger,C+} = \int_{\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{ 0, \frac{v_j^{C+}(x) + h_{j,A,\infty}^{C+}(x) + \delta_{2,j}(x)}{\sqrt{h_{jj,B}^{C+}(x)}} \right\} \right)^2 \mathrm{d}Q(x)$$

We require the following assumption.

Assumption LA:(i) $Q(B_{LA}^{G+}) > 0$, where
$B_{LA}^{G+} = \left\{ x : \sqrt{n} D^{G+}(x)^{-1/2} G_{Lj}^{+}(x) \to h_{j,A,\infty}^{G+}(x) + \delta_{1,j}(x), \ 0 < h_{j,A,\infty}^{G+}(x) + \delta_{1,j}(x) < \infty, \text{ for some } j = 2, ..., k \right\}$.
(ii) $Q(B_{LA}^{C+}) > 0$, where
$B_{LA}^{C+} = \left\{ x : \sqrt{n} D^{C+}(x)^{-1/2} C_{Lj}^{+}(x) \to h_{j,A,\infty}^{C+}(x) + \delta_{2,j}(x), \ 0 < h_{j,A,\infty}^{C+}(x) + \delta_{2,j}(x) < \infty, \text{ for some } j = 2, ..., k \right\}$.

The following result holds.

Theorem 4: *Let Assumptions A1-A4 hold.*

(i) If Assumption LA(i) holds, then under $H_{L,n}^{G+}$:

$$\lim_{n\to\infty} P\left(S_n^{G+} \geq c_{0,n,1-\alpha}^{*G+}\left(\phi_n^{G+}, \overline{h}_{B,n}^{*G+} \right) \right) = P\left(S_{\infty,\delta_1,LG}^{\dagger G+} \geq c_{LG,1-\alpha}\left(h_{A,\infty}^{G+}, h_{B,\infty}^{G+} \right) \right),$$

with $c_{LG,1-\alpha}\left(h_{A,\infty}^{G+}, h_{B,\infty}^{G+} \right)$ *denoting the* $(1-\alpha)$*-th critical value of* $S_{\infty,\delta_1,LG}^{\dagger G+}$*, with* $0 < h_{j,A,\infty}^{G+}(x) + \delta_{1,j}(x) < \infty$*, for some* $j = 2, ..., k$.

(ii) If Assumption LA(ii) holds, then under $H_{L,n}^{C+}$:

$$\lim_{n\to\infty} P\left(S_n^{C+} \geq c_{0,n,1-\alpha}^{*C+}\left(\phi_n^{C+}, \overline{h}_{B,n}^{*C+} \right) \right) = P\left(S_{\infty,\delta_2,LC}^{\dagger C+} \geq c_{LC,1-\alpha}\left(h_{A,\infty}^{C+}, h_{B,\infty}^{C+} \right) \right),$$

with $c_{LC,1-\alpha}\left(h_{A,\infty}^{C+}, h_{B,\infty}^{C+} \right)$ *denoting the* $(1-\alpha)$*-th critical value of* $S_{\infty,\delta_2,LC}^{\dagger C+}$*, with* $0 < h_{j,A,\infty}^{C+}(x) + \delta_{2,j}(x) < \infty$*, for some* $j = 2, ..., k$.

Theorem 4 establishes that our tests have power against $\sqrt{n}-$alternatives, provided that the drifting sequence is bounded away from zero, over a subset of $\mathcal{X}^+$ of non-zero $Q-$measure. Note also that for given loss function, L, the sequence of local alternatives for the White reality check can be defined as:

$$H_{A,n} : \max_{j=2,\ldots,k} \left(E(L(e_1)) - E(L(e_j))\right) = \frac{\lambda}{\sqrt{n}} + o\left(n^{-1/2}\right),\ \lambda > 0. \tag{3.15}$$

For sake of simplicity, suppose that $k = 2$ (this is the well known Diebold and Mariano (1995) test framework). Here,

$$\begin{aligned}
& 0 < \lambda = n^{1/2} E(L(e_1)) - E(L(e_k)) + o(1) \\
= \;& n^{1/2} \int_{-\infty}^{\infty} L(x) \left(f_{1,n}(x) - f_{2,n}(x)\right) \mathrm{d}x \\
= \;& -n^{1/2} \int_{-\infty}^{0} L'(x) \left(F_{1,n}(x) - F_{2,n}(x)\right) \mathrm{d}x \\
& -n^{1/2} \int_{0}^{\infty} L'(x) \left(F_{1,n}(x) - F_{2,n}(x)\right) \mathrm{d}x \\
= \;& n^{1/2} \int_{-\infty}^{0} \left(h_{A,\infty}^{G-}(x) + \delta_1(x)\right) Q(x) \mathrm{d}x + n^{1/2} \int_{0}^{\infty} \left(h_{A,\infty}^{G+}(x) + \delta_1(x)\right) Q(x) \mathrm{d}x,
\end{aligned} \tag{3.16}$$

where $F_{j,n}(x) = F_j(x) + \frac{\delta_{1,j}(x)}{\sqrt{n}}$, and $\delta_1 = \delta_{1,1} - \delta_{1,2}$. Hence, $H_{A,n}$ in (3.15) is equivalent to $H_{LA}^{G+} \cap H_{LA}^{G-}$, whenever *Assumption A0 holds and* $Q(x) = L'(x)sgn(x)$.

Analogously, for any convex loss function, L, which satisfies Assumption A0, $H_{A,n}$ in (3.15) is equivalent to $H_{LA}^{C-} \cap H_{LA}^{C+-}$, whenever $Q(x) = L''(x)sgn(x)$. In fact, it is easy to see that:

$$\begin{aligned}
& 0 < \delta = n^{1/2} E(L(e_1)) - E(L(e_k)) + o(1) \\
= \;& n^{1/2} \int_{-\infty}^{\infty} L(x) \left(f_{1,n}(x) - f_{2,n}(x)\right) \mathrm{d}x \\
= \;& -n^{1/2} \int_{-\infty}^{0} L'(x) \left(F_{1,n}(x) - F_{2,n}(x)\right) \mathrm{d}x - n^{1/2} \int_{0}^{\infty} L'(x) \left(F_{1,n}(x) - F_{2,n}(x)\right) \mathrm{d}x \\
= \;& -L'(x) n^{1/2} \int_{-\infty}^{x} \left(F_{1,n}(z) - F_{2,n}(z)\right) \mathrm{d}z \,\big|_{-\infty}^{0} + n^{1/2} \int_{-\infty}^{0} L''(x) \left(\int_{-\infty}^{x} \left(F_{1,n}(z) - F_{2,n}(z)\right) \mathrm{d}z\right) \mathrm{d}x \\
& + n^{1/2} L'(x) \int_{x}^{\infty} \left(F_{1,n}(z) - F_{2.n}(z)\right) \mathrm{d}z \,\big|_{0}^{\infty} - n^{1/2} \int_{0}^{\infty} L''(x) \left(\int_{x}^{\infty} \left(F_{1,n}(z) - F_{2,n}(z)\right) \mathrm{d}z\right) \mathrm{d}x \\
= \;& n^{1/2} \int_{-\infty}^{0} L''(x) \left(\int_{-\infty}^{x} \left(F_{1,n}(z) - F_{2,n}(z)\right) \mathrm{d}z\right) \mathrm{d}x - n^{1/2} \int_{0}^{\infty} L''(x) \left(\int_{x}^{\infty} \left(F_{1,n}(z) - F_{2,n}(z)\right) \mathrm{d}z\right) \mathrm{d}x \\
= \;& n^{1/2} \int_{-\infty}^{0} \left(\int_{-\infty}^{x} \left(h_{A,\infty}^{C-}(x) + \delta_2(x)\right) \mathrm{d}z\right) Q(x) \mathrm{d}x - n^{1/2} \int_{0}^{\infty} \left(\int_{x}^{\infty} \left(h_{A,\infty}^{C+}(x) + \delta_2(x)\right) \mathrm{d}z\right) Q(x) \mathrm{d}x.
\end{aligned}$$

4 Monte Carlo Experiments

In this section, we evaluate the finite sample performance of GL and CL forecast superiority tests when there are multiple competing sequences of forecast errors, under stationarity. In addition to analyzing the performance of our tests based on S_n^{G+} and S_n^{G-}, (GL forecast superiority) as well as based on S_n^{C+} and S_n^{C-} (CL forecast superiority), we also analyze the performance of the related test statistics from JCS (2017), here called JCS_n^{G+}, JCS_n^{G-}, JCS_n^{C+}, and JCS_n^{C-}. For the sake of brevity, these two classes of tests are called S_n and JCS_n tests, respectively.[8] For each experiment we carry out 1000 Monte Carlo replications, and the number of bootstrap samples is $B = 500$. Additionally, four different values of the smoothing parameter, J_n are examined for the JCS_n tests, including $J_n = \{0.20, 0.35, 0.50, 0.60\}$; and four different values of the uniformity constant, η, are examined for the S_n tests, including $\eta = \{0.0015, 0.002, 0.0025, 0.003\}$.[9] Additionally, for S_n tests, when constructing $\overline{\Sigma}_n^{G+}$(as well as $\overline{\Sigma}_n^{G-}$, etc.), we set $l_n = \text{integer}[n^{0.2}]$ and $\varepsilon = 1e-4$. Finally, when implementing the bootstrap counterpart of S_n, we set $\kappa_n = \sqrt{0.3\log(n)}$ and $c_n = \sqrt{0.4\log(n)/\log(\log(n))}$, following Andrews and Shi (2013, 2017). Sample sizes of $n \in \{300, 600, 900\}$ are generated using each of the following eight data generating processes (DGPs), with independent forecast errors.

DGP1: $e_{1t} \sim i.i.d.N(0,1)$ and $e_{kt} \sim i.i.d.N(0,1)$, $k = 2, 3$.

DGP2: $e_{1t} \sim i.i.d.N(0,1)$ and $e_{kt} \sim i.i.d.N(0,1)$, $k = 2, 3, 4, 5$.

DGP3: $e_{1t} \sim i.i.d.N(0,1)$, $e_{kt} \sim i.i.d.N(0,1)$, $k = 2, 3$ and $e_{kt} \sim i.i.d.N(0, 1.4^2)$, $k = 4, 5$

DGP4: $e_{1t} \sim i.i.d.N(0,1)$, $e_{kt} \sim i.i.d.N(0,1)$, $k = 2, 3$ and $e_{kt} \sim i.i.d.N(0, 1.6^2)$, $k = 4, 5$

DGP5: $e_{1t} \sim i.i.d.N(0,1)$, $e_{kt} \sim i.i.d.N(0, 0.8^2)$, $k = 2, 3$ and $e_{kt} \sim i.i.d.N(0, 1.2^2)$, $k = 4, 5$.

DGP6: $e_{1t} \sim i.i.d.N(0,1)$, $e_{kt} \sim i.i.d.N(0, 0.8^2)$, $k = 2, 3, 4, 5$ and $e_{kt} \sim i.i.d.N(0, 1.2^2)$, $k = 6, 7, 8, 9$.

DGP7: $e_{1t} \sim i.i.d.N(0,1)$, $e_{kt} \sim i.i.d.N(0,1)$, $k = 2, 3$ and $e_{kt} \sim i.i.d.N(0, 0.8^2)$, $k = 4, 5$.

DGP8: $e_{1t} \sim i.i.d.N(0,1)$, $e_{kt} \sim i.i.d.N(0,1)$, $k = 2, 3$ and $e_{kt} \sim i.i.d.N(0, 0.6^2)$, $k = 4, 5$.

DGP9: $e_{1t} \sim i.i.d.N(0,1)$ and $e_{kt} \sim i.i.d.N(0, 0.8^2)$, $k = 2, 3, 4, 5$.

DGP10: $e_{1t} \sim i.i.d.N(0,1)$ and $e_{kt} \sim i.i.d.N(0, 0.6^2)$, $k = 2, 3, 4, 5$.

Additionally, we conducted experiments using DGPs specified with autocorrelated errors. For the sake of brevity, these finding are reported in the supplemental online appendix. Denoting $\widetilde{e}_{i,t} = \varrho\widetilde{e}_{i,t-1} +$

[8] In the construction of the statistics

$$S_n^{G+} = \int_{x\in\mathcal{X}^+} \sum_{j=2}^{k} \left(\max\left\{0, \sqrt{n}\frac{G_{j,n}(x)}{\overline{\sigma}_{jj,n}^G(x)}\right\}\right)^2 dQ(x) \text{ and } S_n^{G-} = \int_{x\in\mathcal{X}^-} \sum_{j=2}^{k} \left(\max\left\{0, \sqrt{n}\frac{G_{j,n}(x)}{\overline{\sigma}_{jj,n}^G(x)}\right\}\right)^2 dQ(x).$$

we set $dQ(x) = \frac{1}{1.5n^{0.6}}$. Thus, $Q(\cdot)$ is still uniform. For inference using our tests, once B is determined, estimate bootstrap $p-$values, $p_{B,n,S_n}^{G+} = \frac{1}{B}\sum_{s=1}^{B} 1\left((S_n^{*G^+} + \eta) \geq S_n^{G^+}\right)$. and $p_{B,n,S_n}^{G-} = \frac{1}{B}\sum_{s=1}^{B} 1\left((S_n^{*G-} + \eta) \geq S_n^{G-}\right)$. Then, use the following rules (Holm, (1979): Reject H_0^{TG} at level α, if $\min\left\{p_{B,n,S_n}^{G+}, p_{B,n,S_n}^{G-}\right\} \leq (\alpha - \eta)/2$. Reject H_0^{TC} at level α, if $\min\left\{p_{B,n,S_n}^{C+}, p_{B,n,S_n}^{C-}\right\} \leq (\alpha - \eta)/2$.

[9] In JCS (2017), the constant that we call J_n is called S_n.

$(1-\varrho^2)^{1/2}\eta_{i,t}$,with $\eta_{kt} \sim i.i.d.N(0,1)$ $i = 1, ..., 5$, the DGPs for these experiments are as follows.

DGP11: $e_{1t} = \widetilde{e}_{1,t}$ and $e_{kt} = \widetilde{e}_{k,t}$, $k = 2, 3, 4, 5$.

DGP12: $e_{1t} = \widetilde{e}_{1,t}$, $e_{kt} = \widetilde{e}_{k,t}$, $k = 2, 3$ and $e_{kt} = 1.4\widetilde{e}_{k,t}$, $k = 4, 5$.

DGP13: $e_{1t} = \widetilde{e}_{1,t}$, $e_{kt} = 0.8\widetilde{e}_{k,t}$, $k = 2, 3$ and $e_{kt} = 1.2\widetilde{e}_{k,t}$, $k = 4, 5$.

DGP14: $e_{1t} = \widetilde{e}_{1,t}$, $e_{kt} = \widetilde{e}_{k,t}$, $k = 2, 3$ and $e_{kt} = 0.6\widetilde{e}_{k,t}$, $k = 4, 5$.

In the above setup, DGPs 1-4 and DGPs 11-12 are used to conduct size experiments, while DGPs 4-10 and DGPs 13-14 are used to conduct power experiments. In all cases, e_{1t} denote the forecast errors from the benchmark model. Note that DGPs 1-2 correspond to the least favorable elements in the null, while in DGPs 3-4 and DGPs 11-12, some models underperform the benchmark. This is the case where we expect significant improvement when using our new tests instead of JCS tests. In DGPs 5-6 and DGP13, one half of the competing models outperform the benchmark model and the other half underperform. In DGPs 7-8, one half of the competing models outperform, while is DGPs 9-10 and DGP14, the competing models all outperform the benchmark model. The above DGPs are similar to those examined in JCS (2017), and are utilized in our experiments because they clearly illustrate the trade-offs associated with using JCS_n and S_n forecast superiority tests.

We now discuss the experimental findings gathered in Tables 1 and 2. All reported results are rejection frequencies based on carrying out the JCS_n and S_n tests using a nominal size equal to 0.1. Turning first to Table 1, note that results in this table are based on JCS_n tests. Summarizing, JCS_n tests have reasonably good size under DGPs 1-2 (the least favorable case under the null). However, they are often undersized (and in some cases severely so) in some sample size / J_n permutations when some models are worse than the benchmark (see DGPs 3-4), as should be expected given that the tests are not asymptotically correctly sized under these two DGPs. Moreover, in these cases the empirical size is non monotonic, in paricular for GL forecast superiority. Turning to Table 2, note that S_n tests, which are asymptotically non conservative, often exhibit better size properties under DGPs 3-4 (compare DGPs 3-4 in Tables 1 and 2) than JCS_n tests. For example, for the CL forecast superiority test the empirical size of the JCS_n test is 0.020 for all values of J_n, when $n = 900$ (see Table 1). The analogous value based on implementation of the S_n test is 0.083, for all values of η (see Table 2). Again, it is worth stressing that this finding comes as no surprise, given that the S_n test is asymptotically non conservative on the boundary of the null hypotheses, while the JCS_n test is conservative. Turning to power, note that the power of the JCS_n test is sometimes quite low relative to that of the S_n test. For example, under DGP7, power is 0.445 for the GL forecast superiority JCS_n test and 0.845 for the CL forecast superiority JCS_n test, when $n = 300$). Note that analogous rejection frequencies for the S_n tests are 0.870 and 0.923 (see Table 2, DGP7, $n = 300$). As expected, thus, S_n tests exhibit improved power relative to JCS_n tests, when some models are worse than the benchmark. All of the above findings pertain to the analysis of DGPS 1-10, in which forecast errors are serially uncorrelated. Results for DGPS 11-14, in which errors

are serially correlated are gathered in the supplemental appendix. Examination of the results for these DGPSs (see Tables Supplemental S1 and S2) are qualitatively the same as those reported on above. Finally, it should be pointed out that the S_n tests are not overly sensitive to the choice of η, and the empirical size of S_n tests appears "best" when η is very small, as should be expected. In conclusion, there is a clear performance improvement when comparing our new robust predictive superiority tests with JCS_n tests.[10]

5 Empirical Illustration: Robust Forecast Evaluation of SPF Expert Pools

In the real-time forecasting literature, predictions from econometric models are often compared with surveys of expert forecasters.[11] Such comparisons are important when assessing the implications associated with using econometric models in policy setting contexts, for example. One key survey dataset collecting expert predictions is the *Survey of Professional Forecasters* (SPF), which is maintained by the Philadelphia Federal Reserve Bank (see Croushore (1993)). This dataset, formerly known as the *American Statistical Association/National Bureau of Economic Research Economic Outlook Survey,* collects predictions on various key economic indicators (including, for example, nominal GDP growth, real GDP growth, prices, unemployment, and industrial production). For further discussion of the variables contained in the SPF, refer to Croushore (1993) and Aiolfi, Capistrán, and Timmermann (2011). The SPF has been examined in numerous papers. For example, Zarnowitz and Braun (1993) comprehensively study the SPF, and find, among other things, that use of the mean or median provides a consensus forecast with lower average errors than most individual forecasts. More recently, Aiolfi, Capistrán, and Timmermann (2011) consider combinations of SPF survey forecasts, and find that equal weighted averages of survey forecasts outperform model based forecasts, although in some cases these mean forecasts can be improved upon by averaging them with mean econometric model-based forecasts. When utilizing European data from the recently released ECB SPF, Genre, Kenny, Meyler, and Timmermann (2013) again find that it is very difficult to beat the simple average. This well known result pervades the macroeconometric forecasting literature, and reasons for the success of such simple forecast averaging

[10] For a discussion of simulation results based on application of the Diebold and Mariano (DM: 1995) test (in which specific loss functions are utilized) in our experimental setup, refer to JCS(2017). Summarizing from that paper, it is clear that when the loss function is unknown, there is an advantage to using our approach of testing for forecast superiority. However the DM test for pairwise comparison or a reality check test for multiple comparisons might yield improved power, for a given loss function. Indeed, under quadratic loss, JCS (2017) show that when the sample size is small, the DM test has better power performance than JCS_n type tests. When the sample size increases, the power difference between the two tests becomes smaller. This is as expected.

[11] See Fair and Shiller (1990), Swanson and White (1997a,b), Aiolfi, Capistrán and Timmermann (2011), and the references cited therein for further discussion.

are discussed in Timmermann (2006). He notes, among other things, that model misspecification related to instability (non-stationarities) and estimation error in situations where there are many models and relatively few observations may account to some degree for the success of simple forecast and model averaging. Our empirical illustration attempts to shed further light on the issue of simple model averaging and its importance in forecasting macroeconomic variables.

Our approach is to address the issue of forecast averaging and combination (called pooling) by viewing the problem through the lens of forecast superiority testing. Our use of loss function robust tests is unique to the SPF literature, to the best of our knowledge. Since we use robust forecast superiority tests, we do not evaluate pooling by using loss function specific tests, such as those discussed in Diebold and Mariano (1995), McCracken (2000), Corradi and Swanson (2003), and Clark and McCracken (2013). Additionally, our approach differs from that taken by Elliott, Timmermann, and Komunjer (2005, 2008), where the rationality of sequences of forecasts is evaluated by determining whether there exists a particular loss function under which the forecasts are rational. We instead evaluate predictive accuracy irrespective of the loss function implicitly used by the forecaster, and determine whether certain forecast combinations are superior when compared against any loss function, regardless of how the forecasts were constructed. In our tests, the benchmarks against which we compare our forecast combinations are simple average and median consensus forecasts. We aim to assess whether the well documented success of these benchmark combinations remains intact when they are compared against other combinations, under generic loss.[12]

In all of our experiments, we utilize SPF predictions of nominal GDP growth. The SPF is a quarterly survey, and the dataset is available at the Philadelphia Federal Reserve Bank (PFRB) website. The original survey began in 1968:Q4, and PFRB took control of it in 1990:Q2; but from that date, there are only around 100 quarterly observations prior to 2018:Q1, where we end our sample. In our analysis we thus use the entire dataset, which, after trimming to account for differing forecast horizons in our calculations, is 166 observations.[13],[14]

For our analysis, we consider 5 forecast horizons (i.e., $h = 0, 1, 2, 3, 4$). The reason we use $h = 0$ for one of the horizons is that the first horizon for which survey participants predict GDP growth is the quarter in which they are making their predictions. In light of this, forecasts made at $h = 0$ are called nowcasts. Moreover, it is worth noting that nowcasts are very important in policy making settings, since

[12] For an interesting discussion of machine learning and forecast combination methods, see Lahiri, Peng, and Zhao (2017); and for a discussion of probability forecasting and calibrated combining using the SPF, see Lahiri, Peng, and Zhao (2015). In these papers, various cases where consensus combinations do not "win" are discussed.

[13] It should be noted that the "timing" of the survey was not known with certainty prior to 1990. However, SPF documentation states that they believe, although are not sure, that the timing of the survey was similar before and after they took control of it.

[14] Note that the number of experts for which forecasts are recorded for each calendar date, was approximately 90 experts during each of the 4 quarters of 1968, while there where only approximately 40 experts in each quarter in 2017. For further details on the SPF dataset, refer to the documentation at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters.

first release GDP data are not available until around the middle of the subsequent quarter. The nominal GDP variable that we examine is called NGDP in the SPF. All test statistics are constructed using NGDP growth rate prediction errors. In particular, assume that one survey participant makes a forecast of NGDP, say $y_{t+h}^f|\mathcal{F}_t$.[15] The associated forecast error is:

$$e_t = \{\ln(y_{t+h}) - \ln(y_t)\} - \left\{\ln(y_{t+h}^f|\mathcal{F}_t) - \ln(y_t)\right\} = \ln(y_{t+h}) - \ln(y_{t+h}^f|\mathcal{F}_t),$$

where the actual NGDP value, y_{t+h}, is reported in the SPF, along with the NGDP predictions of each survey participant. Note that when $h = 0$, $\mathcal{F}_t$ does not include y_t. However, for $h > 0$, $\mathcal{F}_t$ includes y_t. As discussed previously, this is due to the release dates associated with the availability of NGDP data. Figure 1 illustrates some of the key properties of the NGDP data that we utilize. Namely, note that the distributions of the expert forecasts vary over time, and exhibit interesting skewness and kurtosis properties (compare Panels A-D of the figure, and the skewness and kurtosis statistics reported below the plots in the figure). Based on examination of the densities in Figure 1, one might wonder whether "trimming" experts from the panel, say those experts that provided the forecasts appearing in the left tails of the distributions, might improve overall predictive accuracy of the panel. Although this question is not directly addressed in our analysis, we do construct and analyze the performance of various "pools" formed by trimming experts that exhibit sub-par predictive accuracy, for example.

In addition to constructing $S_n^{G+}, S_n^{G-}, S_n^{C-}$, and S_n^{C+} tests in our empirical investigation, we also test for forecast superiority using the JCS_n tests discussed above, which have correct size only under the least favorable case under the null. In particular, we construct $JCS_n^{G+}, JCS_n^{G-}, JCS_n^{C-}$, and JCS_n^{C+} test statistics (see Section 2 and JCS (2017) for further details). All test statistics are calculated using the same parameter values (for B, J_n, η, l_n, and ε) as used in our Monte Carlo experiments. However, results are only reported for $J_n = 0.20$ and $\eta = 0.002$, since our findings remain unchanged when other values of J_n and η from our Monte Carlo experiments are used.

Two different benchmark models are considered, including (i) the arithmetic mean prediction from all participants; and (ii) the median prediction from all participants. Additionally, a variety of alternative model "groups" are considered. In all alternative models, mean and median predictions are again formed, but this time using subsets of the total available panel of experts, chosen in a number of ways, as outlined below.

Group 1 - Experts Chosen Based on Experience: Three expert pools (i.e. three alternative models) consisting of experts with 1, 3, and 5 years of experience.

In all of the remaining groups of combinations, individuals are ranked according to average absolute forecast errors, as well as according to average squared forecast errors. Mean (or median) predictions from these groups are then compared with our benchmark combinations.

[15] Here, $\mathcal{F}_t$ denotes the information set available to the expert forecaster at the time their predictions are made.

Group 2 - Experts Chosen Based on Forecast Accuracy I: Three expert pools consisting of most accurate expert over last 1, 3, and 5 years.
Group 3 - Experts Chosen Based on Forecast Accuracy II: Three expert pools consisting of most accurate group of 3 experts over last 1, 3, and 5 years.
Group 4 - Experts Chosen Based on Forecast Accuracy III: Three expert pools consisting of top 10% most accurate group of experts over last 1, 3, and 5 years.
Group 5 - Experts Chosen Based on Forecast Accuracy III: Three expert pools consisting of top 25% most accurate group of experts over last 1, 3, and 5 years.

Finally, 3 additional groups which combine models from each of Groups 1-5 are analyzed. These include:
Group 6: Five expert pools, including one pool with experts that have 1 year of experience, and 4 additional pools, one from each of Groups 2-5, all defined over the last 1 year.
Group 7: Five expert pools, including one pool with experts that have 3 years of experience, and 4 additional pools, one from each of Groups 2-5, all defined over the last 3 years.
Group 8: Five expert pools, including one pool with experts that have 5 years of experience, and 4 additional pools, one from each of Groups 2-5, all defined over the last 5 years.

As an example of how testing is performed, note that when implementing the S_n^G test using *Group 1,* there are three alternative models. The same is true when implementing tests using *Groups 2-5.* For *Groups 6-8*, tests are implemented using 5 alternative models, where one alternative is taken from each of *Groups 1-5.* Summarizing, we consider: (i) two benchmark models, against which each group of alternatives is compared; (ii) alternative models that are based on either mean or median pooled forecasts for, *Groups 2-8*; (iii) forecast accuracy pools used in *Groups* 1-8 that are based on either average absolute forecast errors or average squared forecast errors; (iii) 5 forecast horizons.

We now discuss our empirical findings. In Tables 3-4, statistics are reported for all forecast superiority tests. Entries are S_n^G, S_n^C, JCS_n^G, and JCS_n^C test statistics reported for forecast horizons $h = 0, 1, 2, 3, 4$. More specifically, $S_n^G = S_n^{G+}$ if $p_{B,n,S_n^{G+}}^{G+} \leq p_{B,n,S_n^{G-}}^{G-}$; otherwise $S_n^G = S_n^{G-}$. The other statistics reported in the tables (i.e., S_n^C, JCS_n^G, and JCS_n^C) are defined analogously. Rejections of the null of no forecast superiority at a 10% level are denoted by a superscript *. In Table 3, the benchmark model is always the arithmetic mean prediction from all participants, and expert pool forecasts are also arithmetic means. Analogously, in Table 4 the benchmark is the median prediction from all participants, and expert pool forecasts are also medians. To understand the layout of the tables, turn to Table 3, and note that for *Group* 1, the 4 statistics defined above (i.e., S_n^G, S_n^C, JCS_n^G, and JCS_n^C) are given, for each forecast horizon, $h = 0, 1, 2, 3$, and 4. Superscripts denote rejection of the null hypothesis based on a particular test. For example, note that application of the JCS_n^G test in *Group* 2 yields a test rejection for horizons $h = 2$ and 4. Turning to the results summarized in the tables, a number of clear conclusions emerge.

First, the majority of test rejections occur for $h = 4$, as can be seen by inspection of the results in both Tables 3 and 4. In particular, note that for $h = 4$, there are 13 test rejections in Table 3 and 11 test rejections in Table 4, across *Groups* 1-8. On the other hand, for all other forecast horizons combined (i.e., $h = \{0, 1, 2, 3\}$), there are 11 test rejections in Table 3 and 8 test rejections in Table 4. This suggests that expert pools which are constructed by "trimming" the least effective experts are most useful for longer horizon forecasts. These findings make sense if one assumes that it is easier to make short term forecasts than long term forecasts. Namely, some experts are simply not "up to the task" when forecasting at longer horizons. Summarizing, our main finding indicates that simple average or median forecasts can be beaten, in cases where forecasts are more difficult to make (i.e., longer horizons). Second, "experience" as measured by the length of time an expert has taken part in the SPF is not a direct indicator of forecast superiority, since there are no rejections of our tests for *Group* 1, when either mean (see Table 3) or median (see Table 4) forecasts are used in our tests. This does not necessarily mean that experience does not matter, at least indirectly (notice that test rejections sometimes occur for *Groups* 6-8, where experience and accuracy traits are combined.[16] Finally, note that Tables S1 and S2 in the supplemental appendix report root mean square forecast errors (RMSFEs) from the benchmark and competing models utilized in our empirical analysis. In these tables, we see that in the majority of cases considered, combination forecasts that utilize the mean have lower RMSFEs than when the median is used for constructing combination forecasts. For example, when comparing the benchmark RMSFEs of *Group* 1 that are reported in Tables S1 and S2, RMSFEs associated with mean combination forecasts (see Table S1) are lower for $h = \{0, 2, 3, 4\}$ than the RMSFEs associated with median combination forecasts (see Table S2). This is interesting, given the clear asymmetry and long left tails associated with the distributions of expert forecasts exhibited in Figure 1, and suggests that outlier forecasts from "less accurate" experts are not overly influential when using measures of central tendency as ensemble forecasts.

Summarizing, we have direct evidence that judicious selection of pools of experts can lead to loss function robust forecast superiority. However, it should be stressed that in this illustration of the testing techniques developed in this paper, we do not consider various combination methods, including Bayesian model averaging, for example. Additionally, we only look at nominal GDP, although the SPF has various other variables in it. Extensions such as these are left to future research.

[16] To explore this finding in more detail, we also constructed additional tables that are closely related to Tables 3 and 4, except that in these tables, RMSFEs are reported for all of the models used in each test (see supplemental appendix, Tables S1 and S2). In these tables, we see that combining experience with prior predictive accuracy can lead to lower RMSFEs, relative to the case where the entire pool of experts is used. However, RMSFEs are even lower for various alternative models for which we only use prior predictive accuracy to select expert pools (compare RMSFEs for *Groups* 3-5 with those for *Groups* 6-8 in the supplemental tables).

6 Concluding Remarks

We develop uniformly valid forecast superiority tests that are asymptotically non conservative, and that are robust to the choice of loss function. Our tests are based on principles of stochastic dominance, which can be interpreted as tests for infinitely many moment inequalities. In light of this, we use tools from Andrews and Shi (2013, 2017) when developing our tests. The tests build on earlier work due to Jin, Corradi, and Swanson (2017), and are meant to provide a class of predictive accuracy tests that are not reliant on a choice of loss function, such as the Diebold and Mariano (1995) test discussed in McCracken (2000). In developing the new tests, we establish uniform convergence (over error support) of HAC variance estimators, and of their bootstrap counterparts. In a Supplement, we also extend the theory of generalized moment selection testing to allow for the presence of non-vanishing parameter estimation error. In a series of Monte Carlo experiments, we show that finite sample performance of our tests is quite good, and that the power of our tests dominates those proposed by JCS (2017). Additionally, we carry out an empirical analysis of the well known Survey of Professional Forecasters, and show that utilizing expert pools based on past forecast quality can lead to loss function robust forecast superiority, when compared with pools that include all survey participants. This finding is particularly prevalent for our longest forecast horizon (i.e., 1-year ahead).

7 Appendix

Proof of Lemma 1: (i) The proof is the same for all j. Thus, let $u_t(x) = (1\{e_{j,t} \leq x\} - F_j(x)) - (1\{e_{1,t} \leq x\} - F_1(x))$, and define

$$\widehat{\widehat{\sigma}}_n^{2,G+}(x) = \frac{1}{n}\sum_{t=1}^{n} u_t^2(x) + 2\frac{1}{n}\sum_{\tau=1}^{l_n} w_\tau u_t(x)u_{t-\tau}(x).$$

We first show that

$$\sup_{x\in\mathcal{X}^+}\left|\widehat{\widehat{\sigma}}_n^{2,G+}(x) - \sigma^{2,G+}(x)\right| = o_p(1),$$

and then we show that

$$\sup_{x\in\mathcal{X}^+}\left|\widehat{\widehat{\sigma}}_n^{2,G+}(x) - \widehat{\sigma}_n^{2,G+}(x)\right| = o_p(1). \tag{7.1}$$

Now,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}^+}\left|\widehat{\widehat{\sigma}}_n^{2,G+}(x) - \sigma^{2,G+}(x)\right| \\
\leq\ &\sup_{x\in\mathcal{X}^+}\left|\frac{1}{n}\sum_{t=1}^{n}\left(u_t^2(x) - \mathrm{E}\left(u_t^2(x)\right)\right) + 2\frac{1}{n}\sum_{\tau=1}^{l_n} w_\tau \sum_{t=1}^{n}\left(u_t(x)u_{t-\tau}(x) - \mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right| \\
&+ \sup_{x\in\mathcal{X}^+}\left|\left(\sigma^2(x) - \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(u_t^2(x)\right) + 2\frac{1}{n}\sum_{\tau=1}^{l_n} w_\tau \sum_{t=1}^{n}\mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right|.
\end{aligned} \tag{7.2}$$

We begin with the first term on the RHS of (7.2). First note that,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}^+}\left|\frac{1}{n}\sum_{t=1}^{n}\left(u_t^2(x) - \mathrm{E}\left(u_t^2(x)\right)\right) + 2\frac{1}{n}\sum_{\tau=1}^{l_n} w_\tau \sum_{t=1}^{n}\left(u_t(x)u_{t-\tau}(x) - \mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right| \\
\leq\ &\sup_{x\in\mathcal{X}^+} 2\sum_{\tau=0}^{l_n}\left|\frac{1}{n}\sum_{t=1}^{n}\left(u_t(x)u_{t-\tau}(x) - \mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right|.
\end{aligned}$$

Now,

$$\begin{aligned}
&\Pr\left(\sup_{x\in\mathcal{X}^+} 2\sum_{\tau=0}^{l_n}\left|\frac{1}{n}\sum_{t=1}^{n}\left(u_t(x)u_{t-\tau}(x) - \mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right| \geq \varepsilon\right) \\
\leq\ &2\sum_{\tau=0}^{l_n}\Pr\left(\sup_{x\in\mathcal{X}^+}\left|\frac{1}{n}\sum_{t=1}^{n}\left(u_t(x)u_{t-\tau}(x) - \mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right| \geq \frac{\varepsilon}{l_n}\right),
\end{aligned}$$

so that we need to show that,

$$\Pr\left(\sup_{x\in\mathcal{X}^+}\left|\frac{1}{n}\sum_{t=1}^{n}\left(u_t(x)u_{t-\tau}(x) - \mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right| \geq \frac{\varepsilon}{l_n}\right) < \frac{\delta}{l_n}.$$

Given Assumption A2, WLOG, we can set $\mathcal{X}^+ = [0, \Delta]$, so that it can be covered by a_n^{-1} balls S_j, $j = 1, ..., \Delta a_n^{-1}$, centered at S_j, with radius a_n. Then,

$$\begin{aligned}
& \sup_{x \in \mathcal{X}^+} \left| \frac{1}{n} \sum_{t=1}^{n} \left(u_t(x) u_{t-\tau}(x) - \mathrm{E}\left(u_t(x) u_{t-\tau}(x) \right) \right) \right| \\
\leq & \max_{j=1,..,\Delta a_n^{-1}} \left| \frac{1}{n} \sum_{t=1}^{n} \left(u_t(s_j) u_{t-\tau}(s_j) - \mathrm{E}\left(u_t(s_j) u_{t-\tau}(s_j) \right) \right) \right| \\
& + \max_{j=1,..,\Delta a_n^{-1}} \sup_{x \in S_j} 2 \left| \left(\frac{1}{n} \sum_{t=1}^{n} u_{t-\tau}(s_j) \left(u_t(x) - u_t(s_j) \right) \right) \right. \\
& \left. - \left(\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\left(u_{t-\tau}(s_j) \left(u_t(x) - u_t(s_j) \right) \right) \right) \right| \\
& + \text{smaller order} \\
= & \; I_n + II_n.
\end{aligned}$$

Now,

$$\begin{aligned}
II_n \leq & \max_{j=1,..,\Delta a_n^{-1}} \sup_{x \in S_j} \left| \frac{1}{n} \sum_{t=1}^{n} u_{t-\tau}(s_j) \left(u_t(x) - u_t(s_j) \right) \right| \\
& + \max_{j=1,..,\Delta a_n^{-1}} \sup_{x \in S_j} \left| \frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\left(u_{t-\tau}(s_j) \left(u_t(x) - u_t(s_j) \right) \right) \right| \\
= & \; II_n^A + II_n^B.
\end{aligned}$$

Given Assumption A1, noting that by Cauchy - Schwarz,

$$\begin{aligned}
& \max_{j=1,..,\Delta a_n^{-1}} \sup_{x \in S_j} \left| \frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\left(u_{t-\tau}(s_j) \left(u_t(s_j) - u_{t-\tau}(s_j) \right) \right) \right| \\
\leq & \max_{j=1,..,\Delta a_n^{-1}} \sup_{x \in S_j} \sqrt{\mathrm{E}\left(u_{t-\tau}(s_j) \right)^2} \max_{j=1,..,a_n^{-1}} \sup_{x \in S_j} \sqrt{\mathrm{E}\left(u_t(s_j) - u_t(x) \right)^2} \\
= & \; O\left(a_n^{1/2} \right),
\end{aligned}$$

for some constant C. Recalling given that $u_t(x) = (1\{e_{j,t} \leq x\} - F_j(x)) - (1\{e_{1,t} \leq x\} - F_1(x))$ and $u_t(s_j)$ stay between -1 and 1

$$\begin{aligned}
& \max_{j=1,..,\Delta a_n^{-1}} \sup_{x \in S_j} \left| \frac{1}{n} \sum_{t=1}^{n} u_{t-\tau}(s_j) \left(u_t(s_j) - u_t(x) \right) \right| \\
\leq \; & 2 \max_{j=1,..,\Delta a_n^{-1}} \sup_{x \in S_j} \frac{1}{n} \sum_{t=1}^{n} |u_t(s_j) - u_t(x)| \\
\leq \; & \frac{2}{n} \sum_{t=1}^{n} 1\{x - a_n \leq e_{1,t} \leq x + a_n\} + \frac{2}{n} \sum_{t=1}^{n} 1\{x - a_n \leq e_{j,t} \leq x + a_n\} \\
& + 2 \sup_{x \in \mathcal{X}^+} (f_1(x) + f_j(x)) \\
= \; & O_p(a_n) = o_p(a_n^{1/2})
\end{aligned}$$

Hence, by Chebyshev inequality

$$l_n \Pr\left(II_n > \frac{\varepsilon}{l_n} \right) = O\left(a_n l_n^3 \right) = o(1),$$

for $a_n = o\left(l_n^{-3}\right)$.

Now, consider I_n. By the Lemma on page 739 of Hansen (2008), setting $a_n = l_n^{-4}$, $m = \frac{\Delta n}{4 l_n^2}$, and $l_n = n^{\delta}$, with $\delta < 1/2$, and recalling that given Assumption A1, $\text{var}\left(\sum_{t=1}^{m} \left(u_t(s_j) u_{t-\tau}(s_j) - \text{E}\left(u_t(s_j) u_{t-\tau}(s_j) \right) \right)\right) \leq Cm$, it follows that for some constant C,

$$\begin{aligned}
& \Pr\left(\max_{j=1,..,a_n^{-1}} \left| \frac{1}{n} \sum_{t=1}^{n} \left(u_t(s_j) u_{t-\tau}(s_j) - \text{E}\left(u_t(s_j) u_{t-\tau}(s_j) \right) \right) \right| \geq \frac{\varepsilon}{l_n} \right) \\
\leq \; & a_n^{-1} \Pr\left(\left| \sum_{t=1}^{n} \left(u_t(s_j) u_{t-\tau}(s_j) - \text{E}\left(u_t(s_j) u_{t-\tau}(s_j) \right) \right) \right| \geq \frac{n\varepsilon}{l_n} \right) \\
\leq \; & 4 a_n^{-1} \left(\exp\left(- \frac{\frac{n^2}{l_n^2} \varepsilon^2}{64 C n + \frac{8}{3} \frac{\Delta n^2}{4 l_n^3}} \right) + \frac{16}{b} l_n^2 \left(\frac{4}{\Delta} \frac{n}{l_n^2} \right)^{-\beta} \right) \\
= \; & a_n^{-1} \exp\left(- \frac{1}{64 C \frac{n}{l_n^2} + \frac{8}{3} \frac{\Delta n^2}{4 l_n^3}} \right) + \frac{64}{b} a_n^{-1} l_n^2 l_n^{2\beta} n^{-\beta} \\
= \; & o(1) + O\left(n^{\delta(6+2\beta)} n^{-\beta} \right) \\
= \; & o(1) \text{ for } \beta > \frac{6\delta}{1 - 2\delta}.
\end{aligned}$$

We now consider the second term on the RHS of (7.2). Note that

$$\begin{aligned}
&\sup_{x\in\mathcal{X}^+}\left|\left(\sigma^{2,G+}(x)-\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(u_t^2(x)\right)+2\frac{1}{n}\sum_{\tau=1}^{l_n}w_\tau\sum_{t=1}^{n}\mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right)\right| \\
\leq\ & 2\sup_{x\in\mathcal{X}^+}\left|\frac{1}{n}\sum_{\tau=1}^{l_n}(1-w_\tau)\sum_{t=1}^{n}\mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right| \qquad (7.3)\\
&+2\sup_{x\in\mathcal{X}^+}\left|\frac{1}{n}\sum_{\tau=l_n+1}^{n}w_\tau\sum_{t=1}^{n}\mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right|.
\end{aligned}$$

The first term on the RHS of (7.3) is $o_p(1)$, by the same argument as that used in Theorem 2 of Newey and West (1987). Also, by Lemma 6.17 in White (1984), for $q>2$,

$$\mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\leq C\tau^{-\beta/2-1/q}\mathrm{var}\left(u_t(x)\right)^{1/2}\mathrm{E}\left\|u_t(x)\right\|^q$$

and

$$\begin{aligned}
&\sup_{x\in\mathcal{X}^+}\left|\frac{1}{n}\sum_{\tau=l_n+1}^{n}w_\tau\sum_{t=1}^{n}\mathrm{E}\left(u_t(x)u_{t-\tau}(x)\right)\right| \\
\leq\ & C\sup_{x\in\mathcal{X}^+}\mathrm{var}\left(u_t(x)\right)^{1/2}\mathrm{E}\left\|u_t(x)\right\|^q\sum_{\tau=l_n+1}^{n}\tau^{-\beta/2-1/q}=o(1),
\end{aligned}$$

as $\beta\delta>1$, given Assumption A1, and noting that q can be taken arbitrarily large because of the boundedness of $u_t(x)$.

Finally, by the same argument as that used in the proof of (7.2), for all j,

$$\sup_{x\in\mathcal{X}^+}\frac{1}{n}\sum_{t=1}^{n}\left(1\left\{e_{j,t}\leq x\right\}-F_j(x)\right)=o_p\left(l_n^{-1}\right).$$

The statement in (7.1) follows immediately.

(ii) By noting that,

$$\begin{aligned}
&\left[e_{j,t}-s_j\right]_+-\left[e_{j,t}-x\right]_+ \\
=\ & (x-s_j)1\{e_t\geq x\}+(x-s_j)\left(1\{e_t\geq x\}-1\{e_t\geq s_j\}\right) \\
&+(e_t-x)\left(1\{e_t\geq s_j\}-1\{e_t\geq x\}\right),
\end{aligned}$$

the statement follows by the same argument as that used in part (i) of the proof.

Proof of Lemma 2: For notational simplicity, we suppress the jj subscript. Also, we suppress the

superscripts C^+ and G^+, as the proof follows by analogous argument. Note that

$$\begin{aligned}
& \sup_{x\in\mathcal{X}^+}\left|\widehat{\sigma}_n^{*2}(x)-\mathrm{E}^*\left(\widehat{\sigma}_n^*(x)\right)\right| \\
\leq\ & \sup_{x\in\mathcal{X}^+}\frac{l_n}{b}\sum_{k=1}^{b}\left|\left(\frac{1}{l_n}\sum_{j=1}^{l_n}u^*_{(k-1)l_n+j}(x)\right)^2-\mathrm{E}^*\left(\left(\frac{1}{l_n}\sum_{j=1}^{l_n}u^*_{(k-1)l_n+j}(x)\right)^2\right)\right| \\
=\ & \sup_{x\in\mathcal{X}^+}\frac{l_n}{b}\sum_{k=1}^{b}\left|\frac{1}{l_n^2}\sum_{j=1}^{l_n}\sum_{i=1}^{l_n}u^*_{(k-1)l_n+j}(x)u^*_{(k-1)l_n+i}(x)-\mathrm{E}^*\left(u^*_{(k-1)l_n+j}(x)u^*_{(k-1)l_n+i}(x)\right)\right|
\end{aligned}$$

Now,

$$\begin{aligned}
& \Pr\left(\sup_{x\in\mathcal{X}^+}\frac{l_n}{b}\sum_{k=1}^{b}\left|\frac{1}{l_n^2}\sum_{j=1}^{l_n}\sum_{i=1}^{l_n}u^*_{(k-1)l_n+j}(x)u^*_{(k-1)l_n+i}(x)-\mathrm{E}^*\left(u^*_{(k-1)l_n+j}(x)u^*_{(k-1)l_n+i}(x)\right)\right|\geq\varepsilon_1 a_n\right) \\
\leq\ & l_n\Pr\left(\sup_{x\in\mathcal{X}^+}\frac{l_n}{b}\sum_{k=1}^{b}\left|\frac{1}{l_n^2}\sum_{j=1}^{l_n}\sum_{i=1}^{l_n}u^*_{(k-1)l_n+j}(x)u^*_{(k-1)l_n+i}(x)-\mathrm{E}^*\left(u^*_{(k-1)l_n+j}(x)u^*_{(k-1)l_n+i}(x)\right)\right|\geq\varepsilon_1\frac{a_n}{l_n}\right).
\end{aligned}$$

It suffices to show that, uniformly in k,

$$\Pr\left(\sup_{x\in\mathcal{X}^+}\left|\frac{1}{l_n^2}\sum_{j=1}^{l_n}\sum_{i=1}^{l_n}u^*_{(k-1)l+j}(x)u^*_{(k-1)l+i}(x)-\mathrm{E}^*\left(u^*_{(k-1)l+j}(x)u^*_{(k-1)l+i}(x)\right)\right|\geq\varepsilon_1\frac{a_n}{l_n}\right)<\frac{\delta}{l_n}.$$

This follows using the same "covering numbers" argument used in the proof of Lemma 1.

Proof of Theorem 1: We again suppress the superscripts G^+ and C^+, as the proof follows by the same argument. We need to show that the statement in Lemma A1 in the Supplement Appendix of Andrews and Shi (2013) holds. Then, the proof of the theorem will follow using the same arguments as those used in the proof of their Theorem 1, as the proof is the same for independent and dependent observations. In fact, our set-up differs from Andrews and Shi (2013) only because we have dependent observations, and because we scale the statistic by a Newey-West variance estimator. For the rest of the proof, our set-up is simpler as we can fix their θ_n at a given value, say zero. It suffices to show that:
(i) $v_n(.)\Rightarrow v(.)$, as a process indexed by $x\in\mathcal{X}^+$, where $v(.)$ is a zero-mean $k-1-$dimensional Gaussian process, with covariance kernel given $\Sigma(x,x')$.
(ii) $\sup_{x,x'\in\mathcal{X}^+}\left\|\overline{h}_{B,n}(x,x')-\overline{h}_B(x,x')\right\|=o_p(1)$.
Now, statement (ii) follows directly from Lemma 1. It remains to show that (i) holds. The key difference between the independent and the dependent cases is that in the former we can rely on the concept of manageability, while in the latter we cannot. Nevertheless, (i) follows if we can show that $v_n(.)$ satisfies an empirical process. Given A1-A3, this follows from Lemma A2 in Jin, Corradi and Swanson (2017).

Proof of Theorem 2: (i) For notational simplicity, we omit the superscript G^+. The proof of this theorem mirrors the proof of Theorem 2(a) in the Supplement of Andrews and Shi (2013). Let $c_0\left(h_{A,n},\alpha\right)$

be the α critical value of $S_n^{\dagger}$, as defined in (3.5). Given Theorem 1(i), it follows that for all $\delta > 0$,

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} P\left(S_n \geq c_0\left(h_{A,n}, \alpha\right) + \delta\right) \leq \alpha.$$

The statement follows if we can show that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} P\left(c^*_{0,n,\alpha}\left(\phi_n, \overline{h}^*_{B,n}\right) \leq c_{0,n,\alpha}\left(h_{A,n}, \overline{h}^*_{B,n}\right)\right) = 0, \tag{7.4}$$

with $c_{0,n,\alpha}\left(h_{A,n}, \overline{h}^*_{B,n}\right)$ defined as $c_0\left(h_{A,n}, \alpha\right)$; but with $\overline{h}^*_{B,n}$ an argument of this function rather than $h_B(x)$; and if we can show that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} P\left(c_{0,n,\alpha}\left(h_{A,n}, \overline{h}^*_{B,n}\right) \leq c_0\left(h_{A,n}, \alpha\right)\right) = 0. \tag{7.5}$$

For $c_n \to \infty$ and $c_n/\kappa_n \to 0$, $\tau_n \to \infty$ and $\tau_n/\kappa_n \to 0$,

$$\begin{aligned}
& \sup_{P\in\mathcal{P}_0} P\left(c^*_{0,n,\alpha}\left(\phi_n, \overline{h}^*_{B,n}\right) \leq c_{0,n,\alpha}\left(h_{A,n}, \overline{h}^*_{B,n}\right)\right) \\
\leq\ & \sup_{P\in\mathcal{P}_0} P\left(-\phi_{j,n}(x) \leq h_{A,j,n}(x), \text{ for some } x \in \mathcal{X}^+ \text{ and some } j = 2,..,k\right) \\
\leq\ & \sup_{P\in\mathcal{P}_0} P\left(\xi_{j,n}(x) < -1 \text{ AND } -c_n \leq h_{A,j,n}(x), \text{ for some } x \in \mathcal{X}^+ \text{ and } j = 2,..,k\right) \\
\leq\ & \sup_{P\in\mathcal{P}_0} P\Big(D(x)^{1/2}\overline{D}^{-1/2}_{jj,n}(x)\left(G_{j,n}(x) - G_j(x)\right) + D(x)^{1/2}\overline{D}^{-1/2}_{jj,n}(x)h_{j,A,n}(x) < -\kappa_n \\
& \text{AND } -c_n \leq h_{A,j,n}(x), \text{ for some } x \in \mathcal{X}^+ \text{ and } j = 2,..,k\Big) \\
\leq\ & \sup_{P\in\mathcal{P}_0} P\Big(-\tau_n + D(x)^{-1/2}\overline{D}^{-1/2}_{jj,n}(x)h_{j,A,n}(x) < -\kappa_n \\
& \text{AND } -c_n \leq h_{A,j,n}(x), \text{ for some } x \in \mathcal{X}^+ \text{ and } j = 2,..,k\Big) \\
& + \sup_{P\in\mathcal{P}_0} P\Big(D(x)^{1/2}\overline{D}^{-1/2}_{jj,n}(x)\left(G_{j,n}(x) - G_j(x)\right) < -\tau_n, \text{ for some } x \in \mathcal{X}^+ \text{ and } j = 2,..,k\Big) \\
\leq\ & \sup_{P\in\mathcal{P}_0} P\Big(-D(x)^{-1/2}\overline{D}^{-1/2}_{jj,n}(x)h_{j,A,n}(x) < -\kappa_n + c_n \\
& \text{AND } -c_n \leq h_{A,j,n}(x), \text{ for some } x \in \mathcal{X}^+ \text{ and } j = 2,..,k\Big) \\
=\ & o(1).
\end{aligned}$$

This establishes that (7.4) holds. Finally, (7.5) follows from Lemma 1 and Lemma 2.

(ii) Recall that $c^*_{0,n,1-\alpha}\left(\phi_n, h_{B,n}\right)$ is the $(1-\alpha)-$ percentile of S^*_n, as defined in (3.11); and define $c^{GMS}_{0,n,1-\alpha}\left(\phi_n, \overline{h}_{B,n}\right)$ to the $(1-\alpha)-$ percentile of S^{GMS}_n, where

$$S^{GMS}_n = \max_{x\in\mathcal{X}^+} \sum_{j=2}^{k} \max\left(\left\{0, \frac{\overline{v}_{j,n}(x) - \phi_{j,n}(x)}{\sqrt{\overline{h}_{B,jj}(x)}}\right\}\right)^2,$$

with $\overline{v}_n = (\overline{v}_{2,n}, ..., \overline{v}_{k,n})'$ is a $k-1$ dimensional Gaussian process, with mean zero and covariance $\overline{h}_{B,n}(x,x') = \widehat{D}_n^{-1/2}(x)\overline{\Sigma}(x,x')\widehat{D}_n^{-1/2}(x')$. Finally, let $v = (v_2, ..., v_k)'$ is a $k-1$ dimensional Gaussian process, with mean zero and covariance $\overline{h}_B(x,x') = D^{-1/2}(x)\overline{\Sigma}(x,x')D^{-1/2}(x')$.We first need to show that

$$c^*_{0,n,1-\alpha}(\phi_n, h_{B,n}) - c^{GMS}_{0,n,1-\alpha}\left(\phi_n, \overline{h}_{B,n}\right) = o_p(1), \tag{7.6}$$

and then to prove that the statement holds when replacing $c^*_{0,n,1-\alpha}(\phi_n, h_{B,n})$ with $c^{GMS}_{0,n,1-\alpha}\left(\phi_n, \overline{h}_{B,n}\right)$.

From Lemma 2, $\widehat{\Sigma}^*_n(x,x') - \widehat{\Sigma}_n(x,x') = o^*_p(1)$, and so $\overline{\Sigma}^*_n(x,x') - \overline{\Sigma}_n(x,x') = o^*_p(1)$. Then, by Theorem 2.3 in Peligrad (1998),

$$v^* \xrightarrow{d^*} v \text{ a.s.-}\omega,$$

where $v^* \xrightarrow{*} v$ denotes weak convergence, conditional on sample. As $\overline{v}_n \Longrightarrow v$, (7.6) follows.

Given Assumption A4, by Lemma B3 in the Supplement of Andrews and Shi (2013), the distribution of $S^\dagger_\infty$, as defined in (3.6), is continuous. It is also strictly increasing and its $(1-\alpha)-$quantile is strictly positive, for all $\alpha < 1/2$. The statement then follows by the same argument as that used in the proof of Theorem 2(b) in the Supplement of Andrews and Shi (2013).

(iii)-(iv) follow by the same arguments as those used in the proof of **(i)** and **(ii)**, respectively. In the case of $S_n^{G^+}$, we rely on the the stochastic equicontinuity of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(1\{e_{1,i} \le x\} - 1\{e_{1,i} \le u\}\right)$, as $|x-u| \to 0$. When considering $S_n^{C^+}$, we need to ensure the stochastic equicontinuity of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \left((e_{1,i}-x)_+ - (e_{1,i}-u)_+\right)$. Now,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \left((e_{1,i}-x)_+ - (e_{1,i}-u)_+\right) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^n (u-x)\,1\{e_{1,i} \ge u\} + \frac{1}{\sqrt{n}}\sum_{i=1}^n (e_{1,i}-u)_+ \left(1\{e_{1,i} \ge x\} - 1\{e_{1,i} \ge u\}\right),
\end{aligned}$$

which, given Assumption 2, is stochastically equicontinous, by the same argument as those used for $S_n^{G^+}$. Hence, Theorem 2.3 in Peligrad (1998) also holds in this case.

Proof of Theorem 3: **(i)** Without loss of generality, let $B^{G+}_{FA} = \{x \in \mathcal{X}^+ : G_2(x) > 0\}$, and note that for all $x \in B^{G+}_{FA}$, $\max\left\{0, \frac{\sqrt{n}G^+_{2,n}(x)}{\overline{\sigma}^{G+}_{22,n}(x)}\right\} = \frac{\sqrt{n}G^+_{2,n}(x)}{\overline{\sigma}^{G+}_{22,n}(x)}$. Thus,

$$\begin{aligned}
S_n^{G+} &= \int_{B^{G+}_{FA}} \sum_{j=2}^k \left(\max\left\{0, \frac{\sqrt{n}G^+_{j,n}(x)}{\overline{\sigma}^{G+}_{jj,n}(x)}\right\}\right)^2 \mathrm{d}Q(x) + \int_{\mathcal{X}^+ \setminus B^{G+}_{FA}} \sum_{j=2}^k \left(\max\left\{0, \frac{\sqrt{n}G^+_{j,n}(x)}{\overline{\sigma}^{G+}_{jj,n}(x)}\right\}\right)^2 \mathrm{d}Q(x) \\
&= \int_{B^{G+}_{FA}} \left(\frac{\sqrt{n}G^+_{2,n}(x)}{\overline{\sigma}^{G+}_{22,n}(x)}\right)^2 \mathrm{d}Q(x) + \int_{B^{G+}_{FA}} \sum_{j=3}^k \left(\max\left\{0, \frac{\sqrt{n}G^+_{j,n}(x)}{\overline{\sigma}^{G+}_{jj,n}(x)}\right\} - \left(\frac{\sqrt{n}G^+_{2,n}(x)}{\overline{\sigma}^{G+}_{22,n}(x)}\right)\right)^2 \mathrm{d}Q(x) \\
&\quad + \int_{\mathcal{X}^+ \setminus B^{G+}_{FA}} \sum_{j=2}^k \left(\max\left\{0, \frac{\sqrt{n}G^+_{j,n}(x)}{\overline{\sigma}^{G+}_{jj,n}(x)}\right\}\right)^2 \mathrm{d}Q(x) \\
&= I_n + II_n + III_n.
\end{aligned}$$

Now, I_n diverges to infinity with probability approaching one, while Theorem 1 ensures that II_n and III_n are $O_p(1)$. Thus, S_n^{G+} diverges to infinity. As S_n^{*G+} is $O_{p^*}(1)$, conditional on the sample, the statement follows.

(ii) Note that $S_n^{C^+}$ can be treated exactly as $S_n^{G^+}$.

Proof of Theorem 4:

(i) Define, $S_{\infty,LA}^{\dagger G+}$ as in (3.6), but with the vector $h_{j,A,\infty}^{G+}(x)$ having at least one component strictly bounded away above from zero, and finite, for all $x \in B_{LA}^{G+}$. Let $\mathcal{P}_{n,LA}^{G+}$ denote the set of probabilities under the sequence of local alternatives. We have that for all $a > 0$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_{n,LA}^{G+}} \left[P\left(S_n^{G+} > a\right) - P\left(S_{\infty,LA}^{\dagger G+} > a\right)\right] = 0,$$

and the distribution of $S_{\infty,LA}^{\dagger G+}$ is continuous at its $(1-\alpha)+\delta$ quintile, for all $0 < \alpha < 1/2$ and $\delta \geq 0$. Also, note that for all $x \in B_{LA}^{G+}$, $\phi_n^{G+} = 0$. The statement then follows by the same argument as that used in the proof of Theorem 2(ii). **(ii)** By the same argument as in part (i).

8 References

Aiolfi, M., C. Capistrán, and A. Timmermann (2011). Forecast Combinations. In M.P. Clements and D.F. Hendry (eds.), **Oxford Handbook of Economic Forecasting**, pp. 355-390, Oxford University Press, Oxford.

Andrews, D.W.K. (1991). Heteroskedasticity and Autocorrelation Robust Covariance Matrix Estimation. *Econometrica,* 59, 817-858.

Andrews, D.W.K. and D. Pollard (1994). An Introduction to Functional Central Limit Theorems for Dependent Stochastic Processes. *International Statistical Review,* 62, 119-132.

Andrews, D.W.K. and X. Shi (2013). Inference Based on Conditional Moment Inequalities. *Econometrica,* 81, 609-666.

Andrews, D.W.K. and X. Shi (2017). Inference Based on Many Conditional Moment Inequalities. *Journal of Econometrics,* 196, 275-287.

Barendse, S. and A.J. Patton (2019). Comparing Predictive Accuracy in the Presence of a Loss Function Shape Parameter. Working Paper, Duke University.

Bierens H.J. (1982). Consistent Model Specification Tests. *Journal of Econometrics,* 20, 105-134.

Bierens H.J. (1990). A Consistent Conditional Moment Tests for Functional Form. *Econometrica,* 58, 1443-1458.

Clark, T. and M. McCracken (2013). Advances in Forecast Evaluation. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), **Handbook of Economic Forecasting Vol. 2**, pp. 1107-1201, Elsevier, Amsterdam.

Corradi, V. and N.R. Swanson (2003). Predictive Density Evaluation. In G. Elliott, C.W.J. Granger and A. Timmermann (eds.), **Handbook of Economic Forecasting Vol. 1**, pp. 197-284, Elsevier, Amsterdam.

Corradi, V. and N.R. Swanson (2007). Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes. *International Economic Review,* 48, 67-109.

Corradi, V. and N. R. Swanson (2013). A Survey of Recent Advances in Forecast Accuracy Comparison Testing, with an Extension to Stochastic Dominance. In X. Chen and N.R. Swanson (eds.), **Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions, Essays in honor of Halbert L. White, Jr.**, pp. 121-144, Springer, New York.

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters, The Federal Reserve Bank of Philadelphia Business Review, November-December, 3-15.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13, 253-263.

Diebold, F.X. and M. Shin (2015). Assessing Point Forecast Accuracy by Stochastic Loss Distance. *Economics Letters,* 130, 37-38.

Diebold, F.X. and M. Shin (2017). Assessing Point Forecast Accuracy by Stochastic Error Distance. *Econometric Reviews,* 36, 588-598.

Elliott, G., I. Komunjer and A. Timmermann (2005). Estimation and Testing of Forecast Rationality under Flexible Loss. *Review of Economic Studies,* 72, 1107-1125.

Elliott, G., I. Komunjer and A. Timmermann (2008). Biases in Macroeconomic Forecasts: Irrationality of Asymmetric Loss? *Journal of the European Economic Association,* 6, 122-157.

Fair, R.C. and R.J. Shiller (1990). Comparing Information in Forecasts from Econometric Models. *American Economic Review*, 80, 375-389.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013). Combining the Forecasts in the ECB Survey of Professional Forecasters: Can Anything Beat the Simple Average. *International Journal of Forecasting*, 29, 108-121.

Gneiting, T. (2011). Making and Evaluating Point Forecast. *Journal of the American Statistical Association,* 106, 746-762.

Granger, C. W. J. (1999). Outline of Forecast Theory using Generalized Cost Functions. *Spanish Economic Review,* 1, 161-173.

Hansen, B.E. (2008). Uniform Convergence Rates for Kernel Estimators with Dependent Data. *Econometric Theory,* 24, 726-748.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics,* 6, 65-70.

Jin, S., V. Corradi and N.R. Swanson (2017). Robust Forecast Comparison. *Econometric Theory,* 33, 1306-1351.

Lahiri, K., H. Peng, and Y. Zhao (2015). Testing the Value of Probability Forecasts for Calibrated Combining. *International Journal of Forecasting*, 31, 113-129.

Lahiri, K., H. Peng, and Y. Zhao (2017). Online Learning and Forecast Combination in Unbalanced Panels. *Econometric Reviews,* 36, 257-288.

Linton, O., K. Song and Y.J. Whang (2010). An Improved Bootstrap Test of Stochastic Dominance. *Journal of Econometrics,* 154, 186-202.

McCracken, M.W. (2000). Robust Out-of-Sample Inference. *Journal of Econometrics*, 99, 195-223.

Newey, W.K. and West, K.D. (1987). A Simple, Positive Definite, Heteroskedasticty and Autocorrelation Consistent Covariance Matrix. *Econometrica,* 55, 703-708.

Patton, A.J. (2019). Comparing Possibly Misspecified Forecasts. *Journal of Business & Economic Statistics,* forthcoming.

Patton, A.J. and A. Timermann (2007). Testing Forecast Optimality under Unknown Loss. *Journal of the American Statistical Association,* v.102, 1172-1184.

Peligrad, M. (1998). On the Blockwise Bootstrap for Empirical Processes for Stationary Sequences. *Annals of Probability,* 26, 877-901.

Swanson, N.R. and H. White (1997a). A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks. *Review of Economics and Statistics*, 79, 1997, 540-550.

Swanson, N.R. and H. White (1997b). Forecasting Economic Time Series Using Adaptive Versus Nonadaptive and Linear Versus Nonlinear Econometric Models. *International Journal of Forecasting,* 13, 1997, 439-461.

Timmermann, A. (2006). Forecast Combinations. In A. Timmermann, C.W.J. Granger, and G. Elliott (eds.), **Handbook of Forecasting Vol. 1**, pp. 135-196**.** North Holland, Amsterdam.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica* 68, 1097-1126.

White, H. (1984). **Asymptotic Theory for Econometricians**. Academic Press, San Diego.

Zarnowitz, V. and P. Braun, (1993). Twenty-Two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance. In J.H. Stock and M.W. Watson (eds.), **Business Cycles, Indicators, and Forecasting**, pp. 11-94, University of Chicago Press, Chicago.

Table 1: Monte Carlo Results for JCS_n^{G+}, JCS_n^{G-}, JCS_n^{C+}, and JCS_n^{C-} Forecast Superiority Tests*

DGP	n	$J_n = 0.20$	$J_n = 0.35$	$J_n = 0.50$	$J_n = 0.65$	$J_n = 0.20$	$J_n = 0.35$	$J_n = 0.50$	$J_n = 0.65$
		GL Forecast Superiority				CL Forecast Superiority			
		Empirical Size							
DGP1	300	0.113	0.100	0.101	0.112	0.113	0.107	0.120	0.115
	600	0.105	0.110	0.099	0.108	0.091	0.089	0.094	0.091
	900	0.102	0.095	0.093	0.096	0.094	0.092	0.082	0.094
DGP2	300	0.110	0.105	0.115	0.113	0.097	0.097	0.092	0.106
	600	0.077	0.073	0.082	0.079	0.090	0.092	0.089	0.081
	900	0.085	0.084	0.086	0.095	0.089	0.101	0.097	0.092
DGP3	300	0.070	0.065	0.065	0.060	0.030	0.030	0.035	0.035
	600	0.050	0.040	0.050	0.045	0.030	0.020	0.030	0.025
	900	0.070	0.070	0.080	0.075	0.020	0.020	0.020	0.020
DGP4	300	0.065	0.070	0.060	0.065	0.010	0.015	0.020	0.015
	600	0.040	0.040	0.040	0.055	0.015	0.015	0.015	0.020
	900	0.070	0.065	0.065	0.065	0.015	0.015	0.015	0.015
		Empirical Power							
DGP5	300	0.496	0.485	0.490	0.477	0.753	0.759	0.745	0.759
	600	0.775	0.771	0.773	0.773	0.991	0.986	0.989	0.981
	900	0.943	0.951	0.947	0.938	1.000	1.000	1.000	1.000
DGP6	300	0.483	0.480	0.476	0.474	0.758	0.741	0.745	0.736
	600	0.768	0.778	0.772	0.774	0.984	0.975	0.981	0.980
	900	0.954	0.949	0.954	0.947	1.000	1.000	1.000	1.000
DGP7	300	0.490	0.475	0.475	0.445	0.875	0.865	0.845	0.855
	600	0.835	0.820	0.820	0.800	0.995	0.995	0.995	0.995
	900	0.975	0.970	0.970	0.965	1.000	1.000	1.000	1.000
DGP8	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DGP9	300	0.643	0.660	0.650	0.629	0.949	0.944	0.937	0.948
	600	0.913	0.885	0.890	0.896	1.000	1.000	1.000	1.000
	900	0.990	0.986	0.984	0.983	1.000	1.000	1.000	1.000
DGP10	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

* Notes: Entries denote rejection frequencies of (JCS_n^{G+},JCS_n^{G-}) tests (i.e., GL forecast superiority) and (JCS_n^{C+},JCS_n^{C-}) tests (i.e., CL forecast superiority) under a variety of data generating processes denoted by DGP1-DGP10. In DGP1-DGP4, no alternative outperforms the benchmark model. In DGP5-DGP10, at least one alternative model outperfroms the benchmark model. Sample sizes include n=300, 600, and 900 observations, as indicated in the second column of entries in the table. Nominal test size is 10%, and tests are carried out using critical values constructed for values of J_n including 0.20, 0.35, 0.50, and 0.65. See Section 4 for complete details.

Table 2: Monte Carlo Results for S_n^{G+}, S_n^{G-}, S_n^{C+}, and S_n^{C-} Forecast Superiority Tests*

DGP	n	$\eta = 0.0015$	$\eta = 0.002$	$\eta = 0.0025$	$\eta = 0.003$	$\eta = 0.0015$	$\eta = 0.002$	$\eta = 0.0025$	$\eta = 0.003$
		GL Forecast Superiority				CL Forecast Superiority			
		Empirical Size							
DGP1	300	0.078	0.078	0.076	0.076	0.095	0.094	0.094	0.091
	600	0.096	0.096	0.095	0.095	0.116	0.116	0.115	0.114
	900	0.120	0.119	0.118	0.117	0.096	0.096	0.095	0.095
DGP2	300	0.068	0.067	0.067	0.066	0.095	0.095	0.095	0.094
	600	0.097	0.096	0.095	0.095	0.096	0.096	0.095	0.094
	900	0.111	0.108	0.108	0.105	0.106	0.105	0.105	0.105
DGP3	300	0.021	0.021	0.021	0.021	0.038	0.038	0.038	0.038
	600	0.070	0.070	0.069	0.068	0.086	0.086	0.085	0.085
	900	0.071	0.070	0.069	0.069	0.083	0.083	0.083	0.082
DGP4	300	0.010	0.010	0.01	0.010	0.041	0.041	0.041	0.040
	600	0.064	0.064	0.064	0.063	0.077	0.077	0.076	0.076
	900	0.069	0.067	0.064	0.063	0.083	0.083	0.083	0.083
		Empirical Power							
DGP5	300	0.911	0.911	0.910	0.909	0.972	0.972	0.971	0.971
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DGP6	300	0.957	0.956	0.956	0.956	0.989	0.989	0.989	0.989
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DGP7	300	0.874	0.872	0.871	0.870	0.925	0.924	0.923	0.923
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DGP8	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DGP9	300	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DGP10	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	900	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

* Notes: Entries denote rejection frequencies of (S_n^{G+},S_n^{G-}) tests (i.e., GL forecast superiority) and (S_n^{C+},S_n^{C-}) tests (i.e., CL forecast superiority) under a variety of data generating processes denoted by DGP1-DGP10. In DGP1-DGP4, no alternative outperforms the benchmark model. In DGP5-DGP10, at least one alternative model outperfroms the benchmark model. Sample sizes include n=300, 600, and 900 observations, as indicated in the second column of entries in the table. Nominal test size is 10%, and tests are carried out using critical values constructed for values of η including 0.0015, 0.002, 0.0025, and 0.0030. See Section 4 for complete details.

Table 3: SPF Forecast Pooling Analysis of Quarterly Nominal GDP Using Mean Benchmark Model and Mean Expert Pool Predictions*

Group	*Statistic*	*Forecast Horizon*				
		$h=0$	$h=1$	$h=2$	$h=3$	$h=4$
Group 1	S_n^G	0.000718	0.000781	0.001284	0.001525	0.024494
	S_n^C	0.000008	0.000020	0.000203	0.000172	0.024470
	JCS_n^G	0.232845	0.232845	0.232845	0.310460	0.024611
	JCS_n^C	0.000357	0.000333	0.000729	0.000745	0.025276
Group 2	S_n^G	0.000820	0.002681	0.002334	0.002451	0.008070
	S_n^C	0.000037	0.001516	0.001983	0.002675	0.003287
	JCS_n^G	0.543305	0.698535	1.164226*	0.698535	1.397071*
	JCS_n^C	0.000115	0.006752	0.009945	0.011544	0.021668
Group 3	S_n^G	0.001975	0.004097	0.005302	0.004784	0.009391
	S_n^C	0.000878	0.001306	0.008277	0.004486	0.014821*
	JCS_n^G	0.465690	0.698535	0.776151	0.620920	1.164226*
	JCS_n^C	0.002955	0.005766	0.011527*	0.010482	0.020510*
Group 4	S_n^G	0.001779	0.005075	0.004359	0.003820	0.008772
	S_n^C	0.000512	0.001760	0.004896	0.004161	0.009259
	JCS_n^G	0.543305	0.853766	0.776151	0.620920	1.241841*
	JCS_n^C	0.002621	0.007716*	0.008856	0.010377	0.022399*
Group 5	S_n^G	0.001540	0.003063	0.005878	0.007226	0.012886*
	S_n^C	0.000507	0.000842	0.008293	0.012770*	0.020682*
	JCS_n^G	0.465690	0.776151	0.620920	0.853766	1.008996*
	JCS_n^C	0.002235	0.004643	0.010388*	0.015376*	0.018873*
Group 6	S_n^G	0.002384	0.007759	0.005390	0.004585	0.012976
	S_n^C	0.000369	0.002493	0.008548	0.007844	0.022130*
	JCS_n^G	0.776151	1.164226*	0.776151	0.698535	1.008996
	JCS_n^C	0.002619	0.008085*	0.009990	0.014823	0.020172
Group 7	S_n^G	0.002284	0.006819	0.008704	0.009044	0.008960
	S_n^C	0.000986	0.002106	0.011060	0.008702	0.007617
	JCS_n^G	0.465690	0.931381*	1.086611*	0.698535	0.931381
	JCS_n^C	0.002803	0.005994*	0.011401*	0.011645	0.012845
Group 8	S_n^G	0.001857	0.004237	0.003911	0.006975	0.016996
	S_n^C	0.000452	0.001066	0.004071	0.007224	0.015145
	JCS_n^G	0.931381	0.776151	0.698535	0.853766	1.397071*
	JCS_n^C	0.002526	0.004420	0.007946	0.008453	0.020293*

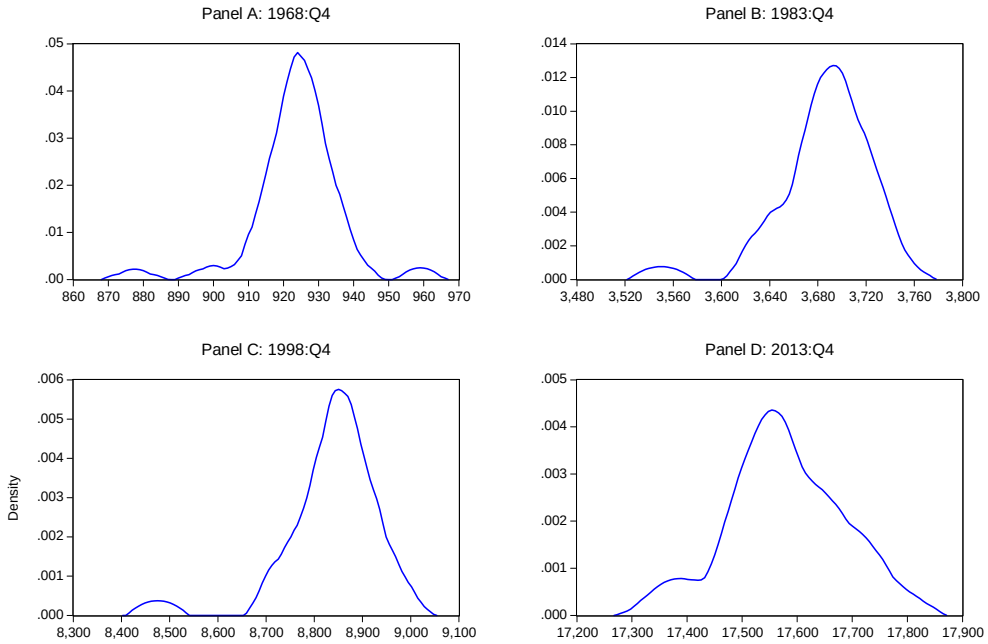
* Notes: Entries are S_n^G, S_n^C, JCS_n^G, and JCS_n^C test statistics reported for forecast horizons $h = 0, 1, 2, 3, 4$. More specifically, $S_n^G = S_n^{G+}$ *if* $p^{G+}_{B,n,S_n^{G+}} \leq p^{G-}_{B,n,S_n^{G-}}$; *otherwise* $S_n^G = S_n^{G-}$. S_n^C, JCS_n^G, and JCS_n^C are defined analogously. Rejections of the null of no forecast superiority at a 10% level are denoted by a superscipt *. See Section 5 for complete details.

Table 4: SPF Forecast Pooling Analysis of Quarterly Nominal GDP Using Median Benchmark Model and Median Expert Pool Predictions*

Group	*Statistic*	*Forecast Horizon*				
		$h=0$	$h=1$	$h=2$	$h=3$	$h=4$
Group 1	S_n^G	0.000563	0.001063	0.001381	0.002021	0.001943
	S_n^C	0.000006	0.000152	0.000556	0.001059	0.000194
	JCS_n^G	0.310460	0.232845	0.232845	0.232845	0.388075
	JCS_n^C	0.000178	0.000907	0.000989	0.002319	0.001603
Group 2	S_n^G	0.000715	0.002496	0.002127	0.002244	0.004061
	S_n^C	0.000055	0.001409	0.001548	0.001152	0.001612
	JCS_n^G	0.465690	0.698535	1.086611*	0.931381	1.008996
	JCS_n^C	0.000858	0.007005	0.009498	0.008086	0.012498
Group 3	S_n^G	0.001448	0.003776	0.004268	0.002254	0.010113*
	S_n^C	0.000318	0.001068	0.003558	0.001985	0.014572*
	JCS_n^G	0.620920	0.853766	0.853766	0.465690	1.008996
	JCS_n^C	0.001480	0.005813	0.010052	0.007434	0.017275*
Group 4	S_n^G	0.001730	0.004888	0.002593	0.002697	0.006291
	S_n^C	0.000587	0.002044	0.002279	0.002268	0.006855
	JCS_n^G	0.698535	0.853766	0.776151	0.776151	1.164226*
	JCS_n^C	0.002062	0.007503*	0.007827	0.012697	0.019182*
Group 5	S_n^G	0.001379	0.003470	0.004909	0.005554	0.009491
	S_n^C	0.000534	0.001044	0.007032	0.009277*	0.017460*
	JCS_n^G	0.388075	0.620920	0.776151	0.776151	0.776151
	JCS_n^C	0.001413	0.005926	0.009500*	0.013223*	0.016788*
Group 6	S_n^G	0.001767	0.005915	0.004387	0.005942	0.008262
	S_n^C	0.000177	0.002169	0.006224	0.009615	0.020157*
	JCS_n^G	0.931381*	0.853766	0.698535	0.853766	0.698535
	JCS_n^C	0.001680	0.007416*	0.009586	0.012696	0.018839
Group 7	S_n^G	0.002128	0.007057	0.008938	0.004600	0.009174
	S_n^C	0.001056	0.002847	0.008537	0.005034	0.008895
	JCS_n^G	0.465690	0.931381	1.008996	0.620920	1.086611*
	JCS_n^C	0.002062	0.007216*	0.010561*	0.007966	0.015189
Group 8	S_n^G	0.001997	0.003098	0.002059	0.004183	0.012612
	S_n^C	0.000379	0.000686	0.001185	0.001744	0.010617
	JCS_n^G	0.620920	0.620920	0.388075	0.698535	1.086611*
	JCS_n^C	0.001764	0.003088	0.003562	0.004157	0.016000*

* Notes: See notes to Table 3.

Figure 1: Kernel Densities of Panel Expert Nominal GDP Forecasts*



Summary Statistics

	1968:Q4	1983:Q4	1998:Q4	2013:Q4
Mean	923.7701	3686.444	8837.940	17578.68
Median	925.0000	3687.000	8846.000	17567.91
Maximum	960.0000	3750.000	8984.600	17781.06
Minimum	875.0000	3550.000	8475.000	17358.72
Std. Dev.	12.16333	37.60961	94.13889	100.0608
Skewness	-0.884233	-1.310238	-1.786877	-0.027824
Kurtosis	7.574713	6.090962	8.287103	2.839074
Jarque-Bera	87.20107	24.63142	52.60335	0.049531
Probability	0.000000	0.000004	0.000000	0.975539
Observations	87	36	31	41

* Notes: Figures are kernel density estimates (Epanechnikov kernel) for various 4 quarter ahead predictions made during the fourth quarters of 1968, 1983, 1998, and 2013. The number of experts in each sample ranges from 31 in 1998 to 87 in 1968, as noted in table of summary statistics below the plots in the figure. See Section 5 for further details.