Forecast dominance testing via sign randomization

Werner Ehm

Heidelberg Institute for Theoretical Studies
Computational Statistics Group
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany[*]
e-mail: wernehm@web.de

and

Fabian Krüger

Heidelberg University
Alfred-Weber-Institute for Economics
Bergheimer Strasse 58
69115 Heidelberg, Germany
e-mail: fabian.krueger@awi.uni-heidelberg.de

Abstract: We propose randomization tests of whether forecast 1 outperforms forecast 2 across a class of scoring functions. This hypothesis is of applied interest: While the prediction context often prescribes a certain class of scoring functions, it is typically hard to motivate a specific choice on statistical or substantive grounds. We investigate the asymptotic behavior of the test statistics under mild conditions, avoiding the need to assume particular dynamic properties of forecasts and realizations. The properties of the one-sided tests depend on a corresponding version of Anderson's inequality, which we state as a conjecture of independent interest. Numerical experiments and a data example indicate that the tests have good size and power properties in practically relevant situations.

Contents

---

[*]Recently retired

1. Introduction

Forecasts of future events and quantities are essential across disciplines. At the same time, forecasts notoriously are imprecise and prone to bias, calling for methods to assess and compare the performance of imperfect predictions both theoretically and on the basis of empirical data. In the context of point forecasts, which we consider in this paper, the appropriate evaluation tool is that of a *consistent scoring function* [21]. A scoring function $S \equiv S(x, y)$ assigns to each forecast x and realization y a real-valued score such that smaller scores correspond to better forecasts. Specifically, let ϕ be a real-valued functional defined on a class $\mathcal{G}$ of possible distributions G of y, such as the mean or a quantile of G. The scoring function is *consistent (for ϕ relative to $\mathcal{G}$)* if $S(\phi(G), G) \leq S(x, G)$ for every $G \in \mathcal{G}$ and forecast x; here $S(x, G) = E_{y \sim G} S(x, y)$. With a consistent scoring function the forecaster can do no better than predict the true functional value, which rewards honest reporting.

For a given functional ϕ, supposed fixed in the following, there generally exists a whole class of consistent scoring functions, or "scores." Characterizations of the respective score classes for various functionals may be found in e.g. [5, 19, 21]. For example, all scores of the form $S(x, y) = \varphi(y) - \varphi(x) - \varphi'(x)\ (y - x)$, φ a convex function with subgradient φ', are consistent for the mean functional [5, 49]. In applied contexts, consistent alternatives to the special case $\varphi(x) = x^2$ of the squared error score were discussed in [22], [37], [50] and others for the binary case $y \in \{0, 1\}$, and in [43] for positive predictands $y \in \mathbb{R}_+$.

The availability of an entire family of scoring functions that are theoretically legitimate comes with the drawback that two scores may produce different forecast rankings even if both are consistent for the same target functional [16, 17, 44]. This lack of robustness is unsatisfactory as there are often no strong

arguments for choosing a particular score. It is therefore natural to ask whether forecast rankings are stable across a family of scores. E.g., given two forecasts x_1, x_2, does x_1 dominate x_2 in the sense that it is superior with respect to *all* consistent scoring functions? In the case of a quantile or expectile functional[1] the complexity of the problem can be much reduced by means of a Choquet representation: here every consistent score S can be represented as a mixture of "elementary" (in fact: extremal) scores S_θ, $\theta \in \mathbb{R}$. That is, for every S there is a nonnegative Borel measure M on $\mathbb{R}$ such that $S(x, y) = \int S_\theta(x, y) \, dM(\theta)$ [16]. The form of the elementary scores depends on the specific functional being studied; for example, in case of the mean functional

$$S_\theta(x, y) = \begin{cases} |y - \theta| & \text{if } \min\{x, y\} \equiv x \wedge y \leq \theta < x \vee y \equiv \max\{x, y\}, \\ 0 & \text{else.} \end{cases} \quad (1.1)$$

The Choquet representation makes it possible to reduce dominance with respect to all consistent scoring functions to dominance with respect to the linearly indexed family of elementary scores $\{S_\theta : \theta \in \mathbb{R}\}$, a substantial simplification. Nevertheless, testing related hypotheses remains a challenging task.

Our focus will be on a very simple test of forecast dominance that goes without any of the common assumptions. In return, the notion of forecast dominance acquires a central role. Initially, it may be defined as follows [16]. Given a family $\mathcal{S} = \{S_\theta : \theta \in \Theta\}$ of (consistent) scoring funcions S_θ we say that *forecast x_1 dominates forecast x_2 at the distribution $G \in \mathcal{G}$ if $S_\theta(x_1, G) \leq S_\theta(x_2, G)$ for every $\theta \in \Theta$*. In such a one-step scenario, the set of all $G \in \mathcal{G}$ satisfying this condition could constitute the hypothesis 'x_1 dominates x_2' (with respect to $\mathcal{S}$). The question how to formulate suitable dominance hypotheses becomes substantially more involved when, as usually, forecasts x_{k1}, x_{k2} are produced step by step and the realizations y_k become known before the next forecast instance. Current work on forecast evaluation and comparison emphasizes the joint dynamic behavior of forecasts and realizations, by using martingale methods [34, 51], the concept of prediction spaces [23, 52], or comparisons of conditional predictive ability [20]. Bootstrap tests of dominance hypotheses are presented in, e.g., [28, 38, 39, 56], based on different dominance concepts and asymptotic frameworks. An account of the related literature addressing the relations with, and differences to the present approach is given in the discussion section 7.

Usually, mathematical analyses proceed from statistical models for the data and the formulation of hypotheses to related tests and their properties. Here we follow a reverse path. We make no assumptions about possible data generating mechanisms; instead we focus on a simple-to-implement test procedure and *ask for hypotheses for which this procedure represents a valid test* (asymptotically, at a given level). We take this route because quite often very little is known about the stochastic nature of the data. In fact, typical forecasting problems have to cope with complex statistical dependencies, structural change, and lim-

---

[1]Expectiles are an asymmetric generalization of the mean; they were introduced by [42] and have recently received attention in financial risk management. We provide a formal definition of expectiles in Section 4.

ited domain knowledge. Thus presumably, most of the usual assumptions do not apply, with largely unknown consequences, and are hard or impossible to check. We therefore have recourse to the classical Fisherian technique of external randomization, which is completely under one's control, and treat everything conditionally on the data (x_{k1}, x_{k2}, y_k), $k = 1, \ldots, n$.

The use of external randomization to compare forecast performance dates back at least to [14]. Here we compare forecast performance across families of scores, rather than with respect to a single scoring function. Concretely, our goal is to elaborate on the sign randomization procedure tentatively proposed in [16, end of Section 3] for testing forecast dominance. The idea is to reject the hypothesis 'forecast 1 dominates forecast 2' if, e.g., $\sup_{\theta \in \Theta} D_n(\theta)$ exceeds some critical value c_n where

$$D_n(\theta) = n^{-1/2} \sum\nolimits_{k \leq n} d_k(\theta), \quad d_k(\theta) = S_\theta(x_{k1}, y_k) - S_\theta(x_{k2}, y_k)$$

is the sum of the single score differences $d_k(\theta)$, properly scaled for the sake of asymptotics. Unfortunately, determination of c_n generally is very diffcult even asymptotically; it appears impossible without making assumptions about the stochastic structure of the data. Our suggestion in [16] was to determine c_n such that $Pr^*[\sup_{\theta \in \Theta} D_n^*(\theta) > c_n] \approx \alpha$, the test level, where

$$D_n^*(\theta) = n^{-1/2} \sum\nolimits_{k \leq n} d_k(\theta)\sigma_k$$

and Pr^* exclusively refers to the i.i.d. ("Rademacher") random variables σ_k assuming the values ± 1 with probability $1/2$ each. This clearly raises questions.

First, how can the randomization distribution be connected to the distribution of the test statistic, particularly when no model assumptions are being made? Secondly, what precisely is to be understood under the hypothesis 'forecast 1 dominates forecast 2'? As explained in Section 3.2, there is in fact a close connection between the two problems that helps to get around both – up to one missing link: The approximate validity of our one-sided tests depends on an unproven variant of the celebrated Anderson's inequality [2]. While for symmetric hypotheses postulating 'no difference in predictive performance' the classical Anderson's inequality provides the necessary link, the asymmetry of dominance hypotheses requires a one-sided version of the inequality which we state as a conjecture that appears of independent interest.

Obviously, dispensing with model asssumptions cannot mean doing without any assumptions. However, as detailed in Section 4.1, asssumptions distantly related to stationarity and (in-)dependence properties of forecasts and observations will enter in a very indirect manner only, via basic asymptotic stability and "moderate local clustering" conditions, respectively, which are fulfilled under virtually any of the standard statistical models; cf. Section 9. Building on this novel asymptotic framework we present, in Section 4, weak convergence results governing the asymptotics of our test statistics in the important special cases of quantile and expectile forecasts. For the overall organization of the paper see the table of contents. R [46] program code to implement the randomization test is available at https://github.com/FK83/fdtest.

2. Testing for forecast dominance – Initial considerations

2.1. Formal setup

Let (x_{k1}, x_{k2}, y_k), $k = 1, \ldots, n$ be a sequence of n triplets where x_{k1}, x_{k2} are two point forecasts each for the subsequent observation y_k. The triplets are considered as random variables on a common probability space $(\Omega, \mathcal{F}, Q)$ endowed with a filtration $\{\mathcal{F}_k, k = 0, \ldots, n\}$ such that (x_{k1}, x_{k2}, y_k) is $\mathcal{F}_k$-measurable for every k, and $\mathcal{F}_0$ is trivial. Given a family $\mathcal{S} = \{S_\theta, \theta \in \Theta\}$ of scoring functions S_θ, we compare the two forecasts via the suitably normalized average difference of their S_θ scores, i.e., we are interested in the stochastic process

$$\theta \mapsto D_n(\theta) = n^{-1/2} \sum\nolimits_{k \leq n} d_k(\theta), \quad d_k(\theta) = S_\theta(x_{k1}, y_k) - S_\theta(x_{k2}, y_k). \quad (2.1)$$

Initially, the family $\mathcal{S}$ may be arbitrary; from Section 4 on it will be specialized to the elementary scores for quantile and expectile forecasts.

2.2. Notions of forecast dominance

Which notion of forecast dominance fits with the randomizarion test in our focus? One possibility to introduce forecast dominance in the present framework is to declare forecast 1 as weakly dominating forecast 2 at Q (with respect to $\mathcal{S}$) if $E_Q D_n(\theta) \leq 0$ for every $\theta \in \Theta$. The same condition furnishes a natural one-sided hypothesis in a testing context:

$$H_-^w : \quad \sup_\theta E_Q D_n(\theta) = \sup_\theta n^{-1/2} \sum\nolimits_{k \leq n} E_Q d_k(\theta) \leq 0. \quad (2.2)$$

In fact, H_-^w stands for all probability measures Q under which $\sup_\theta E_Q D_n(\theta) \leq 0$. Note that the hypothesis does not depend on the scaling constant, which could also be taken as n^{-1}, or the constant 1. The present choice, $n^{-1/2}$, is simply for convenience in regard to the large sample asymptotics to be discussed later on.

The hypothesis H_-^w involves unconditional expectations referring to both the observations y_k and the forecasts $x_{k\ell}$. This form of H_-^w is at odds with the dominance concept of Section 1 which is sequential in nature and makes no assumption about the dynamics of the forecasts, hence is more flexible in this sense. Better accordance with this initial concept is achieved on replacing the unconditional expectations $E_Q d_k(\theta)$ in (2.2) by conditional expectations given the past. This leads upon the following definition of forecast dominance: we say that *forecast 1 dominates forecast 2 at Q* if

$$M_{n,Q}(\theta) = n^{-1/2} \sum\nolimits_{k \leq n} E_Q [d_k(\theta) \,|\, \mathcal{F}_{k-1}] \leq 0 \quad (Q\text{-a.s.}, \theta \in \Theta) \quad (2.3)$$

(a.s. is short for almost surely). The corresponding hypothesis, H_-, then comprises all probabilities Q for which (2.3) holds,

$$H_- : \quad M_{n,Q}(\theta) \leq 0 \quad (Q\text{-a.s.}, \theta \in \Theta). \quad (2.4)$$

Hypothesis H_- is defined by conditions on a random process rather than on a parameter, which is unusual. Nevertheless, it makes for a useful notion of forecast dominance, and it will prove to be the appropriate hypothesis for the randomization test. We also will consider sub-hypotheses of H_- including, specifically, the hypothesis

$$H_-^s : \quad E_Q\left[d_k(\theta)\,|\,\mathcal{F}_{k-1}\right] \le 0 \quad (Q\text{-a.s.}, \ \theta \in \Theta, \ k = 1, \ldots, n).^2 \tag{2.5}$$

The interpretation of H_-^s is straightforward: It says that forecast 1 is at least as good as forecast 2 at each time step. In [20], this dominance concept is referred to as a comparison of *conditional predictive ability.* In this parlance, the hypotheses H_- and H_-^s express superiority of forecast 1 over forecast 2 in terms of, respectively, average and step-by-step conditional predictive ability.

Example 2.1. For illustration we consider the case where the forecasters know one part each of the verifying observation. Specifically, let $y_k = \eta_{k1} + \eta_{k2}$ where $\eta_{k\ell}$, $k \ge 1$, $\ell = 1, 2$ are two independent autoregressive processes of the form $\eta_{k\ell} = a\,\eta_{k-1,\ell} + \epsilon_{k\ell}$ with the same parameter a and independent innovations $\epsilon_{k\ell} \sim \mathcal{N}(0, \tau_\ell^2)$. Suppose that at any instance k, forecaster 1 has access to η_{k1} and the preceding value $\eta_{k-1,2}$ of the second process.[3] If a is known, a natural choice for the prediction of y_k is $x_{k1} = \eta_{k1} + a\,\eta_{k-1,2}$. By definition,

$$x_{k1} = a\,\eta_{k-1,1} + \epsilon_{k1} + a\,\eta_{k-1,2} = a\,y_{k-1} + \epsilon_{k1},$$

and if forecaster 2's prediction similarly is $x_{k2} = \eta_{k2} + a\,\eta_{k-1,1}$, then $x_{k2} = a\,y_{k-1} + \epsilon_{k2}$, and of course, $y_k = a\,y_{k-1} + \epsilon_{k1} + \epsilon_{k2}$. Taking at first squared error as the scoring function, the k-th score difference becomes

$$d_k = (y_k - x_{k1})^2 - (y_k - x_{k2})^2 = \epsilon_{k2}^2 - \epsilon_{k1}^2.$$

Thus if $\tau_1 \ge \tau_2$, say, and the innovations $\epsilon_{k\ell}$ are independent of $\mathcal{F}_{k-1}$ – as in the common case where $\mathcal{F}_k$ is the σ-algebra generated by all triplets (x_{j1}, x_{j2}, y_j), $j \le k$ –, then $E\left[d_k\,|\,\mathcal{F}_{k-1}\right] \le 0$, consistent with the intuition that the forecaster having access to the more variable component should be better off. The case $\tau_1 = \tau_2$ is an instance of a situation where, a priori, none of the two forecasters is believed to outperform the other. Here $E\left[d_k\,|\,\mathcal{F}_{k-1}\right] = 0$, which holds in fact for *every* scoring function S within the present model: by symmetry the joint conditional distributions of $(x_{kj}, y_k) = (a\,y_{k-1} + \epsilon_{kj}, a\,y_{k-1} + \epsilon_{k1} + \epsilon_{k2})$ given $\mathcal{F}_{k-1}$ are identical $(j = 1, 2)$, whence the conditional expectation of $d_k = S(x_{k1}, y_k) - S(x_{k2}, y_k)$ vanishes. In particular, $E\left[d_k(\theta)\,|\,\mathcal{F}_{k-1}\right] = 0$ for every $\theta \in \mathbb{R}$, where $d_k(\theta)$ denotes the score difference with respect to the elementary scoring functions S_θ for the mean value; see (1.1). In this case we even have the following.

Proposition 2.1. *If $\tau_1 > \tau_2$ then $E\left[d_k(\theta) \mid \mathcal{F}_{k-1}\right] \le 0$ for every $\theta \in \mathbb{R}$ and $k \le n$, i.e., H_-^s (and a fortiori H_-) holds.*

---

[2]If the functions $\theta \mapsto d_k(\theta)$ are sufficiently regular, the conditions (2.4), (2.5) hold Q-a.s. *for all $\theta \in \Theta$ simultaneouls*y. We do not further dwell on this technicality.

[3]This setup is similar to the simulation example in [23, Section 4.1], except that our variant includes time series dynamics.

The example indicates that H_- and H_-^s represent meaningful conditions characterizing different forms of forecast dominance. Further examples may be found in the recent systematic study [32] of dominance conditions based on the concept of convex order.

2.3. A fictitious test

For testing the one-sided hypothesis H_- (or H_-^s) it is natural to work with test functionals $T = T(D_n)$ that are *monotone* in the sense that $T(f) \leq T(g)$ whenever $f \leq g$ pointwise on Θ, an interval in $\mathbb{R}$ from now on. Main examples are the supremum functional and integrals of the positive part of D_n, e.g.

$$T_\infty(f) = \sup_{\theta \in \Theta} f(\theta), \quad T_p(f) = \int_\Theta f(\theta)_+^p d\theta \quad (p \geq 1, \ a_+ = a \vee 0). \quad (2.6)$$

These functionals are convex in f, and accordingly, *the generic test functional T is supposed to be monotone and convex* in the real functions f on Θ.

Importantly for the following, in tests of H_- or H_-^s based on such a functional T it suffices to control the error of the first kind at the "boundary" of the hypothesis, where either $M_{n,Q}(\theta) \equiv 0$ or $E_Q[d_k(\theta) \mid \mathcal{F}_{k-1}] \equiv 0$. Indeed, put

$$\widetilde{d}_{k,Q}(\theta) = d_k(\theta) - E_Q[d_k(\theta) \mid \mathcal{F}_{k-1}],$$

and let

$$\widetilde{D}_{n,Q}(\theta) = n^{-1/2} \sum_{k \leq n} \widetilde{d}_{k,Q}(\theta) = D_n(\theta) - M_{n,Q}(\theta)$$

denote the conditionally centered version of D_n. Then by monotonicity

$$Pr_Q[T(D_n) > c] = Pr_Q[T(\widetilde{D}_{n,Q} + M_{n,Q}) > c] \leq Pr_Q[T(\widetilde{D}_{n,Q}) > c] \quad (2.7)$$

for every $Q \in H_-$, as claimed. Thus, *if critical values $\widetilde{c}_n$ could be obtained such that* $\sup_{Q \in H_-} Pr_Q[T(\widetilde{D}_{n,Q}) > \widetilde{c}_n] \approx \alpha,$[4] *the rule 'reject H_- if $T(D_n) > \widetilde{c}_n$' would give us an approximate level-α test of H_-.* For compactness of notation we henceforth drop the index Q, taking the dependence on the underlying probability measure as self-evident.

An initial step toward the determination of critical values is the following proposition, for which we need the Lindeberg condition

(A0) $\limsup_{n \to \infty} n^{-1} \sum_{k \leq n} E\{d_k(\theta)^2 \mathbb{1}_{|d_k(\theta)| > \epsilon\sqrt{n}}\} = 0 \quad (\theta \in \Theta, \ \epsilon > 0).$

Proposition 2.2. *Suppose there is some non-random function $\widetilde{\gamma}$ such that $\widetilde{\gamma}(\theta, \theta) > 0$, $\theta \in \Theta$, and for every pair $\theta_1, \theta_2 \in \Theta$*

$$n^{-1} \sum_{k \leq n} \widetilde{d}_k(\theta_1)\, \widetilde{d}_k(\theta_2) \equiv \widetilde{\gamma}_n(\theta_1, \theta_2) \longrightarrow_p \widetilde{\gamma}(\theta_1, \theta_2) \quad as \ n \to \infty. \quad (2.8)$$

Then under (A0), the finite-dimensional distributions of the process $\widetilde{D}_n$ converge to those of a mean zero Gaussian process $\widetilde{Z}$ with covariance $\widetilde{\gamma}$.

---

[4]The notation '$a_n \approx b_n$' is short for $a_n - b_n \to 0$ as $n \to \infty$.

The proposition suggests that for large n the distribution of the test statistic $T(D_n)$ at the boundary, where $D_n = \widetilde{D}_n$, can be approximated by the distribution of the functional $T(\widetilde{Z})$ on the paths of the Gaussian process $\widetilde{Z}$. Of course, convergence of the finite-dimensional distributions is insufficient for such a conclusion; tightness of the processes $\widetilde{D}_n$ in a suitable function space is required, too. Furthermore, the distribution of $T(\widetilde{Z})$ generally is unknown and may be difficult to determine. And there still is the problem that the process $\widetilde{D}_n$ involves the (sum of the) conditional expectations $E\,[\,d_k(\theta)\,|\,\mathcal{F}_{k-1}]$, which depend on the unknown probability Q and would have to be estimated with sufficient accuracy. In view of these difficulties with the determination of proper critical values $\widetilde{c}_n$ we refer to the hypothetical test rejecting H_- if $T(D_n) > \widetilde{c}_n$ as the "fictitious test."

For the sake of exposition we introduce further sub-hypotheses of H_-[5] besides H_-^s, namely the null-hypothesis H_0 of equal performance on average,

$$H_0: \quad n^{-1/2} \sum\nolimits_{k \leq n} E\,[\,d_k(\theta)\,|\,\mathcal{F}_{k-1}] \equiv M_n(\theta) = 0 \quad (\text{a.s.}, \ \theta \in \Theta), \qquad (2.9)$$

and the null-hypothesis H_0^s of equal performance at every forecast instance,

$$H_0^s: \quad E\,[\,d_k(\theta)\,|\,\mathcal{F}_{k-1}] = 0 \quad (\text{a.s.}, \ \theta \in \Theta, \ k = 1, \dots, n). \qquad (2.10)$$

Evidently, H_0, H_0^s may be regarded as the *boundaries* of the one-sided hypotheses H_-, H_-^s, respectively, representing different forms of equal predictive ability. Note that $H_0^s \subset H_-^s \subset H_- \subset H^w$, and $H_0 \subset H_-$; moreover, that a test that is valid for a given hypothesis remains valid when used as a test of a smaller sub-hypothesis; it may be invalid when used with a larger hypothesis comprising the first.

3. Randomization tests

3.1. General idea

Using a standard randomization device (c.f. Section 7), we let $\sigma_1, \sigma_2, \dots$ be i.i.d. such that $\sigma_k = \pm 1$ with probability $1/2$ each, and define

$$D_n^*(\theta) = n^{-1/2} \sum\nolimits_{k \leq n} d_k(\theta)\sigma_k. \qquad (3.1)$$

We reject H_- "at level α" if $T(D_n) > c_n^*$ where T is a monotone and convex test functional, and c_n^* is determined such that $Pr^*[T(D_n^*) > c_n^*] \approx \alpha$. Here Pr^* exclusively pertains to the random signs σ_k, the data x_{k1}, x_{k2}, y_k being considered as fixed, non-random quantities. Henceforth we refer to this test as

---

[5] In testing for forecast dominance the hypothesis H_- is of our primary interest. The other hypotheses serve to develop the ideas step by step. In particular, the randomization test is not intended, nor apt, to test e.g. H_0^s vs. H_0, or H_-^s vs. H_-.

the *randomization test.* Its rationale is as follows. First, if T is monotone, the argument at (2.7) together with the inclusion $H_0 \subset H_-$ imply

$$\sup_{H_-} Pr\left[T(D_n) > c_n^*\right] = \sup_{H_0} Pr\left[T(D_n) > c_n^*\right], \tag{3.2}$$

so that it suffices to control the error of the first kind across H_0 where there is no systematic difference between the two forecasts. Secondly, if there is no difference in predictive performance between forecasts 1 and 2, changing the labels (i.e., the sign of the d_k) should not affect the distribution of the test statistic. The quotes in "at level α" shall underline that the test has (approximative) level α only *formally;* the actual test level might differ.

While the randomization test has intuitive appeal and is easy to implement, its properties are less clear. For instance, calculating the critical value c_n^* from the randomization distribution tacitly supposes that in the indifference case the distribution of the $\mathbb{R}^n$-valued process $\theta \mapsto (d_1(\theta), \ldots, d_n(\theta))$ is invariant under arbitrary sign changes in the n components (same change for all θ), which is an even stronger hypothesis than H_0^s. This raises questions concerning the approximate range of validity of the test in asymptotic regimes, where fine distinctions between different hypotheses may become inessential.

Initial answers will be obtained through (partially) heuristic reasoning, utilizing relationships between the randomization and the fictitious test. In Section 4 these considerations are complemented by rigorous weak convergence results for the case of quantile and expectile forecasts, which validate the heuristics.

3.2. Test validity: Heuristics, and a conjecture

In part a) of the following proposition it is understood that chance enters in two ways: via the random signs σ_k, and via the statistical nature of the data triplets. In part b) we condition on the data, leaving the σ_k as the sole source of randomness.

Proposition 3.1. *a) Suppose there is a non-random function γ such that $\gamma(\theta, \theta) > 0$, $\theta \in \Theta$, and for every pair $\theta_1, \theta_2 \in \Theta$*

$$n^{-1} \sum\nolimits_{k \leq n} d_k(\theta_1) \, d_k(\theta_2) \equiv \gamma_n(\theta_1, \theta_2) \longrightarrow_p \gamma(\theta_1, \theta_2). \tag{3.3}$$

Then under assumption (A0) the finite-dimensional distributions of the process D_n^* converge to those of a mean zero Gaussian process Z with covariance γ.
b) The latter conclusion also holds under Pr^* (i.e., conditionally on the data) provided that the stochastic convergence (3.3) is replaced by the usual (deterministic) convergence, and the Lindeberg condition (A0) is satisfied without the expectation sign.

Remark 3.1. Regarding part b), note that under Pr^* the $d_k(\theta)$ are known non-random quantities, rendering the expectation sign void. On the other hand, if in (A0) $d_k(\theta)$ everywhere is replaced by $d_k^*(\theta) = d_k(\theta)\sigma_k$, and E by the expectation E^* pertaining to the σ_k only, then the resulting condition (A0*) is

a Lindeberg condition in the usual sense. Anyway, since $|d_k^*(\theta)| = |d_k(\theta)|$, there is no difference between the conditions with and without the expectation sign, and we need not distinguish (A0) and (A0*).

We now address the question for which among the above hypotheses the randomization test is approximately valid, i.e., has size $\lesssim \alpha$. The discussion builds on distributional approximations to be established later on and on an unproven conjecture. The argument still is instructive as it helps delineate the key problem. Let us re-emphasize that in regard to testing for forecast dominance our focus is on the one-sided hypothesis H_-; cf. Footnote 5.

Hypothesis H_0^s.
Under H_0^s we have $d_k(\theta) = \widetilde{d}_k(\theta)$, hence $D_n = \widetilde{D}_n$ and $\gamma_n = \widetilde{\gamma}_n$. Consequently, the limit processes Z and $\widetilde{Z}$ of D_n^* and $\widetilde{D}_n$ are identical in distribution under H_0^s, and so are the limit distributions of any sufficiently regular test statistic (which need not be monotone, for instance). In particular, the critical values c_n^* and $\widetilde{c}_n$ of the randomization and the fictitious test coincide asymptotically. Therefore, since the latter test is, for large n, approximatively valid for testing H_0^s at level α, then so is the former. The point is, of course, that the fictitious test is valid but infeasible, whereas the randomization test is straightforward to implement.

Hypothesis H_0.
Under this hypothesis the above reasoning does not apply because the covariance functions γ and $\widetilde{\gamma}$, hence the limit processes Z and $\widetilde{Z}$, generally are different. Nevertheless, the randomization test remains approximatively valid for the hypothesis H_0 if the test functional T fits with H_0. To substantiate this claim, let us begin by noting that $\gamma = \widetilde{\gamma} + \psi$ with a positive definite function ψ given by the stochastic limit

$$\psi(\theta_1, \theta_2) = p\text{-}\lim \, n^{-1} \sum\nolimits_{k \leq n} E\left[\, d_k(\theta_1) \,|\, \mathcal{F}_{k-1}\right] E\left[\, d_k(\theta_2) \,|\, \mathcal{F}_{k-1}\right]; \qquad (3.4)$$

cf. Lemma 9.1. For the limit processes this means that $Z = \widetilde{Z} + W$ in distribution, where W is an independent centered Gaussian process.

Now in the context of the hypotheses H_0, H_0^s it is natural to consider test functionals T that are *symmetric,* $T(-f) = T(f)$, and convex in f. In other words, the acceptance region $A = \{f : T(f) \leq c\}$ is symmetric and convex. This allows us to control the error probability under H_0 by applying a celebrated inequality. A basic finite-dimensional version of the inequality is as follows.

Anderson's inequality [2, Corollary 3]. *Let X, Y be independent centered $\mathbb{R}^d$-valued Gaussian random variables. Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex and symmetric, i.e. $g(-x) = g(x)$ for every x. Then $Pr[g(X) \leq b] \geq Pr[g(X+Y) \leq b]$ for every $b \in \mathbb{R}$.*

In our case X and Y correspond to $\widetilde{Z}$ and W sampled discretely at d points $\theta_j \in \Theta \subset \mathbb{R}$. Examples of functions g corresponding to test functionals T of interest are $g(x) = \max_i |x_i|$ and $g(x) = \sum_i |x_i|^p$, $p = 1$ or $p = 2$. As the

sampling gets dense, one finds that under H_0 and for symmetric, convex test functionals T one has

$$Pr\left[T(D_n) > c_n^*\right] \overset{(1)}{=} Pr\left[T(\widetilde{D}_n) > c_n^*\right] \approx Pr\left[T(\widetilde{Z}) > c_n^*\right] \quad (3.5)$$

$$\overset{(2)}{\leq} Pr\left[T(Z) > c_n^*\right] \approx Pr^*\left[T(D_n^*) > c_n^*\right] \approx \alpha.$$

Relation (1) holds because $D_n = \widetilde{D}_n$ under H_0, relation (2) by Anderson's inequality along with a standard approximation [2, Proof of Corollary 4], and the three '$\approx$' signs hold by Propositions 2.2 and 3.1, and by construction, respectively. Note that (3.5) in fact implies $\sup_{H_0} Pr\left[T(D_n) > c_n^*\right] \approx \alpha$ since $H_0 \supset H_0^s$ and the admissible error probability is fully exhausted on H_0^s.

Hypotheses H_-, H_-^s.
The test functionals $T = T(f)$ appropriate for these hypotheses are convex and *monotone* in f. The latter property is incompatible with symmetry, which is an essential ingredient of Anderson's inequality. We nevertheless could argue similarly as above if there was a one-sided version of Anderson's inequality. The following would be most helpful.

A one-sided Anderson's inequality? – **Conjecture:** *Let X, Y be independent centered $\mathbb{R}^d$-valued Gaussian random variables. Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex and* monotone *in the sense that $g(x) \leq g(y)$ whenever $x \leq y$ (coordinatewise). Then there is a universal constant $\alpha_0 \in (0, 1/2]$ (bold guess: $\alpha_0 = 1/2$) such that $Pr[g(X) \leq b] \geq Pr[g(X + Y) \leq b]$ whenever $Pr[g(X + Y) \leq b] \geq 1 - \alpha_0$.*

Remark 3.2. An important consequence of this inequality is that *for test levels $\alpha \leq \alpha_0$ the randomization test is (approximately) valid for testing the one-sided hypotheses H_-, H_-^s.* The argument is parallel to (3.5), with two modifications: first, since $M_n \leq 0$ under H_-, H_-^s, instead of (1) we have an inequality, $Pr\left[T(D_n) > c_n^*\right] \leq Pr\left[T(\widetilde{D}_n) > c_n^*\right]$, by the monotonicity of T; secondly, relation (2) now follows from the one-sided Anderson inequality, again up to an approximation as in [2, Proof of Corollary 4].

Remark 3.3. The one-sided Anderson's inequality is *not* needed for the approximative validity of the one-sided randomization test if we only consider (sequences of) probability measures $Q_n \in H_-$ that are *contiguous* [54, p. 87] to some sequence $P_n \in H_0^s$. This is because under P_n we have $\gamma_n = \widetilde{\gamma}_n$, hence $\gamma_n \longrightarrow_p \gamma = \widetilde{\gamma}$, and by contiguity this convergence also takes place under Q_n; cf. Propositions 2.2, 3.1. Therefore the distributions of the limit processes Z and $\widetilde{Z}$ coincide, and the inequality (2) in (3.5) becomes an equality; while the '$=$' sign (1) there has to be replaced by '$\leq$', again by the monotonicity of T. Thus in this case too, $Pr_{Q_n}[T(D_n) > c_n^*] \lessapprox \alpha$, as claimed.

Summary. *Asymptotically, the randomization test is an (approximatively) valid level-α test of the hypotheses H_0, H_0^s. If the conjecture is correct it is also valid for testing the hypotheses H_-, H_-^s.*

It should be emphasized that the above discussion exclusively pertains to the control of the error of the first kind. Regarding power we only mention that the argument in 3.3 may as well be applied to *alternatives* $Q_n \in H_+$ – satisfying $M_{n,Q_n} \geq 0$ Q_n-a.s. for all $\theta \in \Theta$; cf. (2.4) – that are contiguous to a sequence $P_n \in H_0^s$. By monotonicity of T one analogously gets $Pr_{Q_n}[T(D_n) > c_n^*] \gtrsim \alpha$, i.e., *the randomization test is unbiased against such alternatives.* For alternatives Q_n not in H_+, where $\theta \mapsto M_{n,Q_n}(\theta)$ assumes positive as well as negative values, the power issue is subtle and will not be pursued systematically in this paper.

3.3. *Some comments on the conjectured inequality*

In dimension $d = 1$, the inequality is trivial. Convex, monotone acceptance regions then are intervals of the form $(-\infty, b]$, and if $X_i \sim \mathcal{N}(0, v_i)$ $(i = 1, 2)$ with $v_1 \leq v_2$, then obviously $Pr[X_1 \leq b] \geq Pr[X_2 \leq b]$ if and only if $b \geq 0$, i.e., if and only if $Pr[X_2 \leq b] \geq 1/2$. (This fits with the bold guess $\alpha_0 = 1/2$.)

For $d > 1$, a small piece of evidence in favour of the conjecture can be given as follows. Let $A \subset \mathbb{R}^d$ be a convex acceptance region of the form $A \equiv A_{g,b} = \{x \in \mathbb{R}^d : g(x) \leq b\}$ for some convex function g and $b \in \mathbb{R}$. (Monotonicity of g is not required for the argument.) Denote by $G = \mathcal{N}(0, V)$, $G_+ = \mathcal{N}(0, V_+)$ the distributions of the random variables X and $X + Y$, respectively. The matrices V, V_+ and $V_+ - V$ are symmetric and (strictly) positive definite. Put $K = A \cap (-A)$, which intersection is convex and symmetric, and let R denote the complement of the union $A \cup (-A)$. Then for any symmetric probability distribution F on $\mathbb{R}^d$ we have $1 = F(A) + F(-A) - F(K) + F(R) = 2F(A) - F(K) + F(R)$ or $F(A) = (1 + F(K) - F(R))/2$. For the moment being, suppose that R is contained in the set S where the density of G_+ exceeds the density of G. Then $G_+(R) \geq G(R)$, and an application of Anderson's inequality to the set K, $G_+(K) \leq G(K)$, yields the desired conclusion,

$$G_+(A) = (1 + G_+(K) - G_+(R))/2 \leq (1 + G(K) - G(R))/2 = G(A).$$

As for the possible inclusion $R \subset S$, note that in terms of the log densities

$$S = \{x : x'(V^{-1} - V_+^{-1})x > \log(|V_+|/|V|)\}.$$

Now $V_+ > V$ implies $V^{-1} > V_+^{-1}$ in the Loewner order [27, Corollary 7.7.4(a)]. Therefore $\Delta = V^{-1} - V_+^{-1}$ is positive definite, and noting that $L = \log(|V_+|/|V|) > 0$ we find that S is the complement of the ellipsoid $E = \{x : x'\Delta x \leq L\}$. Since $A_{g,b} \uparrow \mathbb{R}^d$ as $b \uparrow \infty$ and g is bounded on E, we have for all large enough b that $S^c = E \subset A \subset R^c$, that is, $R \subset S$. But $b \to \infty$ iff the test level $\alpha \to 0$, so we have proved the following.

Proposition 3.2. *If $A = A_{g,b}$ for some convex function g and $b \in \mathbb{R}$, then there is $\alpha_0 \in (0, 1)$ such that $Pr[g(X) \leq b] \geq Pr[g(X + Y) \leq b]$ whenever $Pr[g(X + Y) \leq b] \geq 1 - \alpha_0$.*

In our case, $X = (\widetilde{Z}(\theta_1), \ldots, \widetilde{Z}(\theta_d))$, $Y = (W(\theta_1), \ldots, W(\theta_d))$ with the θ_j becoming dense. Since the covariance function $\widetilde{\gamma}$ of $\widetilde{Z}$ generally is unknown, we

have no control on the eigenvalues of $V^{-1} - V_+^{-1}$. Proposition 3.2 thus does not guarantee that α_0 stays bounded away from zero uniformly in the pair V, V_+ and all dimensions d, as it is necessary for the one-sided Anderson inequality. This uniformity is the core of the problem.

A proof of the conjecture may require additional assumptions, e.g. invariance of g under coordinate permutations. (Generalizations involving other invariance conditions appear in [11, 41].) Relevant examples include the convex, monotone functions $g(x) = \max_i x_i$, $g(x) = \sum_i (x_i)_+^p$ $(p \geq 1)$, which correspond to test functionals of major interest; cf. (2.6). A proof for such a special case would already be most worthwhile.

Thus far, our numerical experiments in the bivariate case $d = 2$ and simulations with randomly generated covariance matrices for $d > 2$ yielded no counterexample. Needless to say, this is irrelevant for the conjecture.

4. Asymptotics for quantile and expectile forecasts

In principle, the developments so far apply to largely arbitrary functionals ϕ on the class $\mathcal{G}$ of predictive distributions G and related families of consistent scoring functions S_θ. Hereafter we focus on functionals representing a quantile or an expectile. Given $\alpha \in (0, 1)$, the α-expectile of G is defined as the unique solution t to the equation $(1 - \alpha) \int_{-\infty}^t (t - y) \, dG(y) = \alpha \int_t^\infty (y - t) \, dG(y)$ [42]; for $\alpha = 1/2$ one obtains the mean value functional. As usual, $q = \inf\{y : G(y) \geq \alpha\}$ is the (lower) α-quantile of G, which here is identified with its right-continuous CDF. The median of G obtains when $\alpha = 1/2$.

As mentioned in Section 1, forecast dominance with respect to all consistent scoring functions is, for these functionals, equivalent to dominance with respect to a certain linearly indexed family of "elementary" scoring functions S_θ, such that any consistent scoring function can be written as

$$S(x, y) = \int S_\theta(x, y) \, dM(\theta), \tag{4.1}$$

where M is a nonnegative Borel measure on $\mathbb{R}$. Different functions $S(x, y)$ are characterized by different measures M. From [16, Theorem 1], the elementary scores for α-quantiles are

$$S_\theta(x, y) = \{\mathbb{1}_{y<x} - \alpha\} \{\mathbb{1}_{\theta<x} - \mathbb{1}_{\theta<y}\}; \tag{4.2}$$

setting M in (4.1) equal to the Lebesgue measure results in the popular 'piecewise linear' score considered in quantile regression [30]. For α-expectiles,

$$S_\theta(x, y) = |\mathbb{1}_{y<x} - \alpha| \{(y - \theta)_+ - (x - \theta)_+ - (y - x) \mathbb{1}_{\theta<x}\}; \tag{4.3}$$

again, the most popular choice of M is the Lebesgue measure, in which case $S(x, y)$ becomes the asymmetric squared error considered in [42].

The differences $d_k(\theta) = S_\theta(x_{k1}, y_k) - S_\theta(x_{k2}, y_k)$ of the elementary scores are distinguished by a particular property: $d_k(\theta)$ factorizes into the product

of an *identification function* I that depends only on the observation, times a difference of indicator functions depending only on the forecasts. Specifically, $d_k(\theta) = I(\theta, y_k)\{\mathbb{1}_{\theta < x_{k1}} - \mathbb{1}_{\theta < x_{k2}}\}$ [16, Appendix A3]. We shall utilize this fact to establish weak convergence results for quantile and expectile forecasts complementing those of Propositions 2.2, 3.1 about finite-dimensional distributions.

4.1. Conditional weak convergence of D_n^*

The purpose of this section is to establish the approximation $Pr^*[T(D_n^*) > c_n^*] \approx Pr[T(Z) > c_n^*]$ figuring in the display (3.5) that is central to our argument. The asymptotics involves conditioning on the data x_{k1}, x_{k2}, y_k, so that the sign variables σ_k form the only source of randomness. We thus avoid having to make assumptions about the stochastic structure of the data.

Of basic importance are the second (cross-)moments of the process D_n^*,

$$\gamma_n(\theta_1, \theta_2) = E D_n^*(\theta_1) D_n^*(\theta_2) = n^{-1} \sum\nolimits_{k \leq n} d_k(\theta_1) d_k(\theta_2),$$

$$\rho_n(\theta_1, \theta_2)^2 = E\left(D_n^*(\theta_2) - D_n^*(\theta_1)\right)^2 = n^{-1} \sum\nolimits_{k \leq n} \left(d_k(\theta_2) - d_k(\theta_1)\right)^2,$$

and the continuity moduli of the empirical distributions G_n, F_{n1}, F_{n2} of the observations y_k and the forecasts x_{k1}, x_{k2}, respectively.

Put $m_k = |y_k - x_{k1}| \vee |y_k - x_{k2}|$, and let

$$\begin{aligned} \text{for quantiles:} \quad & H_n = G_n + F_{n1} + F_{n2}\,, \\ \text{for expectiles:} \quad & H_n = F_{n1} + F_{n2}\,. \end{aligned}$$

ASSUMPTIONS.

(C1) (3.3) holds: there exists a function γ such that $\gamma(\theta, \theta) > 0$, $\theta \in \mathbb{R}$, and

$$n^{-1} \sum\nolimits_{k \leq n} d_k(\theta_1)\, d_k(\theta_2) = \gamma_n(\theta_1, \theta_2) \longrightarrow \gamma(\theta_1, \theta_2) \quad (n \to \infty,\ \theta_1, \theta_2 \in \mathbb{R}).$$

(C2) There exist numbers $\kappa \in (0, 1)$, $B > 0$, $n_2 \geq 1$ and a sequence $\beta_n \to 0$ such that

$$\sup\nolimits_{0 \leq \theta_2 - \theta_1 \leq r} H_n([\theta_1, \theta_2]) \leq B\, (r \vee \beta_n)^\kappa, \quad r \in [0, 1],\ n \geq n_2.$$

(C3) $$\sup\nolimits_n n^{-1} \sum\nolimits_{k \leq n} m_k^4 \equiv M < \infty \quad \text{(only for expectiles).}$$

(C4) There exist numbers $\nu > 0$, $A > 0$, and $n_1 \geq 1$ such that

$$(F_{n1} + F_{n2})([-\theta, \theta]^c) \leq A\, \theta^{-\nu}, \quad \theta \geq 1,\ n \geq n_1.$$

Discussion of the assumptions. In practice, the test functionals in (2.6) may be evaluated over a bounded interval Θ in the full θ-domain $\mathbb{R}$ and the score difference processes restricted correspondingly. Of course, forecast dominance

then holds only with respect to all mixtures $S = \int S_\theta \, dM(\theta)$ of scores S_θ with a measure M supported by Θ, which would often be considered as sufficient.

Assumption (C1) is a basic asymptotic stability requirement that would hold 'in probability' under virtually any standard statistical model. However, as in part b) of Proposition 3.1 there is no probability governing the data, hence no convergence in probability (see also Remark 3.1).

The uniform Hölder condition assumed in (C2) requires that the data x_{k1}, x_{k2}, y_k are well dispersed and do not heavily cluster locally. The lower bound $r \geq \beta_n$ on the width of the increments is unavoidable because of the (asymptotically small) jumps of the empirical CDFs.

Assumption (C3) only matters for expectiles. To substantiate it, one may argue that reasonable forecasts should covary with the observations, which would limit the deflections of the quantities m_k. (C3) is stronger than boundedness on average of the m_k^2, which appears as the minimal condition to impose. In return, it implies a Lindeberg type condition holding uniformly in θ,

$$\lim_{n \to \infty} \sup_\theta \, n^{-1} \sum_{k \leq n} d_k(\theta)^2 \, \mathbb{1}_{|d_k(\theta)| > \epsilon \sqrt{n}} = 0 \quad \text{for every } \epsilon > 0. \quad (4.4)$$

Assumption (C4) restrains the large fluctuations of the forecasts x_{k1}, x_{k2} and allows us to control the tail behavior of the functions $\theta \mapsto ED_n^*(\theta)^2$.

Altogether, the assumptions appear weak as well as natural for the quantile and expectile functionals and for continuously distributed data. They only pertain to quantities computable from the data and do not presuppose any statistical model. On the other hand, if a probabilistic model is assumed, they hold with arbitrarily high probability in many of the customary settings. See the corresponding discussion in Section 9, where (C2), (C4) are verified under conventional stationarity assumptions.

Hereafter, ℓ_0^∞ denotes the space of all bounded measurable functions on $\mathbb{R}$ vanishing at infinity equipped with the sup-norm [54]. The sample paths of D_n^* are in ℓ_0^∞ since the elementary scoring functions of quantiles and expectiles are piecewise linear and vanish outside the smallest interval including all forecasts $x_{k\ell}$. In order to avoid problems related to the jumps of D_n^*, *we henceforth assume that the critical values c_n^* are calculated using a continuous version $\bar{D}_n$ of D_n^* obtained by linear interpolation of the D_n^*-values on the grid $\{j\beta_n : j \in \mathbb{Z}\}$, where β_n is as in Assumption (C2). That is, the c_n^* are supposed to satisfy $Pr^* [T(\bar{D}_n) > c_n^*] \approx \alpha$.* Since the grid becomes arbitrarily fine as n gets large this modification is immaterial; see Lemma 8.2.

Theorem 4.1. *Under the assumptions (C1) to (C4) the processes $\{\bar{D}_n(\theta), \theta \in \mathbb{R}\}$ converge weakly in ℓ_0^∞ to a mean zero Gaussian process $\{Z(\theta), \theta \in \mathbb{R}\}$ with covariance function γ and continuous sample paths.*

The theorem states weak convergence under Pr^*, i.e., conditionally on the data. A fortiori, this convergence holds unconditionally as long as the limit covariance function γ, hence the limit process Z, is unique. See Proposition 3.1 for a related discussion.

As a consequence of the theorem, $T(\bar{D}_n)$ converges weakly in distribution to $T(Z)$ for any continuous functional on the space ℓ_0^∞. This covers the supremum statistic $T_\infty(f) = \sup_{\theta \in \mathbb{R}} f(\theta)$ as one special case of interest. Other examples such as the integral type functionals $T_p(f) = \int_{\mathbb{R}} f(\theta)_+^p \, d\theta$ $(p = 1, 2)$ require a sharpening of assumption (C4) for the control of the tail masses.

Corollary 4.1. *Assume (C1) to (C4). Then $T_1(\bar{D}_n)$ converges weakly in distribution to $T_1(Z)$ if the exponent ν in (C4) satisfies $\nu > 2$ in the quantile, and $\nu > 4$ in the expectile case. For the functional T_2 the corresponding conditions are $\nu > 1$ in the quantile, and $\nu > 2$ in the expectile case.*

The conditions on ν can be dropped if the integration in the functionals T_1, T_2 is restricted to a *finite* interval $\Theta \subset \mathbb{R}$ (cf. (2.6)). The simplification occurs because, in this case, the functionals are continuous on the space $\ell^\infty(\Theta)$ of all bounded measurable functions on Θ endowed with the sup-norm, and because weak convergence in ℓ_0^∞ implies weak convergence in $\ell^\infty(\Theta)$.

4.2. Weak convergence of $\widetilde{D}_n$

Here the focus is on the approximation $Pr\,[T(\widetilde{D}_n) > c_n^*] \approx Pr\,[T(\widetilde{Z}) > c_n^*]$ in (3.5), whose verification completes our proof of the validity of the randomization test (apart from the conjecture). The setting is unconditional, so 'Pr' refers to some underlying probability measure governing the joint stochastic behavior of the data triplets. For simplicity we only deal with weak convergence on finite intervals Θ, that is, in $\ell^\infty(\Theta)$.

The necessary distinction between the quantile and the expectile case is a bit tedious. We denote the sequentially conditioned versions of the empirical data distributions as

$$G_n^c(J) = \frac{1}{n} \sum_{k \le n} Pr[y_k \in J \mid \mathcal{F}_{k-1}], \ F_{n\ell}^c(J) = \frac{1}{n} \sum_{k \le n} Pr[x_{k\ell} \in J \mid \mathcal{F}_{k-1}],$$

$\ell = 1, 2$, J an interval, and put as earlier $H_n^c = G_n^c + F_{n1}^c + F_{n2}^c$ in the quantile, and $H_n^c = F_{n1}^c + F_{n2}^c$ in the expectile case. Note that $F_{n\ell}^c = F_{n\ell}$ in the common case of forecasts $x_{k\ell}$ that are $\mathcal{F}_{k-1}$-measurable. The following assumptions are similar to those in the previous section, except that convergence is 'in probabiliy' and expectations are being taken at the appropriate places. A justification of assumption (A2) is given in Section 9.

ASSUMPTIONS.

(A1) (2.8) holds: there exists a function $\widetilde{\gamma}$ such that $\widetilde{\gamma}(\theta, \theta) > 0$ for all θ, and

$$n^{-1} \sum_{k \le n} \widetilde{d}_k(\theta_1) \, \widetilde{d}_k(\theta_2) \ = \ \widetilde{\gamma}_n(\theta_1, \theta_2) \longrightarrow_p \widetilde{\gamma}(\theta_1, \theta_2) \quad (n \to \infty, \ \theta_1, \theta_2 \in \mathbb{R}).$$

Given any $b > 1$ there exists a number $p \ge 1$ such that (A2) and (A3) hold:

(A2)   There are numbers $B > 0$, $n_2 \geq 1$ and a sequence $\beta_n \to 0$ such
       that for both $K_n = H_n$ and $K_n = H_n^c$

$$\sup\nolimits_{0 \leq \theta_2 - \theta_1 \leq r} E K_n([\theta_1, \theta_2])^p \leq B\, (r \vee \beta_n)^b, \quad r \in [0, 1],\ n \geq n_2.$$

(A3) $\sup_n n^{-1} \sum_{k \leq n} E m_k^{4p} < \infty$ (only for expectiles).

As before we consider a continuous version of $\widetilde{D}_n$ obtained by linear in-
terpolation on the grid $\{j\beta_n\ :\ j \in \mathbb{Z}\}$, β_n as in Assumption (A2), which
we denote as $\hat{D}_n$. Accordingly, the approximation to be established becomes
$Pr\left[T(\hat{D}_n) > c_n^*\right] \approx Pr\left[T(\widetilde{Z}) > c_n^*\right]$.

Theorem 4.2. *Assume (A1) to (A3). For every bounded interval $\Theta \subset \mathbb{R}$ the
processes $\{\hat{D}_n(\theta),\ \theta \in \mathbb{R}\}$ converge weakly in $\ell^\infty(\Theta)$ to a mean zero Gaussian
process $\{\widetilde{Z}(\theta),\ \theta \in \mathbb{R}\}$ with covariance function $\widetilde{\gamma}$ and continuous sample paths.*

For bounded Θ the functionals T_∞, T_1, T_2, and in fact all T_p, $1 \leq p \leq \infty$ (see
Equation 2.6) are continuous on $\ell^\infty(\Theta)$. Therefore the desired approximation is
immediate from the following.

Corollary 4.2. *Assume (A1) to (A3). Then for any $1 \leq p \leq \infty$ the random
variable $T_p(\hat{D}_n)$ converges weakly in distribution to $T_p(Z)$.*

5. Monte Carlo simulations

Here we study the randomization test in finite sample scenarios involving mean
(i.e., expectile) and quantile forecasts. The test statistics under examination are
the positive part integrals $T_p(D_n) = \int D_n(\theta)_+^p\, d\theta$, $p = 1, 2$, considered as tests
of the hypothesis H_-^s saying that method 1 dominates method 2 at each time
step. All simulation results are based on $1\,000$ Monte Carlo iterations.

5.1. Mean forecasts

We first present simulation results for the illustrative example from Section 2.2.
One of the variances is fixed, $\tau_1 = 1$, while τ_2 is varied. By Proposition 2.1
the hypothesis H_-^s is satisfied if $\tau_2 \leq 1$, and violated otherwise. We consider
samples of $n = 200$ observations each, which in an economic context is empiri-
cally relevant for quarterly time series data focusing on the postwar period. The
top panel of Figure 1 summarizes our results for the case where the regression
parameter $a = 0.4$; similar results obtain for other values of a. The figure shows
that the performance of the test is quite satisfactory: It comes close to its nom-
inal level 5% at the boundary of the hypothesis ($\tau_2 = 1$), and it is conservative
in its interior ($\tau_2 < 1$), as predicted by the conjectured one-sided Anderson in-
equality. The part of the figure in which $\tau_2 > 1$ yields evidence on the power of
the test. Naturally, we find that the power increases monotonically in τ_2 (i.e.,
clearer violations of the hypothesis imply higher rejection rates). Furthermore,
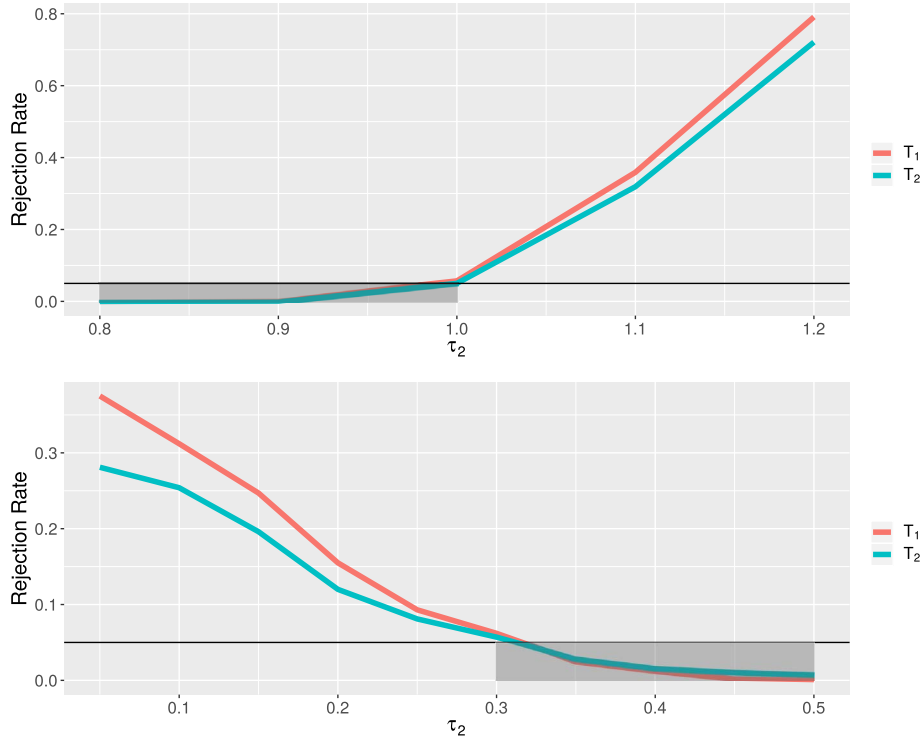the functional T_1 has a slightly higher power than T_2.

FIG 1. *Size and power of the randomization test for mean forecasts (top panel) and quantile forecasts (bottom panel). Rejections are at 5% level, marked by horizontal line. In each panel, the dark gray area indicates the parameter range for which the hypothesis H_-^s is true, such that the rejection rate should be at most 5%. Test statistics T_1 and T_2 are defined at (2.6). Results are based on 1 000 Monte Carlo iterations; within each iteration, the test is computed based on 1 000 simulated sign randomizations. See Sections 5.1 and 5.2 for further details.*

5.2. Quantile forecasts

We take the observations y_k to follow an AR(1)-GARCH(1,1) process in the spirit of [6]:

$$
\begin{aligned}
y_k &= 0.03 + 0.05\, y_{k-1} + s_k \varepsilon_k \\
s_k^2 &= 0.05 + 0.9\, s_{k-1}^2 + 0.05\, s_{k-1}^2 \varepsilon_{k-1}^2
\end{aligned}
$$

with independent "shocks" $\varepsilon_k \sim \mathcal{N}(0,1)$. The parametrization follows [44, Section 3], thereby intending to replicate the empirical features of daily stock returns. Forecasters are asked to state the $\alpha = 0.05$-quantile of the process, conditional on the information $\mathcal{F}_{k-1}$ available up to and including time $k-1$.

To devise a simulation model for two imperfect forecasts, let us first conceive of an oracle. If the oracle knew the data generating mechanism and the initial values s_0, ϵ_0, y_0, she could successively compute s_k from the observations y_j, $j <$

k. Let $\mathcal{F}_k$ denote the σ-algebra generated by the variables s_0, ϵ_0, y_j, $j \leq k$. Then (assuming the regression parameters are known) $y_k \mid \mathcal{F}_{k-1} \sim \mathcal{N}(m_k, s_k^2)$, where we write $m_k = 0.03 + 0.05\, y_{k-1}$ for convenience. Thus for our oracle, the ideal quantile forecast would be the α-quantile of the conditional distribution of y_k, namely $x_{k,ideal} = m_k + s_k z_\alpha$ where $z_\alpha = \Phi^{-1}(\alpha)$ is the ideal forecast in standard units. This leads us to mimic lack of knowledge and forecast errors by assuming that the issued forecasts are of the form $x_{k\ell} = m_k + s_k z_{k\ell}$ ($\ell = 1, 2$) where the $z_{k\ell}$ are random perturbations of z_α that are independent among themselves and from all other variables. Specifically, we assume that the deflections from z_α are Gaussian in the log odds scale,

$$z_{k\ell} = \Phi^{-1}\left[\frac{1}{1 + e^{-\beta - u_{k\ell}}}\right] \quad (\beta = \log\frac{\alpha}{1 - \alpha},\ u_{k\ell} \sim \mathcal{N}(0, \tau_\ell^2),\ \ell = 1, 2). \quad (5.1)$$

Intuitively, forecast 1 should be better than forecast 2 if $\tau_1 < \tau_2$, since the deflections from the ideal forecast are then smaller for forecast 1. It can indeed be shown that H_-^s holds if and only if $\tau_1 \leq \tau_2$; cf. end of Section 9.

In our simulations (bottom panel of Figure 1), $\tau_1 = 0.3$ is fixed, and τ_2 varies from 0.05 to 0.5. The quantile level is $\alpha = 0.05$, and the sample size is 2 000. Both choices are in line with the empirical case study in Section 6.2, where we analyze daily financial return data. Again, as in the previous example, the course of the power as a function of τ_2 supports our claim that the randomization test is approximatively valid for testing H_-^s.

6. Case studies

6.1. Mean forecasts

For a practical application of the randomization test we consider the recession probability forecasts studied in [48], using the updated data set analyzed by [16, Section 4]. The data set covers $n = 186$ quarterly observations from 1968 to 2014, and two competing forecasting methods: Judgmental forecasts from a survey of professional forecasters (SPF), and forecasts from a simple statistical model (Probit). Both forecasts are one quarter ahead, and are out-of-sample.[6] As shown in [16, Figure 6], the survey based forecasts attain better elementary expectile scores for most thresholds $\theta \in [0, 1]$. We specifically consider two test problems where either the survey forecast or the model based forecast dominates the respective other one under the maintained hypothesis.

The top panel of Table 6.1 summarizes the results, which are based on 1 000 simulated sign randomizations. The hypothesis that Probit dominates SPF is rejected at the one percent significance level. By contrast, there is no evidence against the hypothesis that SPF dominates Probit. These results conform with

---

[6]The statistical model is re-estimated recursively at each forecast date in order to mimic a realistic forecast situation. The forecast data set is available within the R package murphy-diagram [29].

those of [16] who consider informal (pointwise) confidence intervals. Remarkably, the randomization test here proves powerful enough to yield interpretable conclusions in a relatively small data set.

TABLE 1
Randomization test results for recession probability forecasts (top panel) and quantile forecasts of stock returns (bottom panel).

Mean forecasts (recession probabilities)

Hypothesis (H_-)	Test statistic	P-value
SPF dominates Probit	T_1	0.987
SPF dominates Probit	T_2	0.986
Probit dominates SPF	T_1	0.010
Probit dominates SPF	T_2	0.002

Quantile forecasts (stock returns, $\alpha = 0.05$)

Hypothesis (H_-)	Test statistic	P-value
QR_{RV} dominates QR_{ABS}	T_1	0.646
QR_{RV} dominates QR_{ABS}	T_2	0.916
QR_{ABS} dominates QR_{RV}	T_1	0.000
QR_{ABS} dominates QR_{RV}	T_2	0.000

6.2. Quantile forecasts

In a second case study we consider quantile forecasts of daily returns y_k of the Dow Jones Industrial Average (DJIA), using data that is freely available at http://realized.oxford-man.ox.ac.uk/. Quantiles at low levels α are commonly used as measures for financial risk, and are referred to as Value-at-Risk at level α (e.g. [40], Sections 1 and 2). We specifically consider prediction of the five percent quantile of y_k, given information until the previous business day $k-1$. In a first specification, which we denote by QR_{RV}, the predicted quantile is given by

$$x_{k1} = \hat{\beta}_0 + \hat{\beta}_1 |RV_{k-1}|, \tag{6.1}$$

where RV_{k-1} is the so-called realized volatility computed from intra-daily data (e.g. [1]). We obtain parameter estimates $\hat{\beta}_0, \hat{\beta}_1$ by quantile regression [30], based on a rolling window of 2 000 observations.[7] Recent evidence [60] suggests that the specification in (6.1) compares favorably to a number of more complicated alternatives. Our second specification (QR_{ABS}) is analogous to (6.1), except that it employs the lagged absolute return $|y_{k-1}|$ in place of realized volatility $|RV_{k-1}|$. The two specifications are motivated by the fact that realized volatility

---

[7]The first rolling window ends on November 19, 2008; the last window ends on May 4, 2017. We use each rolling window to compute parameter estimates and form a forecast for the next business day. Our data set thus covers 1 964 forecast/realization pairs that we use for evaluation. Our implementation of quantile regression is based on the function rq from the R package quantreg [31].

and absolute returns proxy for the variability of financial returns, which is well known to fluctuate over time (cf. Section 5.2). Both measures should thus be informative about the quantiles of y_k, given $\mathcal{F}_{k-1}$.

The bottom panel of Table 6.1 presents the results of the comparison. We find no evidence against the hypothesis that QR_{RV} dominates QR_{ABS}; however, we clearly reject the converse hypothesis that QR_{ABS} dominates QR_{RV}. This suggests that intra-daily information encoded in realized volatility contains more predictive content than daily returns.[8] Similar conclusions were reached in [58]. As in the first case study, the results are qualitatively robust across the two test statistics T_1, T_2. In summary, the Monte Carlo simulations and the case studies point to the potential usefulness of the proposed randomization test.

7. Discussion

We have studied randomization type tests of hypotheses implying that a quantile or expectile forecast is superior to a competitor, uniformly across all (or a large class of) consistent scoring functions. Variants of this topic recently have gained considerable interest, particularly in the econometrics literature. Tests of dominance relations in quantile and expectile forecasts are studied in [56] using the bootstrap, while in [58] the authors focus on the so-called expected shortfall functional, relying on a combination of pointwise tests and multiple testing corrections. These two papers are closest to the present work in that they base forecast comparisons on consistent scoring functions – arguably the proper concept for this purpose [21] – and their mixture representation [16]. We next provide a more detailed discussion of related literature.

Dominance concept. Our dominance concept aims at the direct comparison of forecasts on a purely empirical basis. This is distinct from more theoretically oriented concepts like nested or overlapping information sets [26, 32], or the notion that forecast x_1 'encompasses' forecast x_2 [18, Section 17.1]. In a nutshell, forecast encompassing asks whether a user who has access to both forecasts may safely ignore x_2, judging from some given criterion. By contrast, our notion of dominance asks whether a user who must choose between the forecasts would prefer x_1 over x_2, in regard to any admissible criterion. Tests of *stochastic* dominance are considered in [28, 38, 39]. Analogously to forecast dominance, stochastic dominance stands for superiority (of some procedure compared to another) that holds uniformly across a whole class of criteria. In stochastic dominance, this class comprises those functions of e, the residuals or forecast errors from some regression type model, that increase as e moves away from zero, or else, are convex in e. In forecast dominance as here, the relevant criteria are the consistent scoring functions, or some subclass thereof.

Modeling assumptions. Our stance is to try to avoid assumptions about possible data generating mechanisms as far as possible, on the grounds given in the introduction. For a similar view see [20] and [28, p. 1308 and Sect. 5], where

---

[8]We obtain the same result when using the lagged squared return, y_{k-1}^2, rather than the lagged absolute return, $|y_{k-1}|$, in the second specification.

the common stationarity assumption is weakened to distributional heterogeneity. Consequently our framework puts no restrictions on the type of forecasts or their connection with the observations, and it allows for great freedom regarding their dynamics. The lack of an explicit statistical model and the need to harmonize the underlying statistics with the surrogate external randomization lead us to state forecast dominance hypotheses in terms of *conditional* expectations expressing conditional predicitive ability, as proposed in [20]. The step-by-step character of the forecast scheme renders this an attractive, natural alternative to the familiar formulations using unconditional expectations; see Example 2.1 or the simulation model in Section 5.2. We note, however, that our hypotheses differ from those of [20]: there the focus is on single-criterion two-sided (equality) testing, whereas in the present paper the focus is on simultaneous one-sided (inequality) testing. In the setup of [20], simple least squares regressions yield attractive tests for the null-hypothesis of equal conditional predictive ability that are not available in our case.

Test procedures. The idea of external randomization has long played an important role in statistics (see e.g. [15, Chapter 4.4], as well as [9] on the related idea of permutation), and it is central to our approach. Randomization for the task of forecast comparison was used in, e.g., [13, 14], and it has been applied to ensure that a well-defined limiting distribution exists in the first place [4, 10, 53]. Our test construction is similar in spirit to the 'conditional p-value' approach of [24] and the 'wild bootstrap' proposed in [35]. In the context of functional data, the use of maximum and integral type test statistics like ours is standard; for weighted versions cf. [39]. Since the limit distributions are unknown, the determination of critical values makes it necessary to resort to resampling methods, customarily various forms of (block [33]) bootstrap as in [28, 38, 39, 56]. While this is often the method of choice, its application in the present one-sided, high-dimensional context is not without problems. These partly are due to the nonstandard asymptotics of the bootstrap-based tests resulting from degenerate limit processes; see e.g. [28, 39, 56]. Noteworthily at this point, the Gaussian limit processes in our setting are entirely regular.

Test size control. Intuition and classical test theory suggest that in order to control the error of the first kind, it might suffice to control it on the boundary of the hypothesis. Unfortunately, what constitutes the boundary is subtle, and a focus on least favorable cases is inadequate [25]. For an extensive discussion of these points in a different framework (stochastic dominance) see Linton et al. [39], who emphasize the importance, and difficulty, of a uniform control of the test size (see also [28, p. 1320]) and develop a sophisticated bootstrap procedure for this purpose (in the i.i.d. case). Still, even there uniformity is achieved only if certain subsets of the hypothesis are excluded. We are actually not aware of any fully satisfactory result in this regard; neither is the issue clarified in the present paper. However, our approach suggests a potentially elegant solution at least: if the one-sided Anderson's inequality were true, our tests would be valid uniformly on H_-.[9] The discussion in Section 3.2 elaborates the central role of our

---

[9]To be read as: uniformly on those parts of H_- where the weak convergence in Theorems

corresponding conjecture. A resolution of the issue, whether in the positive or in the negative, would certainly be of great interest. We may point out, however, that independently of the final status of the conjecture, the randomization test does behave properly for probabilities that are contiguous to the strict null-hypothesis H_0^s; cf. Remark 3.3.

8. Proofs

Proof of Proposition 2.1. We may refer to [32, Example 2.3], wherein our earlier proof is much simplified. To see the relevance of Theorem 2.1 of that paper for the elementary scores note that the mean value is an $\alpha = 1/2$-expectile whence eq. (4.3) assumes the form $S_\theta = \varphi(y) - \varphi(x) - (y - x)\varphi'(x)$ with the convex function $\varphi(x) = \frac{1}{2}(x - \theta)_+$. $\qquad\square$

Proof of Proposition 2.2. Given $\theta_1, \ldots, \theta_m \in \mathbb{R}$, $c_1, \ldots, c_m \in \mathbb{R}$, put $X_n = \sum_j c_j \widetilde{D}_n(\theta_j)$ and $V = \sum_{i,j} c_i c_j \widetilde{\gamma}(\theta_i, \theta_j)$. It suffices to show that the distribution of X_n converges to $\mathcal{N}(0, V)$. We have $X_n = \sum_{k=1}^n X_{kn}$ where

$$X_{kn} = \sum_j c_j \, n^{-1/2} \, \widetilde{d}_k(\theta_j), \quad k = 1, \ldots, n.$$

In order to apply [36, Theorem 2.3] to the martingale difference array $\{X_{kn}\}$, we note at first that (2.8) implies

$$\sum_{k \le n} X_{kn}^2 = \sum_{i,j} c_i c_j \, n^{-1} \sum_{k \le n} \widetilde{d}_k(\theta_i) \, \widetilde{d}_k(\theta_j) \longrightarrow_p \sum_{i,j} c_i c_j \, \widetilde{\gamma}(\theta_i, \theta_j) = V. \tag{8.1}$$

Thus if

$$E\left(\max_{k \le n} X_{kn}^2\right) \to 0 \quad \text{as } n \to \infty \tag{8.2}$$

holds, the two conditions (a), (b) of [36, Theorem 2.3] are satisfied, and in view of (8.1) we are done. (We may asssume $V = 1$.) Now

$$\max_{k \le n} X_{kn}^2 \le \frac{2m}{n} \sum_j c_j^2 \left\{ \max_{k \le n} d_k(\theta_j)^2 + \max_{k \le n} E[d_k(\theta_j)^2 \mid \mathcal{F}_{k-1}] \right\}$$

by Jensen's inequality, and since m and the c_j are fixed, it suffices to show that n^{-1} times the expectation of the two maxima in curly brackets tends to zero for every j. Let $\epsilon > 0$. For any θ we have

$$\begin{aligned}
&E\left(\max_{k \le n} n^{-1} E[d_k(\theta)^2 \mid \mathcal{F}_{k-1}]\right) \\
&\le E\left(\max_{k \le n} n^{-1} E[d_k(\theta)^2 \mathbb{1}_{|d_k(\theta)| > \epsilon\sqrt{n}} \mid \mathcal{F}_{k-1}]\right) \\
&\quad + E\left(\max_{k \le n} n^{-1} E[d_k(\theta)^2 \mathbb{1}_{|d_k(\theta)| \le \epsilon\sqrt{n}} \mid \mathcal{F}_{k-1}]\right) \\
&\le n^{-1} E\left(\sum_{k \le n} E[d_k(\theta)^2 \mathbb{1}_{|d_k(\theta)| > \epsilon\sqrt{n}} \mid \mathcal{F}_{k-1}]\right) + \epsilon^2 \\
&= n^{-1} \sum_{k \le n} E\{d_k(\theta)^2 \mathbb{1}_{|d_k(\theta)| > \epsilon\sqrt{n}}\} + \epsilon^2.
\end{aligned}$$

4.1, 4.2 holds uniformly. Clearly, such restrictions always apply when asymptotics is involved.

The same upper bound holds for $E\left(n^{-1}\max_{k\leq n}d_k(\theta)^2\right)$. Since ϵ was arbitrary, (8.2) follows by assumption (C0). □

Proof of Proposition 3.1. The proof follows the same lines as the proof of Proposition 2.2. It suffices to replace $\widetilde{d}_k(\theta)$ by $d_k(\theta)\sigma_k$, define $\mathcal{F}_k$ as the σ-algebra generated by the random variables $\sigma_1,\ldots,\sigma_k$, and observe that $|\sigma_k|=1$ and $E\left[d_k(\theta)\sigma_k\mid\mathcal{F}_{k-1}\right]=0$. See Remark 3.1. □

Toward the proofs of Theorem 4.1 and Corollary 4.1, we assume throughout that (C1) to (C4) are fulfilled. We begin with some first consequences of the assumptions. Constants generally depend on whether they refer to quantiles or expectiles, which is indicated by subscripts q, e, respectively. Recall that expectations and probabilities are conditional, and refer to the random signs only. For simplicity we nevertheless use the "un-starred" symbols E, Pr.

Lemma 8.1. *(i) There are constants $L_{q,e}$ such that*

$$\sup_n\int_{-\infty}^{\infty}ED_n^*(\theta)^2d\theta\leq L_{q,e}.\tag{8.3}$$

(ii) There are constants $A_{q,e}$ and $\nu_q=\nu$, $\nu_e=\nu/2$ (cf. (C4)) such that

$$ED_n^*(\theta)^2\leq A_{q,e}\,|\theta|^{-\nu_{q,e}},\ |\theta|\geq1,\ n\geq n_1\,.\tag{8.4}$$

(iii) There are constants $B_{q,e}$ and $\lambda_q=\kappa/2$, $\lambda_e=\kappa/4$ (cf. (C2)) such that

$$\omega_n(r)=\sup_{0\leq\theta_2-\theta_1\leq r}\rho_n(\theta_1,\theta_2)\leq B_{q,e}\,(r\vee\beta_n)^{\lambda_{q,e}},\ r\in[0,1],\ n\geq n_2.\tag{8.5}$$

Proof. For generalized quantiles the individual score differences are of the form

$$d_k(\theta)=I(\theta,y_k)\,\delta_k(\theta),\qquad\delta_k(\theta)=\mathbb{1}_{\theta<x_{k1}}-\mathbb{1}_{\theta<x_{k2}}\tag{8.6}$$

where $I(\theta,y_k)$ is the respective identification function. Specifically for α-quantiles, the identification function is $I(\theta,y)=\mathbb{1}_{y\leq\theta}-\alpha$, whence $|d_k(\theta)|\leq|\delta_k(\theta)|$. For α-expectiles, $I(\theta,y)=(1-\alpha)\,(\theta-y)_+-\alpha\,(y-\theta)_+$, whence

$$|d_k(\theta)|\leq|y_k-\theta|\,|\delta_k(\theta)|\leq\{|y_k-x_{k1}|\vee|y_k-x_{k2}|\}\,|\delta_k(\theta)|=m_k\,|\delta_k(\theta)|\,.\tag{8.7}$$

The second inequality is easily seen to follow from the fact that $|\delta_k(\theta)|$ equals one if θ lies between x_{k1} and x_{k2}, and is zero otherwise. This observation also shows that $\int\delta_k(\theta)^2d\theta=|x_{k1}-x_{k2}|\leq2m_k$, whence by Hölder's inequality and (C3)

$$\int ED_n^*(\theta)^2d\theta\ =\ n^{-1}\sum_{k\leq n}\int d_k(\theta)^2d\theta\leq n^{-1}\sum_{k\leq n}2(m_k)^{s+1}\leq2M^{\frac{s+1}{4}}$$

where $s=0$ and $s=2$ for quantiles and expectiles, respectively, which is (i).

Similarly, if $|\theta|\geq1$, $n\geq n_1$, using (C4) we get for quantiles

$$ED_n^*(\theta)^2\leq n^{-1}\sum_{k\leq n}|\delta_k(\theta)|\leq(F_{n1}+F_{n2})([-|\theta|,|\theta|]^c)\leq A\,|\theta|^{-\nu}\,,$$

while for expectiles, (8.7), (C3), and Cauchy-Schwarz give

$$
\begin{aligned}
ED_n^*(\theta)^2 \;=\; n^{-1}\sum\nolimits_{k\le n} d_k(\theta)^2 &\le \left\{ \left[n^{-1}\sum\nolimits_{k\le n} m_k^4 \right]\left[n^{-1}\sum\nolimits_{k\le n} |\delta_k(\theta)| \right] \right\}^{1/2} \\
&\le \{MA\,|\theta|^{-\nu}\}^{1/2},
\end{aligned}
\tag{8.8}
$$

which settles (ii). As for the increments, let $\theta_1 < \theta_2$ and put $\delta_k(\theta_1,\theta_2) = \delta_k(\theta_2) - \delta_k(\theta_1)$. Writing

$$
d_k(\theta_2) - d_k(\theta_1) = \{I(\theta_2,y_k) - I(\theta_1,y_k)\}\delta_k(\theta_i) + I(\theta_j,y_k)\delta_k(\theta_1,\theta_2)
$$

with either $i=1, j=2$ or $i=2, j=1$, whichever is more convenient, we get for α-quantiles

$$
|d_k(\theta_2) - d_k(\theta_1)| \le \mathbb{1}_{\theta_1 < y_k \le \theta_2} + |\delta_k(\theta_1,\theta_2)|,
\tag{8.9}
$$

and for α-expectiles

$$
\begin{aligned}
&|d_k(\theta_2) - d_k(\theta_1)| \tag{8.10}\\
&\le |\theta_2 - \theta_1|(|\delta_k(\theta_1)| \vee |\delta_k(\theta_2)|) + \{|y_k - \theta_1| \wedge |y_k - \theta_2|\}\,|\delta_k(\theta_1,\theta_2)|\\
&\le |\theta_2 - \theta_1|(|\delta_k(\theta_1)| \vee |\delta_k(\theta_2)|) + \{|y_k - x_{k1}| \vee |y_k - x_{k2}|\}\,|\delta_k(\theta_1,\theta_2)|.
\end{aligned}
$$

The last inequality may be verified similarly as at (8.7) on observing that $|\delta_k(\theta_1,\theta_2)| = 1$ if exactly one of x_{k1}, x_{k2} lies in the interval $[\theta_1,\theta_2)$, and is zero otherwise. This observation also shows that

$$
n^{-1}\sum\nolimits_{k\le n} |\delta_k(\theta_1,\theta_2)| \le F_{n1}([\theta_1,\theta_2)) + F_{n2}([\theta_1,\theta_2)).
\tag{8.11}
$$

For quantiles we then get by (8.9)

$$
\begin{aligned}
\rho_n(\theta_1,\theta_2)^2 &\le 2\{G_n((\theta_1,\theta_2]) + F_{n1}([\theta_1,\theta_2)) + F_{n2}([\theta_1,\theta_2))\}\\
&\le 2H_n([\theta_1,\theta_2]),
\end{aligned}
\tag{8.12}
$$

while for expectiles the estimates (8.10), (8.11) and Cauchy-Schwarz give similarly as at (8.8)

$$
\rho_n(\theta_1,\theta_2)^2 \le 2\,(\theta_2 - \theta_1)^2 + 2\,\{MH_n([\theta_1,\theta_2]\}^{1/2}.
\tag{8.13}
$$

Assertion (iii) thus follows from (C2).   □

Lemma 8.2. *Up to adjustments of the constants, the assertions of Lemma 8.1 also hold for the interpolated processes $\bar{D}_n$, with the following improvement of (iii):*

$$
\sup\nolimits_{0\le \theta_2 - \theta_1 \le r} E(\bar{D}_n(\theta_2) - \bar{D}_n(\theta_1))^2 \le C_{q,e}\, r^{\lambda_{q,e}} \quad (r \in [0,1],\ n \ge n_2)
\tag{8.14}
$$

(i.e., with $r^{\lambda_{q,e}}$ rather than $(r \vee \beta_n)^{\lambda_{q,e}}$). Furthermore,

$$
\lim\nolimits_{n\to\infty}\sup\nolimits_\theta E(\bar{D}_n(\theta) - D_n^*(\theta))^2 = 0,
\tag{8.15}
$$

$$
\lim\nolimits_{n\to\infty} E\int (\bar{D}_n(\theta) - D_n^*(\theta))^2 d\theta = 0.
\tag{8.16}
$$

Proof. For convenience we intermediately write $\beta_n \equiv \epsilon$. Given θ, there is exactly one $\ell \in \mathbb{Z}$ and $w \in [0,1)$ such that $\theta = (1-w)\ell\epsilon + w(\ell+1)\epsilon$. By Jensen's inequality

$$
\begin{aligned}
& E(\bar{D}_n(\theta) - D_n^*(\theta))^2 \\
\le\ & n^{-1} \sum_{k \le n} \left\{ w\,[d_k((\ell+1)\epsilon) - d_k(\theta)]^2 + (1-w)[d_k(\theta) - d_k(\ell\epsilon)]^2 \right\} \\
\le\ & \omega_n(\epsilon) \equiv \omega_n(\beta_n),
\end{aligned}
$$

which proves (8.15). Turning to (8.16), let us write $\Delta_k(\theta)$ for the k-th term in the above sum. We first consider the quantile case. Recalling that ℓ and w are uniquely determined by θ we get by (8.9),

$$
\Delta_k(\theta) \le 2\,\mathbb{1}_{\ell\epsilon < y_k \le (\ell+1)\epsilon} + 2\,|\delta_k(\ell\epsilon, (\ell+1)\epsilon)|.
$$

The right-hand side is always ≤ 4, and it vanishes except if both θ and any of y_k, x_{k1}, or x_{k2} lie in the interval $[\ell\epsilon, (\ell+1)\epsilon)$. Thus for fixed k there are at most 3 intervals of length ϵ on which the function $\theta \mapsto \Delta_k(\theta)$ is non-zero. Consequently, $\int \Delta_k(\theta)d\theta \le 12\epsilon$, so taking the average over k settles the quantile case. A similar reasoning applies in the expectile case. By (8.10),

$$
\Delta_k(\theta) \le 2\,\epsilon^2 \{|\delta_k(\ell\epsilon)| \vee |\delta_k((\ell+1)\epsilon)|\} + 2\,m_k^2\,|\delta_k(\ell\epsilon, (\ell+1)\epsilon)|.
$$

The term $|\delta_k(\ell\epsilon)| \vee |\delta_k((\ell+1)\epsilon)| \le 1$ is nonzero at most if $\theta \in [x_{k1} \wedge x_{k2} - \epsilon,\ x_{k1} \vee x_{k2} + \epsilon]$. Therefore

$$
\int \Delta_k(\theta)d\theta \le 2\,\epsilon^2\,(|x_{k2} - x_{k1}| + 2\epsilon) + 4\,m_k^2\,\epsilon \le 2\,\epsilon^2\,(2m_k + 2\epsilon) + 4\,m_k^2\,\epsilon,
$$

so averaging over k and using (C3) gives (8.16).

Straightforward estimates yield the uniform Hölder condition (8.14) at first for points θ_1, θ_2 belonging to the same interval $[\ell\epsilon, (\ell+1)\epsilon)]$, then belonging to two adjacent intervals, finally for points with one or more such intervals in between, where we may apply (8.5). The analogs of assertions 1 and 2 in Lemma 8.1 are obvious. $\qquad\square$

Proof of Theorem 4.1. Convergence of the finite-dimensional distributions being clear from Proposition 3.1, (4.4), and (8.15), we only need to prove (stochastic) asymptotic equicontinuity [45, 54] and the uniform vanishing at infinity of the sample paths of $\bar{D}_n$. Without loss of generality we may assume $n \ge n_1 \vee n_2$ (cf. (C4), (C2)). Distinguishing between quantiles and expectiles is not necessary here, so we omit the subscripts q, e in the quantities appearing in Lemma 8.1 and 8.2. Moreover, by Lemma 8.2 quantities initially referring to D_n^* such as ρ_n or ω_n may also be used with $\bar{D}_n$, with the same bounds.

Let $u > 0$. For any set $T_0 \subset \mathbb{R}$, let $N_n(u, T_0)$ denote the minimal cardinality of a subset $T \subset T_0$ such that $\min_{t \in T} \rho_n(\theta, t) \le u$ for every $\theta \in T_0$. Given $b > 1$, pick $t_j \in [-b, b]$ equidistant with spacing $r = 2(u/B)^{1/\lambda}$. By (8.14), the minimal ρ_n-distance of any $\theta \in [-b, b]$ to the resulting set T is $\le \omega_n(r/2) \le u$,

whence $N_n(u, [-b, b]) \le Kbu^{-1/\lambda}$. Here and subsequently we write K for any independent finite constant, whose value may thus change from instance to instance.

By (8.4) and Lemma 8.2 there is $\nu > 0$ such that

$$\rho_n(\theta, b)^2 \le 2\{E\bar{D}_n(\theta)^2 + E\bar{D}_n(b)^2\} \le (Kb^{-\nu})^2, \quad \theta > b, \tag{8.17}$$

and similarly for $\theta < -b$. Therefore, with $b = (K/u)^{1/\nu}$ we have $\min_{t \in T} \rho_n(\theta, t) \le u$ for every $\theta \in \mathbb{R}$ and thus

$$N_n(u, \mathbb{R}) \le Ku^{-1/\nu - 1/\lambda}, \quad u > 0. \tag{8.18}$$

Let

$$\Omega_n(r) = \sup\{|\bar{D}_n(\theta_2) - \bar{D}_n(\theta_1)| : \rho_n(\theta_1, \theta_2) \le r, \theta_1, \theta_2 \in \mathbb{R}\}, \ r > 0.$$

By [59, Lemma 1.2] applied with $x_i(s) \equiv d_k(\theta)/\sqrt{n}$ and $p = 1$ we have

$$E\,\Omega_n(r) \le K \int_0^{r/4} (\log N_n(u, \mathbb{R}))^{1/2}\,du \le K \int_0^r (\log u^{-1/\tau - 1/\lambda})^{1/2}\,du$$

for all $r \in [0, 1]$. Therefore $E\,\Omega_n(r_n) \to 0$ if $r_n \to 0$, which implies asymptotic equicontinuity on $\mathbb{R}$ with respect to the semi-metrics ρ_n.

There are two further consequences. First, we already know that for every $\eta > 0$ there is $b > 0$ such that $E\bar{D}_n(b)^2 \le \eta^2$ and $\rho_n(b, \theta) \le \eta$ for every $\theta \in (b, \infty)$. Thus

$$|\bar{D}_n(\theta)| \le |\bar{D}_n(b)| + |\bar{D}_n(\theta) - \bar{D}_n(b)| \le |\bar{D}_n(b)| + \Omega_n(\eta)$$

and so

$$\sup_{|\theta| > b, \, \theta \in \mathbb{R}} |\bar{D}_n(\theta)| = o_p(1) \quad \text{as} \quad b \to \infty.$$

Secondly, by (8.14)

$$\begin{aligned}\widetilde{\Omega}_n(r) &= \sup\{|\bar{D}_n(\theta_2) - \bar{D}_n(\theta_1)| : |\theta_1 - \theta_2| \le r, \theta_1, \theta_2 \in \mathbb{R}\} \\ &\le \Omega_n(\omega_n(r)) \le \Omega_n(Kr^\lambda),\end{aligned}$$

whence $E\,\widetilde{\Omega}_n(r_n) \to 0$ if $r_n \to 0$, implying asymptotic equicontinuity also with respect to the standard metric. It follows that the processes $\bar{D}_n$ converge weakly in ℓ_0^∞ to the specified Gaussian process Z, which by the asymptotic equicontinuity can be assumed to have continuous sample paths. $\square$

Proof of Corollary 4.1. By (8.4) and Lemma 8.2, $\int_{|\theta| > b} E\bar{D}_n(\theta)^2\,d\theta \to 0$ as $b \to \infty$ under the given conditions. Consequently, $T_2(\bar{D}_n)$ equals $\int_{|\theta| \le b} \bar{D}_n(\theta)_+^2\,d\theta$ up to the arbitrarily small contribution from the tails, and weak convergence follows by Theorem 4.1. The same argument, up to an application of Jensen's inequality, applies to T_1. $\square$

Proof of Theorem 4.2. To prove stochastic equicontinuity we use the classical Kolmogorov moment criterion. In view of the linear interpolation it suffices to show that there exist positive constants ξ, K, and $\eta > 1$, such that for all θ_1, θ_2 in the grid $\{j\beta_n : j \in \mathbb{Z}\}$ one has

$$E\,|\widetilde{D}_n(\theta_2) - \widetilde{D}_n(\theta_1)|^\xi \le K|\theta_2 - \theta_1|^\eta. \tag{8.19}$$

Let such a pair θ_1, θ_2 be fixed. Since the partial sums

$$S_k = n^{-1/2}\sum\nolimits_{j \le k}(\widetilde{d}_j(\theta_2) - \widetilde{d}_j(\theta_1)), \quad k = 1, \ldots, n$$

represent a martingale with respect to the filtration $\{\mathcal{F}_k\}$, Burkholder's inequality [7, Theorem 9] gives

$$
\begin{aligned}
E\,|S_n|^{2p} &= E|\widetilde{D}_n(\theta_2) - \widetilde{D}_n(\theta_1)|^{2p} \\
&\le N_p\,E\left\{n^{-1}\sum\nolimits_{k \le n}(\widetilde{d}_k(\theta_2) - \widetilde{d}_k(\theta_1))^2\right\}^p
\end{aligned}
$$

for any $p \ge 1$, with a universal constant N_p. Now

$$
\begin{aligned}
\widetilde{\rho}_n(\theta_1, \theta_2)^2 &:= n^{-1}\sum\nolimits_k(\widetilde{d}_k(\theta_2) - \widetilde{d}_k(\theta_1))^2 \\
&\le 2\rho_n(\theta_1, \theta_2)^2 + 2n^{-1}\sum\nolimits_k\{E\,[\,|d_k(\theta_2) - d_k(\theta_1)|\mid\mathcal{F}_{k-1}]\}^2 \tag{8.20}
\end{aligned}
$$

which may be further estimated as in the proof of Lemma 8.1. We first consider the expectile case. Putting $\delta_k = \delta_k(\theta_1, \theta_2)$ we get from (8.10) that the last term is bounded by a constant times the sum of $(\theta_2 - \theta_1)^2$ plus the term

$$
\begin{aligned}
n^{-1}\sum\nolimits_k\{E[\,m_k\,|\delta_k|\mid\mathcal{F}_{k-1}]\}^2 &\le n^{-1}\sum\nolimits_k E[\,m_k^2\mid\mathcal{F}_{k-1}]\,E[\,\delta_k^2\mid\mathcal{F}_{k-1}] \\
&\le \left\{n^{-1}\sum\nolimits_k(E[\,m_k^2\mid\mathcal{F}_{k-1}])^2\right\}^{1/2}\left\{n^{-1}\sum\nolimits_k E[\,|\delta_k|\mid\mathcal{F}_{k-1}]\right\}^{1/2} \\
&\le \left\{n^{-1}\sum\nolimits_k E[\,m_k^4\mid\mathcal{F}_{k-1}]\right\}^{1/2}\{H_n^c([\theta_1, \theta_2])\}^{1/2}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
E\,\widetilde{\rho}_n(\theta_1, \theta_2)^{2p} &\le K\left[|\theta_2 - \theta_1|^{2p} + E\left\{\left(n^{-1}\sum\nolimits_k m_k^4\right)H_n([\theta_1, \theta_2])\right\}^{p/2}\right. \\
&\qquad\qquad\left. + E\left\{\left(n^{-1}\sum\nolimits_k E[\,m_k^4\mid\mathcal{F}_{k-1}]\right)H_n^c([\theta_1, \theta_2])\right\}^{p/2}\right] \\
&\le K\left[|\theta_2 - \theta_1|^{2p} + \left\{E\left(n^{-1}\sum\nolimits_k m_k^4\right)^p E\,H_n([\theta_1, \theta_2]^p\right\}^{1/2}\right. \\
&\qquad\qquad\left. + \left\{E\left(n^{-1}\sum\nolimits_k E[\,m_k^4\mid\mathcal{F}_{k-1}]\right)^p E\,H_n^c([\theta_1, \theta_2]^p\right\}^{1/2}\right] \\
&\le K\left[|\theta_2 - \theta_1|^{2p} + \sqrt{M_{4p}}\left(E\,H_n([\theta_1, \theta_2]^p + E\,H_n^c([\theta_1, \theta_2]^p)^{1/2}\right]
\end{aligned}
$$

where $M_{4p} = n^{-1}\sum_k E\,m_k^{4p}$. So given $\eta > 1$, putting $b = 2\eta$ we may choose $p \ge 1$ in (A2), (A3) such that (8.19) is satisfied with $\xi = 2p$. This settles the expectile case. The quantile case can be dealt with similarly starting from (8.20). Given $\eta > 1$ one puts $b = \eta$ and uses (8.9), (8.12), then (A2), (A3). $\qquad\square$

9. Additional material

Lemma 9.1. *Let $h_k(\theta) = E\left[d_k(\theta)\,|\,\mathcal{F}_{k-1}\right]$. Under the conditions (2.8), (3.3), and (C3) we have*

$$\gamma = \widetilde{\gamma} + \psi \quad where \quad \psi(\theta_1, \theta_2) = p\text{-}\lim\, n^{-1} \sum\nolimits_{k \le n} h_k(\theta_1) h_k(\theta_2). \qquad (9.1)$$

Proof. By (2.8), (3.3), and

$$\begin{aligned}
\gamma_n(\theta_1, \theta_2) &= n^{-1} \sum\nolimits_{k \le n} \{\widetilde{d}_k(\theta_1) + h_k(\theta_1)\}\,\{\widetilde{d}_k(\theta_2) + h_k(\theta_2)\} \\
&= \widetilde{\gamma}_n(\theta_1, \theta_2) + n^{-1} \sum\nolimits_{k \le n} h_k(\theta_1) h_k(\theta_2) \\
&\quad + n^{-1} \sum\nolimits_{k \le n} h_k(\theta_2) \widetilde{d}_k(\theta_1) + n^{-1} \sum\nolimits_{k \le n} h_k(\theta_1) \widetilde{d}_k(\theta_2)
\end{aligned}$$

it suffices to show that e.g. the last term, to be denoted R_n, tends to zero in quadratic mean. But $ER_n = 0$ because $E\left[\widetilde{d}_k(\theta_2)\,|\,\mathcal{F}_{k-1}\right] = 0$ and $h_k(\theta_1)$ is $\mathcal{F}_{k-1}$-measurable. Similarly, $ER_n^2 \to 0$: the off-diagonal terms in the double sum vanish, and by Jensen's and Cauchy's inequalities and (C3) the sum of the diagonal terms is $O(n)$. □

Justification of (C2), (C4). We will show that the conditions (C2), (C4) are satisfied with probability arbitrarily close to one under common probability models for the data. Possible dependencies within the triplets (x_{k1}, x_{k2}, y_k) do not matter because (C2), (C4) effectively pertain to the marginal CDFs F_{n1}, F_{n2}, G_n only. However, it is natural in the prediction setting to allow for serial dependence. Specifically, suppose that the predictions x_{k1}, x_{k2} and the observations y_k each form a strictly stationary sequence, defined for all $k \in \mathbb{Z}$.

CONDITION (C2). To verify (C2) for the empirical CDF G_n of the observations, e.g., we may invoke an estimate by W. B. Wu applying to certain causal processes of the form $y_k = J(\cdots, \epsilon_{k-1}, \epsilon_k)$, where J is measurable and the ϵ_k, $k \in \mathbb{Z}$ are i.i.d. random variables. As an immediate consequence of [55, Theorem 2] one has, under the conditions given there, that

$$E\left[\sup\nolimits_{0 \le t-s \le r} (G_n(t) - G_n(s))^2\right] = O(n^{-1}\, r^{1-2/q}), \qquad r \ge \beta_n,$$

where $2 < q < 4$ and β_n is a sequence tending to zero sufficiently slowly (not faster than $n^{-1}\,(\log n)^{2q/(q-2)}$). Markov's inequality then gives

$$Pr\left[\sup\nolimits_{0 \le t-s \le r} |G_n(t) - G_n(s)| \ge Kr^\kappa\right] = O(n^{-1}\, r^{1-2/q-2\kappa}), \quad r \ge \beta_n,\, K > 0.$$

Putting $r = 2^{-\ell}$ and summing over $\ell = 0, 1, \ldots$ we find that for any positive $\kappa < 1/2 - 1/q$ we have with probability $1 - O(n^{-1})$ that

$$\sup\nolimits_{0 \le t-s \le 2^{-\ell}} |G_n(t) - G_n(s)| \le K(2^{-\ell})^\kappa \quad \text{for every } \ell \text{ with } 2^{-\ell} \ge \beta_n.$$

Thus (C2) holds with probability tending to one for the processes in question.

For an alternative justification let us consider the more common case where the y_k form a strong (α-)mixing sequence. In a first step we apply covariance inequalities due to E. Rio [47] yielding an estimate of the variance of the increments of G_n. Specifically, suppose that the mixing coefficients α_n decay as $n^{-\varrho}$ for some $\varrho > 1$. Let G denote the common CDF of the y_k. Then as a consequence of [47, Theorem 1.2] we get that for some finite constant K

$$\mathrm{Var}(G_n(t) - G_n(s)) \leq K n^{-1}(G(t) - G(s))^{1-1/\varrho}, \quad s < t, \, n \geq 1. \qquad (9.2)$$

Proof. (Cf. [47, pp. 590,591].) We have $G_n(t) - G_n(s) = n^{-1}\sum_{1 \leq k \leq n} \xi_k$, $\xi_k = \mathbb{1}_{s < y_k \leq t}$. The common 'quantile function' Q_ξ of the ξ_k is readily seen to be given by $Q_\xi(u) = 1$ if $u < Pr[s < y_k \leq t] = G(t) - G(s) \equiv \Delta$, and $= 0$ otherwise. Since $\alpha^{-1}(u) = \sum_k \mathbb{1}_{\alpha_k > u} = O(u^{-1/\varrho})$ it follows that the term

$$\int_0^1 \alpha^{-1}(u) Q_\xi(2u)^2 \, du \leq K \int_0^\Delta u^{-1/\varrho} \, du = O(\Delta^{1-1/\varrho}),$$

whence (9.2) follows. $\qquad \square$

Further by [47, Theorem 1.2], the limits of the sequences $n \, \mathrm{Var}(G_n(t) - G_n(s))$ and $n \, \mathrm{Var}(G_n(t))$, hence also of $n \, \mathrm{Cov}(G_n(s), G_n(t))$, exist for all s, t. The latter, for instance, is given by the absolutely convergent sum

$$\lim_{n \to \infty} n \, \mathrm{Cov}(G_n(s), G_n(t)) = \sum_{k \in \mathbb{Z}} \mathrm{Cov}(\mathbb{1}_{y_0 \leq s}, \mathbb{1}_{y_k \leq t}) \equiv \Lambda(s, t). \qquad (9.3)$$

Analogous expressions hold for the other limits.

We next appeal to the weak convergence of the processes $n^{1/2}(G_n(t) - G(t))$, $t \in \mathbb{R}$ to the mean zero Gaussian process $V(t)$, $t \in \mathbb{R}$ with covariance function Λ from (9.3), which follows from a stronger (almost sure) approximation result cited below. By the convergence of moments, the increments of V satisfy the analog of (9.2),

$$E(V(t) - V(s))^2 \leq K(G(t) - G(s))^{1-1/\varrho}.$$

Since V is Gaussian, an application of the well-known Garsia-Rodemich-Rumsey Lemma (see e.g. [3]) along with an intermediate time change implies that there exists a positive constant K such that for any $\delta < (1 - 1/\varrho)/2$ the process V satisfies with probability one a pathwise Hölder condition of the form

$$|V(t) - V(s)| \leq K(G(t) - G(s))^\delta, \quad s < t \quad \text{(almost surely, 'a.s.')}. \qquad (9.4)$$

Now suppose that the CDF G is uniformly Hölder continuous with index $\kappa \in (0, 1]$. Then by (9.4), V also fulfils, for any $\eta < \kappa(1 - 1/\varrho)/2$,

$$|V(t) - V(s)| \leq K(t - s)^\eta, \quad s < t \quad \text{(a.s.)}.$$

In order to transfer this pathwise Hölder condition to the processes G_n we may apply a "Hungarian type" strong approximation result for the empirical process of a stationary sequence. As a consequence of [57, Theorem] or [12, Theorem

2.1], there exists a sequence of Gaussian processes $V_n(t)$, $t \in \mathbb{R}$, all copies of V, all defined on a common probability space carrying also the y_k, such that

$$\sup_{t \in \mathbb{R}} |G_n(t) - G(t) - n^{-1/2} V_n(t)| = o(n^{-1/2}) \qquad \text{(a.s.)}. \qquad (9.5)$$

It follows that a.s. for all s, t with $|s - t| \le r$

$$\begin{aligned}
|G_n(t) - G_n(s)| &\le |G(t) - G(s)| + n^{-1/2} |V_n(t) - V_n(s)| + o(n^{-1/2}) \\
&= O(r^\kappa) + O(n^{-1/2} r^\eta) + o(n^{-1/2}).
\end{aligned}$$

This reduces to $O(r^\kappa)$ if we set $\beta_n = n^{-1/(2\kappa)}$, since then $n^{-1/2} = \beta_n^\kappa \le r^\kappa$ for $r \ge \beta_n$. The empirical processes F_{n1}, F_{n2} can be treated analogously. Thus we have shown that assumption (C2) is fulfilled with probability one under the indicated conditions, namely (sufficiently) strong mixing of the y_k and x_{k1}, x_{k2}, and Hölder continuity of their respective marginal CDFs. $\qquad \square$

CONDITION (C4). Suppose that the predictions x_{k1} form a strongly mixing sequence with the common marginal CDF F_1. Suppose, furthermore, that we have a strong approximation of the empirical processes F_{n1} as in the previously discussed case. Using the same notation V_n for the approximating Gaussian processes, we then have as in (9.5)

$$\sup_{\theta \in \mathbb{R}} |F_{n1}(\theta) - F_1(\theta) - n^{-1/2} V_n(\theta)| = o(n^{-1/2}) \quad \text{(a.s.)}.$$

Arguing as from (9.2) to (9.4) we get $|V_n(-\theta)| \le K F_1(-\theta)^\delta$ for all $\theta \ge 0$ (a.s.), where again δ may be any positive number less than $(1 - 1/\varrho)/2$ and $\varrho > 1$ has the same meaning for the x_{k1} as it had for the y_k.

We now assume that $\int |x|^q dF_1(x) < \infty$ for some $q > 0$. Then

$$\begin{aligned}
F_{n1}(-\theta) &\le F_1(-\theta) + n^{-1/2} |V_n(-\theta)| + o(n^{-1/2}) \\
&= O(\theta^{-q}) + O(n^{-1/2} \theta^{-q\delta}) + o(n^{-1/2}), \quad \theta \ge 1 \quad \text{(a.s.)}
\end{aligned}$$

For $1 \le \theta \le n^{1/q}$, hence $n^{-1/2} \le \theta^{-q/2}$, we have

$$F_{n1}(-\theta) = O(\theta^{-q}) + O(\theta^{-q/2 - q\delta}) + o(\theta^{-q/2}),$$

which certainly is $o(\theta^{-\nu})$ if we set $\nu = q/2$. An analogous estimate for the right tail gives $F_{n1}([-\theta, \theta]^c) = o(\theta^{-\nu})$ uniformly in the range $1 \le \theta \le n^{1/q}$ (a.s.). On the other hand, if $\theta > n^{1/q}$ then

$$Pr\left[F_{n1}([-\theta, \theta]^c) > 0\right] = Pr\left[\max_{k \le n} |x_{k1}| > \theta\right] \le n\theta^{-q} \int_{|x| > n^{\frac{1}{q}}} |x|^q dF_1(x),$$

which is $o(1)$ as $n \to \infty$. It follows that with probability tending to one we have $F_{n1}([-\theta, \theta]^c) = o(\theta^{-\nu})$ for all $\theta \ge 1$. Thus under the indicated assumptions, the tail condition (C4) is fulfilled with arbitrarily high probability if the marginal CDFs F_1, F_2 have a finite absolute moment of the order $q = 2\nu$. $\qquad \square$

Justification of assumption (A2). In view of general Poisson approximation results for frequencies of rare events (e.g., [8]) we may expect that under

broad conditions allowing for dependent observations the number N of data falling into a small interval J of length r is roughly Poisson distributed with parameter of the order nr^κ, where $\kappa \in (0,1]$ characterizes the (maximal) clustering of the data points. The p-th moment ($p \geq 1$) of the Poisson distribution with mean m is $O(m + m^p)$ uniformly in m. Thus if $nr^\kappa \geq 1$ we may expect that $EH_n(J)^p = O(n^{-p}(nr^\kappa)^p) = O(r^{p\kappa})$, while for $nr^\kappa \leq 1$ we should have $EH_n(J)^p = O(n^{-p}nr^\kappa) = O(n^{1-p}r^\kappa)$. Now given $b > 1$, choose $p > b/\kappa$ and put $\beta_n = n^{-(p-1)/(b-\kappa)}$. Noting that $\beta_n \leq n^{-1/\kappa}$, we find that in case $nr^\kappa \geq 1$ we have both $EH_n(J)^p = O(r^{p\kappa}) = O(r^b)$ and $r \geq \beta_n$. In case $nr^\kappa \leq 1$ we have $EH_n(J)^p = O(n^{1-p}r^\kappa)$, which is $O(\beta_n^{b-\kappa}r^\kappa) = O(r^b)$ if $r \geq \beta_n$. It follows that $EH_n(J)^p = O(r^b)$ whenever $r \geq \beta_n$, so that (A2) indeed would hold under quite general conditions. $\qquad\square$

Analysis of the quantile forecast example (Section 5.2). The difference of the elementary quantile scores is

$$d_k(\theta) = S_\theta(x_{k1}, y_k) - S_\theta(x_{k2}, y_k) = (\mathbb{1}_{y_k \leq \theta} - \alpha)\{\mathbb{1}_{\theta < x_{k1}} - \mathbb{1}_{\theta < x_{k2}}\}.$$

Taking our assumptions into account and passing to standard units on writing $t_k = (\theta - m_k)/s_k$ (and $z_{k\ell} = (x_{k\ell} - m_k)/s_k$), we get

$$h_k(\theta) := E[d_k(\theta) \mid \mathcal{F}_{k-1}] = (\Phi(t_k) - \alpha)\{Pr[t_k < z_{k1}] - Pr[t_k < z_{k2}]\} \quad (9.6)$$

where Pr refers to the $z_{k\ell}$ (resp. $u_{k\ell}$), everything else being considered as non-random (given $\mathcal{F}_{k-1}$). We henceforth omit the index k and use the abbreviation $h_k(\theta) \equiv h$.

Since $t < z_\ell$ iff $\log[\Phi(t)/(1 - \Phi(t))] - \log[\alpha/(1 - \alpha)] < u_\ell$, we have with $\lambda(p) = \log[p/(1 - p)]$ that

$$Pr[t < z_\ell] = 1 - \Phi([\lambda(\Phi(t)) - \lambda(\alpha)]/\tau_\ell) = \Phi([\lambda(\alpha) - \lambda(\Phi(t))]/\tau_\ell).$$

Suppose at first that $\Phi(t) - \alpha < 0$. Then $\lambda(\alpha) - \lambda(\Phi(t)) > 0$, so $\tau_1 < \tau_2$ implies

$$Pr[t < z_1] = \Phi([\lambda(\alpha) - \lambda(\Phi(t))]/\tau_1) > \Phi([\lambda(\alpha) - \lambda(\Phi(t))]/\tau_2) = Pr[t < z_2],$$

and hence $h < 0$, by (9.6). Analogously, $Pr[t < z_1] < Pr[t < z_2]$ if $\Phi(t) - \alpha > 0$. It follows that $h < 0$ in each case (and for all k, θ), proving that H_-^s holds iff $\tau_1 \leq \tau_2$. $\qquad\square$

Acknowledgments

References

[1] Andersen, T. G., Bollerslev, T. , Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* **71**, 579-625. MR1958138

[2] Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6** 170-176. MR0069229

[3] Azmoodeh, E., Sottinen, E., Viitasaari, L., and Yazigi, A. (2014). Necessary and sufficient conditions for Hölder continuity of Gaussian processes. *Statist. Probab. Lett.* **94** 230-235. MR3257384

[4] Bandi, F. M., and Corradi, V. (2014). Nonparametric nonstationary tests. *Econometric Theory* **30** 127-149. MR3177794

[5] Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inform. Theory* **51** 2664-2669. MR2246384

[6] Bollerslev, T. (1986). Generalized autoregressive conditional heteroscdasticity. *J. Econometrics* **31** 307-327. MR0853051

[7] Burkholder, D. L. (1966). Martingale transforms. *Ann. Math. Statist.* **37** 1494-2004. MR0208647

[8] Chen, L. Y. H., and Röllin, A. (2013). Approximating dependent rare events. *Bernoulli* **19** 1243-1267. MR3102550

[9] Chung, E., and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484-507. MR3099111

[10] Corradi, V., and Swanson, N. R. (2006). The effect of data transformation on common cycles, cointegration and unit root tests. Monte Carlo results and a simple test. *J. Econometrics* **132** 195-229. MR2271396

[11] Das Gupta, S. (1976). A generalization of Anderson's theorem on unimodal functions. *Proc. Amer. Math. Soc.* **60** 85-91. MR0425050

[12] Dedecker, J., Merlevède, F., and Rio, E. (2013). Strong approximation results for the empirical process of stationary sequences. *Ann. Probab.* **41** 3658-3696. MR3127895

[13] DelSole, T., and Tippett, M. K. (2014). Comparing forecast skill. *Mon. Weather Rev.* **142** 4658-4678.

[14] Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253-263. MR3303732

[15] Efron, B., and Hastie, T. (2016). Computer age statistical inference: Algorithms, evidence, and data science. Cambridge University Press, Cambridge. MR3523956

[16] Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations, and forecast rankings (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 505-562. MR3506792

[17] Elliott, G., Ghanem, D., and Krüger, F. (2016). Forecasting conditional probabilities of binary outcomes under misspecification. *Rev. Econ. Stat.* **98** 742-755.

[18] Elliott, G., and Timmermann, A. (2016). Economic forecasting. Princeton University Press, Princeton.

[19] Fissler, T., and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Ann. Statist.* **44** 1680-1707. MR3519937

[20] Giacomini, R., and White, H. (2006). Tests of conditional predictive ability. *Econometrica* **74** 1545-1578. MR2268409

[21] Gneiting, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746-762. MR2847988

[22] Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359-378. MR2345548

[23] Gneiting, T., and Ranjan, R. (2013). Combining predictive distributions. *Electron. J. Stat.* **7** 1747-1782. MR3080409

[24] Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64** 413-430. MR1375740

[25] Hansen, P. R. (2005). A test for superior predictive ability. *J. Business Econ. Statist.* **23** 365-380. MR2206007

[26] Holzmann, H., and Eulert, M. (2014). The role of the information set for forecasting with applications to risk management. *Ann. Appl. Statist* **8** 595-621. MR3192004

[27] Horn, R. A., and Johnson, C. R. (2013). *Matrix Analysis,* 2nd ed. Cambridge University Press, Cambridge. MR2978290

[28] Jin, S., Corradi, V., and Swanson, N. R. (2017). Robust forecast comparison. *Econometric Theory* **33** 1306-1351. MR3715003

[29] Jordan, A., and Krüger, F. (2016). *murphydiagram: Murphy Diagrams for Forecast Comparisons.* R package, version 0.11, available at https://cran.r-project.org/web/packages/murphydiagram/index.html.

[30] Koenker, R. (2005). *Quantile Regression.* Cambridge University Press. MR2268657

[31] Koenker, R. (2017). *quantreg: Quantile Regression.* R package, version 5.33, available at https://cran.r-project.org/web/packages/quantreg/index.html.

[32] Krüger, F. and Ziegel, J. F. (2018). Generic conditions for forecast dominance. Preprint, available at https://arxiv.org/abs/1805.09902.

[33] Künsch, H. R. (1989). The jackknife abd the bootstrap for general stationary observations. *Ann. Statist.* **17** 2356-2382. MR1015147

[34] Lai, T. L., Gross, S. T., and Shen, D. B. (2011). Evaluating probability forecasts. *Ann. Statist.* **39** 2356-2382. MR2906871

[35] Mammen, E. (1993). Boostrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21** 255-285. MR1212176

[36] McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* **2** 620-628. MR0358933

[37] Merkle, E. C. and Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decis. Anal.* **10** 292-304. MR3150456

[38] Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *Rev. Econom. Stud.* **72** 735-765. MR2148141

[39] Linton, O., Song, K., and Whang, Y.-J. (2010). An improved bootstrap test of stochastic dominance. *J. Econometrics* **154** 108-202. MR2558960

[40] McNeil, A., Frey, R., and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools,* revised ed. Princeton University Press, Princeton. MR3445371

[41] Mudholkar, G. S. (1966). The integral of an invariant unimodal function over an invariant convex set–an inequality and applications. *Proc. Amer. Math. Soc.* **17** 1327-1333. MR0207928

[42] Newey, W. K., and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819-847. MR0906565

[43] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *J. Econometrics* **160** 246-256. MR2745881

[44] Patton, A. J. (2016). Comparing possibly misspecified forecasts. Preprint, Duke University.

[45] Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Institute of Mathematical Statistics, Hayward. MR1089429

[46] R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

[47] Rio, E. (1993). Covariance inequalities for strongly mixing processes. *Ann. Inst. Henri Poincaré* **29** 587-597. MR1251142

[48] Rudebusch, G. D., and Williams, J. C. (2009). The puzzle of the enduring power of the yield curve. *J. Bus. Econom. Stat.* **27** 492-503. MR2572036

[49] Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **71** 783-801. MR0331571

[50] Schuford, E. H., Albert, A., and Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika* **31** 125-145.

[51] Seillier-Moisewitsch, F., and Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.* **88** 355-359. MR1212496

[52] Strähl, C., and Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electron. J. Stat.* **11** 608-639. MR3619318

[53] Trapani, L. (2016). Testing for (in)finite moments. *J. Econometrics* **191**, 57-68. MR3434435

[54] van der Vaart, A. W. (1998). *Asympotic Statistics.* Cambridge University Press, Cambridge. MR1652247

[55] Wu, W.B. (2008). Empirical processes of stationary sequences. *Statist. Sinica* **18** 313-333. MR2384990

[56] Yen, T.-J., and Yen, Y.-M. (2018). Testing forecast accuracy of expectiles and quantiles with the extremal consistent loss functions. Preprint, available at https://arxiv.org/abs/1707.02048.

[57] Yoshihara, K.-I. (1979). Note on an almost sure invariance principle for some empirical processes. *Yokohama Math. J.* **27** 105-110. MR0560618

[58] Ziegel, J. F., Krüger, F., Jordan, A., and Fasciati, F. (2017). Murphy Diagrams: Forecast evaluation of Expected Shortfall. Preprint, available at https://arxiv.org/abs/1705.04537.

[59] Ziegler, K. (1997). Functional central limit theorems for triangular arrays of

function-indexed processes under uniformly integrable entropy conditions. *J. Multivariate Anal.* **62** 233-272 MR1473875

[60] Žikeš, F., and Baruník, J. (2016). Semi-parametric conditional quantile models for financial returns and realized volatility. *J. Financ. Economet.* **14** 185-226.