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.

Keywords and phrases: Comparative forecast evaluation, hypothesis testing, randomization.

Received November 2017.

Contents

1 Introduction ......................................................... 3759
2 Testing for forecast dominance – Initial considerations ........ 3762
  2.1 Formal setup .................................................. 3762
  2.2 Notions of forecast dominance .............................. 3762
  2.3 A fictitious test ............................................... 3764
3 Randomization tests ............................................. 3765

∗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 \( \phi \) be a real-valued functional defined on a class \( G \) of possible distributions \( G \) of \( y \), such as the mean or a quantile of \( G \).

The scoring function is consistent (for \( \phi \) relative to \( G \)) if \( S(\phi(G), G) \leq S(x, G) \) for every \( G \in G \) 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 \( \phi \), 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 \( S(x, y) = \varphi(y) - \varphi(x) - \varphi'(x)(y - x) \), \( \varphi \) a convex function with subgradient \( \varphi' \), 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 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 $S_\theta$, $\theta \in \mathbb{R}$. 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

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

(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 functions $S_\theta$ 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-
Randomization test

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

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

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 difficult 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

\[
D_n^*(\theta) = n^{-1/2} \sum_{k \leq n} d_k(\theta)\sigma_k
\]

and \(Pr^*\) exclusively refers to the i.i.d. (“Rademacher”) random variables \(\sigma_k\) assuming the values \(\pm 1\) 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 assumptions cannot mean doing without any assumptions. However, as detailed in Section 4.1, assumptions 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 \((x_k^1, x_k^2, y_k), k = 1, \ldots, n\) be a sequence of \(n\) triplets where \(x_k^1, x_k^2\) 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 \((x_k^1, x_k^2, y_k)\) is \(\mathcal{F}_k\)-measurable for every \(k\), and \(\mathcal{F}_0\) is trivial. Given a family \(S = \{S_\theta, \theta \in \Theta\}\) of scoring functions \(S_\theta\), we compare the two forecasts via the suitably normalized average difference of their \(S_\theta\) scores, i.e., we are interested in the stochastic process \(\theta \mapsto D_n(\theta) = \frac{n-1}{2} \sum_{k \leq n} d_k(\theta), \quad d_k(\theta) = S_\theta(x_k^1, y_k) - S_\theta(x_k^2, y_k).\) (2.1)

Initially, the family \(S\) 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 randomization 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 \(S\)) 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:

\[
H^- : \sup_\theta E_Q D_n(\theta) = \sup_\theta n^{-1/2} \sum_{k \leq n} E_Q d_k(\theta) \leq 0. \tag{2.2}
\]

In fact, \(H^-\) 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 \(H^-\) involves unconditional expectations referring to both the observations \(y_k\) and the forecasts \(x_k^\ell\). This form of \(H^-\) 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

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

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

\[
H^- : M_{n,Q}(\theta) \leq 0 \quad (Q\text{-a.s.}, \theta \in \Theta). \tag{2.4}
\]
Proposition 2.1. If for the mean value; see (1.1). In this case we even have the following.

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

$$H^* : E_Q \{ d_k(\theta) | \mathcal{F}_{k-1} \} \leq 0 \quad (Q\text{-a.s., } \theta \in \Theta, \ k = 1, \ldots, n).$$

(2.5)

The interpretation of $H^*$ 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 $H^*$ 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 $\eta_{k}, k \geq 1, \ell = 1, 2$ 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 $\epsilon_{k\ell} \sim \mathcal{N}(0, \tau^2)$. Suppose that at any instance $k$, forecaster 1 has access to $\eta_{k1}$ and the preceding value $\eta_{k-1,2}$ of the second process. 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 $\tau_1 \geq \tau_2$, 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 $\sigma$-algebra generated by all triplets $(x_{j1}, x_{j2}, y_j)$, $j \leq k$, then $E \{ d_k | \mathcal{F}_{k-1} \} \leq 0$, 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 $E \{ d_k | \mathcal{F}_{k-1} \} = 0$, 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, $E \{ d_k(\theta) | \mathcal{F}_{k-1} \} = 0$ for every $\theta \in \mathbb{R}$, where $d_k(\theta)$ denotes the score difference with respect to the elementary scoring functions $S_0$ for the mean value; see (1.1). In this case we even have the following.

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

---

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

3This setup is similar to the simulation example in [23, Section 4.1], except that our variant includes time series dynamics.
Proposition 2.2. Suppose there is some non-random function $\gamma$ 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 $H_+$) 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 $\Theta$, 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.

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

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 $\Theta$.

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

$$\tilde{d}_{k,Q}(\theta) = d_k(\theta) - E_Q\{d_k(\theta) | \mathcal{F}_{k-1}\},$$

and let

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

denote the conditionally centered version of $D_n$. Then by monotonicity

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

for every $Q \in H_-$, as claimed. Thus, if critical values $\tilde{c}_n$ could be obtained such that $\sup_{Q \in H_-} Pr_Q[T(\tilde{D}_{n,Q}) > \tilde{c}_n] \approx \alpha \downarrow$ the rule ‘reject $H_-$ if $T(D_n) > \tilde{c}_n$’ would give us an approximate level-$\alpha$ 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) $\limsup_{n \to \infty} n^{-1} \sum_{k \leq n} E\{d_k(\theta)^2 1_{|d_k(\theta)| > \epsilon \sqrt{n}}\} = 0 \quad (\theta \in \Theta, \ \epsilon > 0)$. \quad \text{(2.8)}$

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

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

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

$^4$The notation ‘$a_n \approx b_n$’ is short for $a_n - b_n \to 0$ as $n \to \infty$.\end{flushright}
The proposition suggests that for large \( n \) the distribution of the test statistic \( T(D_n) \) at the boundary, where \( D_n = \tilde{D}_n \), can be approximated by the distribution of the functional \( T(\tilde{Z}) \) on the paths of the Gaussian process \( \tilde{Z} \). Of course, convergence of the finite-dimensional distributions is insufficient for such a conclusion; tightness of the processes \( \tilde{D}_n \) in a suitable function space is required, too. Furthermore, the distribution of \( T(\tilde{Z}) \) generally is unknown and may be difficult to determine. And there still is the problem that the process \( \tilde{D}_n \) 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 \( \tilde{c}_n \) we refer to the hypothetical test rejecting \( H_- \) if \( T(D_n) > \tilde{c}_n \) as the “fictitious test.”

For the sake of exposition we introduce further sub-hypotheses of \( H_- \) besides \( H^*_- \), namely the null-hypothesis \( H_0 \) of equal performance on average,

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

and the null-hypothesis \( H^*_0 \) of equal performance at every forecast instance,

\[
H^*_0 : \quad E[d_k(\theta) | \mathcal{F}_{k-1}] = 0 \quad (\text{a.s., } \theta \in \Theta, \; k = 1, \ldots, n). \tag{2.10}
\]

Evidently, \( H_0, \; H^*_0 \) may be regarded as the boundaries of the one-sided hypotheses \( H_- \), \( H^*_- \), respectively, representing different forms of equal predictive ability. Note that \( H^*_0 \subset H^*_- \subset H_- \subset H^w_- \); 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 \( \sigma_1, \sigma_2, \ldots \) be i.i.d. such that \( \sigma_k = \pm 1 \) with probability 1/2 each, and define

\[
D_n^*(\theta) = n^{-1/2} \sum_{k \leq n} d_k(\theta)\sigma_k. \tag{3.1}
\]

We reject \( H_- \) “at level \( \alpha \)” 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 \( \sigma_k \), the data \( x_{k1}, x_{k2}, y_k \) being considered as fixed, non-random quantities. Henceforth we refer to this test as
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

\[ \sup_{H_-} \Pr [T(D_n) > c_n^+] = \sup_{H_0} \Pr [T(D_n) > c_n^*], \quad (3.2) \]

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 $\alpha$” shall underline that the test has (approximative) level $\alpha$ 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 $\theta$), 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 $\sigma_k$, and via the statistical nature of the data triplets. In part b) we condition on the data, leaving the $\sigma_k$ as the sole source of randomness.

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

\[ n^{-1} \sum_{k \leq n} d_k(\theta_1) d_k(\theta_2) \equiv \gamma_n(\theta_1, \theta_2) \rightarrow_p \gamma(\theta_1, \theta_2). \quad (3.3) \]

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 $\gamma$.

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 $\sigma_k$ 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 \(\leq \alpha\). 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 \(H_0^*\).**

Under \(H_0^\ast\) we have \(d_k(\theta) = \tilde{d}_k(\theta)\), hence \(D_n = \tilde{D}_n\) and \(\gamma_n = \tilde{\gamma}_n\). Consequently, the limit processes \(Z\) and \(\tilde{Z}\) of \(D_n^*\) and \(\tilde{D}_n\) are identical in distribution under \(H_0^\ast\), 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 \(\tilde{c}_n\) of the randomization and the fictitious test coincide asymptotically. Therefore, since the latter test is, for large \(n\), approximatively valid for testing \(H_0^\ast\) at level \(\alpha\), 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 \(\gamma\) and \(\tilde{\gamma}\), hence the limit processes \(Z\) and \(\tilde{Z}\), 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 \(\gamma = \tilde{\gamma} + \psi\) with a positive definite function \(\psi\) given by the stochastic limit

\[
\psi(\theta_1, \theta_2) = p\text{-lim} \, n^{-1} \sum_{k \leq n} E[d_k(\theta_1) \mid F_{k-1}] \, E[d_k(\theta_2) \mid F_{k-1}]; \tag{3.4}
\]

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

Now in the context of the hypotheses \(H_0, H_0^\ast\) 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 \(\tilde{Z}\) 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

$$\Pr[T(D_n) > c_n^*] \overset{(1)}{=} \Pr[T(\tilde{D}_n) > c_n^*] \approx \Pr[T(\tilde{Z}) > c_n^*]$$

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

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

**Hypotheses $H_-, H_*$.**

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 $H_-, H_*$. The argument is parallel to (3.5), with two modifications: first, since $M_n \leq 0$ under $H_-, H_*$, instead of (1) we have an inequality, $\Pr[T(D_n) > c_n^*] \leq \Pr[T(\tilde{D}_n) > c_n^*]$, 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 $P_n \in H_0^*$. This is because under $P_n$ we have $\gamma_n = \tilde{\gamma}_n$, hence $\gamma_n \to_p \gamma = \tilde{\gamma}$, 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 $\tilde{Z}$ coincide, and the inequality (2) in (3.5) becomes an equality; while the ‘=’ sign (1) there has to be replaced by ‘≤’, again by the monotonicity of $T$. Thus in this case too, $\Pr_{Q_n}[T(D_n) > c_n^*] \leq \alpha$, as claimed.

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

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 \( M_n, Q_n \geq 0 \) a.s. for all \( \theta \in \Theta \); cf. (2.4) – that are contiguous to a sequence \( P_n \in H_0^* \). By monotonicity of \( T \) one analogously gets \( \Pr_{Q_n}[T(D_n) > c_n^\alpha] \geq \alpha \), 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 \( X_i \sim \mathcal{N}(0, \nu_i) \) (\( i = 1, 2 \)) with \( \nu_1 \leq \nu_2 \), 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 \( A \subseteq \mathbb{R}^d \) be a convex acceptance region of the form \( A = A_{g,b} = \{ x \in \mathbb{R}^d : g(x) \leq b \} \) 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 \( R \subseteq S \), 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 \( E^c = E \subseteq A \subseteq R^c \), that is, \( R \subseteq S \). 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, \( X = (\tilde{Z}(\theta_1), \ldots, \tilde{Z}(\theta_d)), Y = (W(\theta_1), \ldots, W(\theta_d)) \) with the \( \theta_i \) becoming dense. Since the covariance function \( \tilde{\gamma} \) of \( \tilde{Z} \) generally is unknown, we
have no control on the eigenvalues of $V^{-1} - V_+^{-1}$. Proposition 3.2 thus does not guarantee that $a_0$ 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 $\phi$ on the class $G$ of predictive distributions $G$ and related families of consistent scoring functions $S_{\theta}$. Hereafter we focus on functionals representing a quantile or an expectile. Given $\alpha \in (0, 1)$, the $\alpha$-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) $\alpha$-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 $S_{\theta}$, such that any consistent scoring function can be written as

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

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 $\alpha$-quantiles are

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

setting $M$ in (4.1) equal to the Lebesgue measure results in the popular ‘piecewise linear’ score considered in quantile regression [30]. For $\alpha$-expectiles,

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

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...
Randomization test

of an identification function $I$ that depends only on the observation, times a difference of indicator functions depending only on the forecasts. Specifically, $d_k(\theta) = I(\theta, y_k)\{1_{\theta < x_{k1}} - 1_{\theta < x_{k2}}\}$ [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 $\sigma_k$ 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 $D_n^*$,

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

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

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

for quantiles: $H_n = G_n + F_{n1} + F_{n2},$

for expectiles: $H_n = F_{n1} + F_{n2}.$

Assumptions.

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

$$n^{-1} \sum_{k \leq n} d_k(\theta_1)d_k(\theta_2) = \gamma_n(\theta_1, \theta_2) \rightarrow \gamma(\theta_1, \theta_2) \ (n \rightarrow \infty, \ \theta_1, \theta_2 \in \mathbb{R}).$$

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

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

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

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

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

Discussion of the assumptions. In practice, the test functionals in (2.6) may be evaluated over a bounded interval $\Theta$ in the full $\theta$-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 \( S_\theta \) with
a measure \( M \) supported by \( \Theta \), 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 ar-

Theorem 4.1. Under the assumptions (C1) to (C4) the processes \( \{ \bar{D}_n(\theta), \theta \in \mathbb{R} \} \) converge weakly in \( \ell_0^\infty \) to a mean zero Gaussian process \( \{ Z(\theta), \theta \in \mathbb{R} \} \) with
covariance function \( \gamma \) and continuous sample paths.
Randomization test

As a consequence of the theorem, \( T(\tilde{D}_n) \) converges weakly in distribution to \( T(Z) \) for any continuous functional on the space \( \ell^\infty_0 \). 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 \( T_p(f) = \int_{\mathbb{R}} f(\theta) \, d\theta \) \((p = 1, 2)\) require a sharpening of assumption (C4) for the control of the tail masses.

**Corollary 4.1.** Assume (C1) to (C4). Then \( T_1(\tilde{D}_n) \) converges weakly in distribution to \( T_1(Z) \) if the exponent \( \nu \) 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 \( \nu \) 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 \( \Theta \) endowed with the sup-norm, and because weak convergence in \( \ell^\infty_0 \) implies weak convergence in \( \ell^\infty(\Theta) \).

**4.2. Weak convergence of \( \tilde{D}_n \)**

Here the focus is on the approximation \( Pr[T(\tilde{D}_n) > c_n^*] \approx Pr[T(\tilde{Z}) > c_n^*] \) 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 \( \Theta \), 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

\[
G_n^c(J) = \frac{1}{n} \sum_{k \leq n} Pr[y_k \in J \mid F_{k-1}], \quad F_n^c(J) = \frac{1}{n} \sum_{k \leq n} Pr[x_{k\ell} \in J \mid F_{k-1}],
\]

\( \ell = 1, 2, J \) 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 \( F_n^c_{n\ell} = F_{n\ell}^c \) in the common case of forecasts \( x_{k\ell} \) that are \( F_{k-1} \)-measurable. The following assumptions are similar to those in the previous section, except that convergence is ‘in probability’ 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 \( \tilde{\gamma} \) such that \( \tilde{\gamma}(\theta, \theta) > 0 \) for all \( \theta \), and

\[
n^{-1} \sum_{k \leq n} \tilde{d}_k(\theta_1) \tilde{d}_k(\theta_2) = \tilde{\gamma}_n(\theta_1, \theta_2) \to_p \tilde{\gamma}(\theta_1, \theta_2) \quad (n \to \infty, \theta_1, \theta_2 \in \mathbb{R}).
\]

Given any \( b > 1 \) there exists a number \( p \geq 1 \) such that (A2) and (A3) hold:
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 $K_n = H_n^c$

\[
\sup_{0 \leq \theta_2 - \theta_1 \leq r} EK_n([\theta_1, \theta_2])^p \leq B (r \vee \beta_n)^b, \quad r \in [0, 1], \ n \geq n_2.
\]

As before we consider a continuous version of $\tilde{D}_n$ obtained by linear interpolation on the grid $\{j \beta_n : j \in \mathbb{Z}\}$, $\beta_n$ as in Assumption (A2), which we denote as $\hat{D}_n$. Accordingly, the approximation to be established becomes

\[
\Pr\left[T(\hat{D}_n) > c^*_n\right] \approx \Pr\left[T(\tilde{Z}) > c^*_n\right].
\]

**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 $\{\tilde{Z}(\theta) : \theta \in \mathbb{R}\}$ with covariance function $\tilde{\gamma}$ and continuous sample paths.

For bounded $\Theta$ the functionals $T_\infty, T_1, T_2$, 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 1000 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 $\tau_2$ 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 empirically 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 nominal 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 inequality. 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 $\tau_2$ (i.e., clearer violations of the hypothesis imply higher rejection rates). Furthermore, the functional $T_1$ has a slightly higher power than $T_2.$
5.2. Quantile forecasts

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

\[
\begin{align*}
y_k &= 0.03 + 0.05 y_{k-1} + s_k \varepsilon_k \\
s_k^2 &= 0.05 + 0.9 s_{k-1}^2 + 0.05 s_{k-1}^2 \varepsilon_{k-1}^2
\end{align*}
\]

with independent “shocks” $\varepsilon_k \sim N(0, 1)$. 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 $s_0, \epsilon_0, y_0$, she could successively compute $s_k$ from the observations $y_j, j < k$. 

---

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 1000 Monte Carlo iterations; within each iteration, the test is computed based on 1000 simulated sign randomizations. See Sections 5.1 and 5.2 for further details.
$F_k$ denote the $\sigma$-algebra generated by the variables $s_0, \epsilon_0, y_j; j \leq k$. Then (assuming the regression parameters are known) $y_k \mid F_{k-1} \sim N(m_k, s_k^2)$, 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 $\alpha$-quantile of the conditional distribution of $y_k$, namely $x_{k,\text{ideal}} = m_k + s_k z_\alpha$ 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 $x_{k,\ell} = m_k + s_k z_{\ell}$ ($\ell = 1, 2$) where the $z_{\ell}$ are random perturbations of $z_\alpha$ that are independent among themselves and from all other variables. Specifically, we assume that the deflections from $z_\alpha$ are Gaussian in the log odds scale,

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

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 $H_+^\tau$ 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 $\tau_2$ varies from 0.05 to 0.5. The quantile level is $\alpha = 0.05$, and the sample size is 2000. 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 $\tau_2$ supports our claim that the randomization test is approximatively valid for testing $H_+^\tau$.

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. 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 1000 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)**

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Test statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPF dominates Probit</td>
<td>$T_1$</td>
<td>0.987</td>
</tr>
<tr>
<td>SPF dominates Probit</td>
<td>$T_2$</td>
<td>0.986</td>
</tr>
<tr>
<td>Probit dominates SPF</td>
<td>$T_1$</td>
<td>0.010</td>
</tr>
<tr>
<td>Probit dominates SPF</td>
<td>$T_2$</td>
<td>0.002</td>
</tr>
</tbody>
</table>

**Quantile forecasts (stock returns, α = 0.05)**

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Test statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>QR$<em>{RV}$ dominates QR$</em>{ABS}$</td>
<td>$T_1$</td>
<td>0.646</td>
</tr>
<tr>
<td>QR$<em>{RV}$ dominates QR$</em>{ABS}$</td>
<td>$T_2$</td>
<td>0.916</td>
</tr>
<tr>
<td>QR$<em>{ABS}$ dominates QR$</em>{RV}$</td>
<td>$T_1$</td>
<td>0.000</td>
</tr>
<tr>
<td>QR$<em>{ABS}$ dominates QR$</em>{RV}$</td>
<td>$T_2$</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**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/](http://realized.oxford-man.ox.ac.uk/). Quantiles at low levels $\alpha$ are commonly used as measures for financial risk, and are referred to as Value-at-Risk at level $\alpha$ (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

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

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. 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...
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 $F_{k-1}$.

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. 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

---

8We 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 predictive 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 \( H^- \). The discussion in Section 3.2 elaborates the central role of our

---

9To 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 \( H_0^n \); 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 \( S_0 = \varphi(y) - \varphi(x) - (y - x)\varphi'(x) \) with the convex function \( \varphi(x) = \frac{1}{2} (x - \theta)_+ \).

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

\[
X_{kn} = \sum_j c_j n^{-1/2} \tilde{a}_k(\theta_j), \quad k = 1, \ldots, n.
\]

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

\[
\sum_{k \leq n} X_{kn}^2 = \sum_{i,j} c_i c_j n^{-1} \sum_{k \leq n} \tilde{d}_k(\theta_i) \tilde{d}_k(\theta_j) \to_p \sum_{i,j} c_i c_j \tilde{\gamma}(\theta_i, \theta_j) = V.
\]  

Thus if

\[
E \left( \max_{k \leq n} X_{kn}^2 \right) \to 0 \quad \text{as } n \to \infty
\]

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

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

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 \( \theta \) we have

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

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 $\epsilon$ 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 $d_k(\theta)$ by $d_k(\theta) \sigma_k$, define $F_k$ as the $\sigma$-algebra generated by the random variables $\sigma_1, \ldots, \sigma_k$, and observe that $|\sigma_k| = 1$ and $E \left[ d_k(\theta) \sigma_k \mid F_{k-1} \right] = 0$. 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

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

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

$$
ED_n^*(\theta)^2 \leq A_{q,e} |\theta|^{-\nu_{q,e}}, \quad |\theta| \geq 1, \quad n \geq n_1.
$$

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

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

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

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

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

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

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

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

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

Similarly, if $|\theta| \geq 1, \quad n \geq n_1$, 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
\[
ED_n^*(\theta)^2 = n^{-1} \sum_{k \leq n} d_k(\theta)^2 \leq \left\{ \left[ n^{-1} \sum_{k \leq n} m_k^2 \right] \left[ n^{-1} \sum_{k \leq n} |\delta_k(\theta)| \right] \right\}^{1/2} \leq \left\{ MA|\theta^{-\nu}\right\}^{1/2},
\]
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
\[
d_k(\theta_2) - d_k(\theta_1) = \{I(\theta_2, y_k) - I(\theta_1, y_k)\} \delta_k(\theta_1) + I(\theta_2, y_k) \delta_k(\theta_1, \theta_2)
\]
with either $i = 1, j = 2$ or $i = 2, j = 1$, whichever is more convenient, we get for $\alpha$-quantiles
\[
|d_k(\theta_2) - d_k(\theta_1)| \leq 1_{\theta_1 < y_k \leq \theta_2} + |\delta_k(\theta_1, \theta_2)|,
\]
and for $\alpha$-expectiles
\[
|d_k(\theta_2) - d_k(\theta_1)| \leq |\theta_2 - \theta_1|(|\delta_k(\theta_1)| \vee |\delta_k(\theta_2)|) + \{|y_k - \theta_1| \wedge |y_k - \theta_2|\} |\delta_k(\theta_1, \theta_2)|
\]
\[
\leq |\theta_2 - \theta_1|(|\delta_k(\theta_1)| \vee |\delta_k(\theta_2)|) + \{|y_k - x_{k1}| \lor |y_k - x_{k2}|\} |\delta_k(\theta_1, \theta_2)|.
\]
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
\[
n^{-1} \sum_{k \leq n} |\delta_k(\theta_1, \theta_2)| \leq F_n([\theta_1, \theta_2]) + F_n([\theta_1, \theta_2]).
\]

For quantiles we then get by (8.9)
\[
\rho_n(\theta_1, \theta_2)^2 \leq 2 \{G_n([\theta_1, \theta_2]) + F_n([\theta_1, \theta_2])\} \leq 2H_n([\theta_1, \theta_2]),
\]
while for expectiles the estimates (8.10), (8.11) and Cauchy-Schwarz give similarly as at (8.8)
\[
\rho_n(\theta_1, \theta_2)^2 \leq 2 (\theta_2 - \theta_1)^2 + 2 \{MH_n([\theta_1, \theta_2])\}^{1/2}.
\]

Assertion (iii) thus follows from (C2).

\textbf{Lemma 8.2.} Up to adjustments of the constants, the assertions of Lemma 8.1 also hold for the interpolated processes $\tilde{D}_n$, with the following improvement of (iii):
\[
\sup_{0 \leq \theta_2 - \theta_1 \leq r} E(\tilde{D}_n(\theta_2) - \tilde{D}_n(\theta_1))^2 \leq C_{q,e} r^{\lambda'\nu}\quad (r \in [0, 1], \ n \geq n_2)
\]
\text{(i.e., with $r^{\lambda'\nu}$ rather than $(r \lor \beta_n)^{\lambda'\nu}$). Furthermore,}
\[
\lim_{n \to \infty} \sup_{\theta} E(\tilde{D}_n(\theta) - D_n^*(\theta))^2 = 0,
\]
\[
\lim_{n \to \infty} E \int (\tilde{D}_n(\theta) - D_n^*(\theta))^2 d\theta = 0.
\]
Proof. For convenience we immediately write \( \beta_n \equiv \epsilon \). Given \( \theta \), 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

\[
E(D_n(\theta) - D_n^*(\theta))^2 \leq n^{-1} \sum_{k \leq n} \left\{ w[d_k((\ell + 1) \epsilon) - d_k(\theta)]^2 + (1 - w)[d_k(\theta) - d_k(\ell \epsilon)]^2 \right\}
\]

\[
\leq \omega_n(\epsilon) \equiv \omega_n(\beta_n),
\]

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 \( \ell \) and \( w \) are uniquely determined by \( \theta \) we get by (8.9),

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

The right-hand side is always \( \leq 4 \), and it vanishes except if both \( \theta \) 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 \( \epsilon \) on which the function \( \theta \mapsto \Delta_k(\theta) \) is non-zero. Consequently, \( \int \Delta_k(\theta) d\theta \leq 12 \epsilon \), so taking the average over \( k \) settles the quantile case. A similar reasoning applies in the expectile case. By (8.10),

\[
\Delta_k(\theta) \leq 2 \epsilon^2 \{|\delta_k(\ell \epsilon)| \lor |\delta_k((\ell + 1) \epsilon)|\} + 2 m_k^2 |\delta_k(\ell \epsilon, (\ell + 1) \epsilon)|.
\]

The term \( |\delta_k(\ell \epsilon)| \lor |\delta_k((\ell + 1) \epsilon)| \) is nonzero at most if \( \theta \in \left[ x_{k1} \land x_{k2} - \epsilon, x_{k1} \lor x_{k2} + \epsilon \right] \). Therefore

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

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

Straightforward estimates yield the uniform Hölder condition (8.14) at first for points \( \theta_1, \theta_2 \) 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.

\[ \square \]

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 \( D_n \). Without loss of generality we may assume \( n \geq n_1 \lor n_2 \) (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 \( \rho_n \) or \( \omega_n \) may also be used with \( D_n \), 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 \( \min_{t \in T} \rho_n(\theta, t) \leq u \) for every \( \theta \in T_0 \). Given \( b > 1 \), pick \( t_1 \in [-b, b] \) equidistant with spacing \( r = 2(b/B)^{1/\lambda} \). By (8.14), the minimal \( \rho_n \)-distance of any \( \theta \in [-b, b] \) to the resulting set \( T \) is \( \leq \omega_n(r/2) \leq u \),
whence \( N_n(u, [-b, b]) \leq K b u^{-1/\lambda} \). 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

\[
\rho_n(\theta, b)^2 \leq 2 \{ E \tilde{D}_n(\theta)^2 + E \tilde{D}_n(b)^2 \} \leq (K b^{-\nu})^2, \quad \theta > b, \tag{8.17}
\]

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

\[
N_n(u, \mathbb{R}) \leq K u^{-1/\nu - 1/\lambda}, \quad u > 0. \tag{8.18}
\]

Let

\[
\Omega_n(r) = \sup \{ |\tilde{D}_n(\theta_2) - \tilde{D}_n(\theta_1)| : \rho_n(\theta_1, \theta_2) \leq r, \theta_1, \theta_2 \in \mathbb{R} \}, \quad r > 0.
\]

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

\[
E \Omega_n(r) \leq K \int_0^{r/4} (\log N_n(u, \mathbb{R}))^{1/2} du \leq K \int_0^{r} (\log u^{-1/\nu - 1/\lambda})^{1/2} du
\]

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

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

\[
|\tilde{D}_n(\theta)| \leq |\tilde{D}_n(b)| + |\tilde{D}_n(b) - \tilde{D}_n(\theta)| \leq |\tilde{D}_n(b)| + \Omega_n(\eta)
\]

and so

\[
\sup_{|\theta| > b, \theta \in \mathbb{R}} |\tilde{D}_n(\theta)| = o_{p}(1) \quad \text{as} \quad b \rightarrow \infty.
\]

Secondly, by (8.14)

\[
\tilde{\Omega}_n(r) = \sup \{ |\tilde{D}_n(\theta_2) - \tilde{D}_n(\theta_1)| : |\theta_1 - \theta_2| \leq r, \theta_1, \theta_2 \in \mathbb{R} \}
\leq \Omega_n(\omega_n(r)) \leq \Omega_n(K r^\lambda),
\]

whence \( E \tilde{\Omega}_n(r_n) \rightarrow 0 \) if \( r_n \rightarrow 0 \), implying asymptotic equicontinuity also with respect to the standard metric. It follows that the processes \( \tilde{D}_n \) converge weakly in \( \ell_{0, \mathbb{R}}^{\infty} \) 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, \( \int_{|\theta| > b} E \tilde{D}_n(\theta)^2 \ d\theta \rightarrow 0 \) as \( b \rightarrow \infty \) under the given conditions. Consequently, \( T_2(\tilde{D}_n) \) equals \( \int_{|\theta| \leq b} \tilde{D}_n(\theta)^2 \ d\theta \) 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 \(\xi, K\), and \(\eta > 1\), such that for all \(\theta_1, \theta_2\) in the grid \(\{j\beta_n : j \in \mathbb{Z}\}\) one has

\[
E|\tilde{D}_n(\theta_2) - \tilde{D}_n(\theta_1)|^\xi \leq K|\theta_2 - \theta_1|^p. \tag{8.19}
\]

Let such a pair \(\theta_1, \theta_2\) be fixed. Since the partial sums

\[
S_k = n^{-1/2} \sum_{j \leq k} (\tilde{d}_j(\theta_2) - \tilde{d}_j(\theta_1)), \quad k = 1, \ldots, n
\]

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

\[
E|S_n|^{2p} = E|\tilde{D}_n(\theta_2) - \tilde{D}_n(\theta_1)|^{2p} \leq N_p E \left\{ n^{-1} \sum_{k \leq n} (\tilde{a}_k(\theta_2) - \tilde{a}_k(\theta_1))^2 \right\}^p
\]

for any \(p \geq 1\), with a universal constant \(N_p\). Now

\[
\tilde{\rho}_n(\theta_1, \theta_2)^2 := n^{-1} \sum_k (\tilde{a}_k(\theta_2) - \tilde{a}_k(\theta_1))^2
\]

\[
\leq 2\tilde{\rho}_n(\theta_1, \theta_2)^2 + 2n^{-1} \sum_k \left\{ E|d_k(\theta_2) - d_k(\theta_1)| \mid \mathcal{F}_{k-1} \right\}^2 \tag{8.20}
\]

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

\[
n^{-1} \sum_k \left\{ E|m_k \mid \mathcal{F}_{k-1} \right\}^2 \leq n^{-1} \sum_k E[m_k^2 \mid \mathcal{F}_{k-1}] E[|\delta_k| \mid \mathcal{F}_{k-1}]
\]

\[
\leq \left\{ n^{-1} \sum_k E[m_k^2 | \mathcal{F}_{k-1}] \right\}^{1/2} \left\{ n^{-1} \sum_k E[|\delta_k| | \mathcal{F}_{k-1}] \right\}^{1/2}
\]

\[
\leq \left\{ n^{-1} \sum_k E[m_k^4 | \mathcal{F}_{k-1}] \right\}^{1/2} \{H_n^c([\theta_1, \theta_2])\}^{1/2}.
\]

It follows that

\[
E\tilde{\rho}_n(\theta_1, \theta_2)^{2p} \leq K \left[ \theta_2 - \theta_1 \right]^{2p} + E \left\{ \left( n^{-1} \sum_k m_k^4 \right) H_n([\theta_1, \theta_2]) \right\}^{p/2}
\]

\[
+ E \left\{ \left( n^{-1} \sum_k E[m_k^4 | \mathcal{F}_{k-1}] \right) H_n^c([\theta_1, \theta_2]) \right\}^{p/2}
\]

\[
\leq K \left[ \theta_2 - \theta_1 \right]^{2p} + \left\{ E \left( n^{-1} \sum_k m_k^4 \right)^p E H_n([\theta_1, \theta_2]) \right\}^{1/2}
\]

\[
+ \left\{ E \left( n^{-1} \sum_k E[m_k^4 | \mathcal{F}_{k-1}] \right)^p E H_n^c([\theta_1, \theta_2]) \right\}^{1/2}
\]

\[
\leq K \left[ \theta_2 - \theta_1 \right]^{2p} + \sqrt{M_4 p} \left( E H_n([\theta_1, \theta_2]) + E H_n^c([\theta_1, \theta_2]) \right)^{1/2}
\]

where \(M_4 = n^{-1} \sum_k E m_k^4\). So given \(\eta > 1\) putting \(b = 2\eta\) we may choose \(p \geq 1\) 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). \(\Box\)
9. Additional material

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

\[
\gamma = \gamma + \psi \quad \text{where} \quad \psi(\theta_1, \theta_2) = p \cdot \lim n^{-1} \sum_{k \leq n} h_k(\theta_1) h_k(\theta_2).
\]  

(9.1)

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

\[
\gamma_n(\theta_1, \theta_2) = n^{-1} \sum_{k \leq n} \{ \tilde{d}_k(\theta_1) + h_k(\theta_1) \} \{ \tilde{d}_k(\theta_2) + h_k(\theta_2) \}
\]

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 \( E[\tilde{d}_k(\theta_2) | \mathcal{F}_{k-1}] = 0 \) 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 \( O(n) \).

\[\square\]

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 \( \epsilon_k, k \in \mathbb{Z} \) are i.i.d. random variables. As an immediate consequence of [55, Theorem 2] one has, under the conditions given there, that

\[
E \left[ \sup_{0 \leq s \leq r} (G_n(t) - G_n(s))^2 \right] = O(n^{-1} r^{1-2/q}), \quad r \geq \beta_n,
\]

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

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

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

\[
\sup_{0 \leq t \leq 2^{2-\ell}} |G_n(t) - G_n(s)| \leq K(2^{-\ell})^{\kappa} \quad \text{for every } \ell \text{ with } 2^{-\ell} \geq \beta_n.
\]

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 (\( \alpha \))-mixing sequence. In a first step we apply covariance inequalities due to E. Rio \cite{47} yielding an estimate of the variance of the increments of \( G_n \). Specifically, suppose that the mixing coefficients \( \alpha_n \) decay as \( n^{-\varrho} \) for some \( \varrho > 1 \). Let \( G \) denote the common CDF of the \( y_k \). Then as a consequence of \cite{47, Theorem 1.2} we get that for some finite constant \( K \)

\[
\text{Var}(G_n(t) - G_n(s)) \leq Kn^{-1}(G(t) - G(s))^{1-1/\varrho}, \quad s < t, \ n \geq 1.
\] (9.2)

**Proof.** (Cf. \cite[pp. 590,591]{47}.) We have \( G_n(t) - G_n(s) = n^{-1} \sum_{1 \leq k \leq n} \xi_k \), \( \xi_k = 1_{s < y_k \leq t} \). The common ‘quantile function’ \( Q_{\xi} \) of the \( \xi_k \) 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 \( \alpha^{-1}(u) = \sum_k 1_{\xi_k > u} = O(u^{-1/\varrho}) \) 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.

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

\[
\lim_{n \to \infty} n \text{Cov}(G_n(s), G_n(t)) = \sum_{k \in \mathbb{Z}} \text{Cov}(1_{y_0 \leq s}, 1_{y_k \leq t}) = \Lambda(s, t).
\] (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 \( \Lambda \) 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. \cite{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

\[
|V(t) - V(s)| \leq K(G(t) - G(s))^\delta, \quad s < t \quad \text{(almost surely, ‘a.s.’)}.
\] (9.4)

Now suppose that the CDF \( G \) is uniformly Hölder continuous with index \( \kappa \in (0, 1] \). Then by (9.4), \( V \) also fulfills, 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 \cite[Theorem]{57} or \cite[Theorem]{12}, Theorem
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

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

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

$$|G_n(t) - G_n(s)| \leq |G(t) - G(s)| + n^{-1/2}|V_n(t) - V_n(s)| + o(n^{-1/2})$$

This reduces to $O(r^\kappa)$ if we set $\beta_n = n^{-1/(2\kappa)}$, since then $n^{-1/2} = \beta_n^\kappa \leq r^\kappa$ for $r \geq \beta_n$. 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.

**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)

$$\sup_{\theta \in \mathbb{R}} |F_n1(\theta) - F(\theta) - n^{-1/2}V_n(\theta)| = o(n^{-1/2}) \quad \text{(a.s.)}$$

Arguing as from (9.2) to (9.4) we get $|V_n(-\theta)| \leq KF_1(-\theta)^\delta$ for all $\theta \geq 0$ (a.s.), where again $\delta$ 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

$$F_n1(-\theta) \leq F_1(-\theta) + n^{-1/2} |V_n(-\theta)| + o(n^{-1/2})$$

$$= O(\theta^{-\varrho}) + O(n^{-1/2} \theta^{-\varrho} \vartheta) + o(n^{-1/2}), \quad \theta \geq 1 \quad \text{(a.s.)}$$

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

$$F_n1(-\theta) = O(\theta^{-\varrho}) + O(\theta^{-q/2-\varrho}) + o(\theta^{-q/2}),$$

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 $1 \leq \theta \leq n^{1/q}$ (a.s.). On the other hand, if $\theta > n^{1/2}$ then

$$Pr[F_n1([-\theta, \theta]^c) > 0] = Pr[\max_{k \leq n} |x_{k1}| > \theta] \leq n\theta^{-q} \int_{|x| > n^{q/4}} |x|^q dF_1(x),$$

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 $\theta \geq 1$. 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 $q = 2\nu$.

**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 $nr^\alpha$, where $\alpha \in (0, 1]$ 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 $nr^\alpha \geq 1$ we may expect that $EH_n(J)^p = O(n^{-p}(nr^\alpha)^p) = O(r^{bp})$, while for $nr^\alpha \leq 1$ we should have $EH_n(J)^p = O(n^{-p}nr^\alpha) = O(n^{1-p}r^p)$. Now given $b > 1$, choose $p > b/\alpha$ and put $\beta_n = n^{-(p-1)/(b-\alpha)}$. Noting that $\beta_n \leq n^{-1/\alpha}$, we find that in case $nr^\alpha \geq 1$ we have both $EH_n(J)^p = O(r^{bp}) = O(r^b)$ and $r \geq \beta_n$. In case $nr^\alpha \leq 1$ we have $EH_n(J)^p = O(n^{1-p}r^p)$, which is $O(\beta_n^{b-\alpha}r^\alpha)$ 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.

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

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

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

$$h_k(\theta) := E[d_k(\theta) | F_{k-1}] = (\Phi(t_k) - \alpha) \cdot \{ Pr[t_k < z_{k1}] - Pr[t_k < z_{k2}] \} \quad \text{(9.6)}$$

where $Pr$ refers to the $z_{kt}$ (resp. $u_{kt}$), everything else being considered as non-random (given $F_{k-1}$). 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 $k$, $\theta$), proving that $H_+^\alpha$ holds iff $\tau_1 \leq \tau_2$.

**Acknowledgments**

This work was funded by the European Union Seventh Framework Programme under grant agreement 290976. The Klaus Tschira Foundation provided infrastructural support at the Heidelberg Institute for Theoretical Studies (HITS). We thank the referees, Tilmann Gneiting, Alexander Jordan, and Wolfgang Polonik for valuable comments and discussions.
References


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


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