

Nonmanipulable Bayesian Testing*

Colin Stewart[†]

August 2009

Abstract

This paper considers the problem of testing an expert who makes probabilistic forecasts about the outcomes of a stochastic process. I show that, as long as uninformed experts do not learn the correct forecasts too quickly, a likelihood test can distinguish informed from uninformed experts with high prior probability. The test rejects informed experts on some data-generating processes; however, the set of such processes is topologically small. These results contrast sharply with many negative results in the literature.

1 Introduction

In settings ranging from economics and politics to meteorology, ostensible experts offer probabilistic forecasts of a sequence of events. Testing the reliability of these forecasts can be problematic since probabilistic forecasts of individual events are not generally falsifiable. It is tempting to think that this problem should disappear with the collection of enough data; that is, that an appropriately designed test could reliably distinguish true experts

*I am grateful to Martin Osborne, Marcin Peški, Jakub Steiner, and seminar participants at SFU for helpful comments. This work is supported by a SSHRC research grant.

[†]Department of Economics, University of Toronto. Email: colin.stewart@utoronto.ca

from charlatans by comparing a long sequence of forecasts to the realized sequence of events. However, recent work has shown that under very general conditions, no such test exists (see Olszewski and Sandroni (2008) and Shmaya (2008)). Any test that passes experts who know the true data-generating process can also be passed on *any* realized data by experts who know nothing, but who choose their forecasts appropriately.

Negative results of this type are typically based on the assumption that the tester is non-Bayesian and completely agnostic about the data-generating process. If instead the tester forms beliefs about the true process, then effective testing is generally possible. In this paper, I construct a test that delivers the correct verdict with high prior probability as long as uninformed experts, by updating the tester's prior, are unlikely to quickly learn the correct forecasts. In addition, the test is robust in the sense that informed experts fail only on a topologically small set of data-generating processes.

Consider a tester who believes that there is some true distribution P generating the data, but who faces uncertainty about P captured by some distribution μ . The distinction between the distributions μ and P can be understood in terms of the standard distinction between *epistemic* and *aleatory* uncertainty. Epistemic uncertainty results from a lack of knowledge of a physical process that could in principle be eliminated. Aleatory uncertainty is inherent to the environment. In this terminology, the distribution P captures the aleatory uncertainty whereas the prior distribution μ captures the tester's epistemic uncertainty.

Two natural requirements arise for a tester who forms beliefs μ over the data-generating process. First, true experts should pass the test with high probability relative to these beliefs. More precisely, if the data are generated by a distribution P , an expert forecasting according to P should pass the test on a set of outcomes with large P -probability, except perhaps for a set of distributions P having small μ -probability. Second, an expert who does not know the true process—but does know the tester's prior—should not be able to

pass the test. That is, no matter what mixed forecasting strategy σ a false expert uses, she should fail the test with high (σ, P) -probability for a set of distributions P having large μ -probability.

Designing a test with both of these properties is not possible for every prior μ . For instance, if μ assigns probability $1/2$ to each of two distributions P_1 and P_2 , then a true expert must pass with high probability on both distributions, in which case a false expert can also pass with probability close to $1/2$ simply by forecasting according to either one of P_1 or P_2 .

I show below that there exists a likelihood ratio test satisfying both desiderata whenever forecasts derived from the tester's prior do not converge to the true probabilities quickly, or do so only with small μ -probability. The test is based on a comparison of the likelihood assigned to the realized data by the expert's forecasts to that assigned by the tester's forecasts (obtained through Bayesian updating of μ given the realized data to date). The expert passes the test if two conditions are satisfied: (i) the likelihood associated with the expert's forecasts is eventually higher than that associated with the tester's, and (ii) the expert's forecasts do not converge to those of the tester too quickly.¹ For a true expert, condition (i) is almost surely satisfied whenever condition (ii) holds. Thus a true expert fails the test only when the tester learns the true probabilities independently of the expert's forecasts (in which case the tester may have no need for the expert). A false expert, on the other hand, cannot manipulate this test. Regardless of the false expert's forecasting strategy, she fails the test with probability 1 with respect to the tester's prior and her own randomization. Intuitively, condition (ii) forces the false expert to make forecasts that are very unlikely to outperform the tester's own when the data-generating process is distributed according to the tester's beliefs.

¹More precisely, the sum of the squared differences between the two forecasts diverges.

One might worry that allowing true experts to fail with small probability with respect to a fixed prior nonetheless allows them to fail on a large set of distributions, and thus fails to be robust to incorrect priors. It turns out that, regardless of the prior μ , the true expert fails the likelihood ratio test with positive probability only on a topologically small set of distributions (in the sense of Baire category). Therefore, the test also has desirable properties from the perspective of a non-Bayesian tester who wants to avoid manipulation by false experts without failing true experts on a large set of distributions. Such a tester can use the test corresponding to a prior μ chosen so that, under this prior, the tester does not learn the true probabilities too quickly. The set of distributions on which true experts fail the test is then small in both a topological and a measure-theoretic sense.

For certain simple priors, constructing tests that are effective with respect to those priors is easy. For example, if the prior is non-atomic and assigns probability 1 to a set of i.i.d. processes, then the tester only needs to check whether the frequency of outcomes converges to the expert’s first-period forecast. The results of this paper go beyond this observation in two significant ways. First, I show that designing such a test is possible for any prior provided that the uninformed expert is truly uninformed (in the sense that she cannot learn the correct forecasts quickly with positive prior probability). Second, tests such as that described above for i.i.d. distributions typically rely heavily on the prior insofar as true experts fail the test on a topologically large set of distributions—albeit a set to which the prior assigns small probability. The test proposed here does not suffer from this deficiency; a true expert passes the test for “most” distributions.

2 Related literature

Previous literature has identified an important distinction for tests that are required to pass true experts on every data-generating process. Dekel and Feinberg (2006) and Olszewski

and Sandroni (2009a) have obtained strong positive results if the tester can ask the expert to report the entire distribution P over infinite sequences of data. In many real-world settings, however, the tester can observe only the single-period forecasts made by the expert along the realized data sequence. In this case, Olszewski and Sandroni (2008) and Shmaya (2008) have shown that *every* test that passes true experts can be manipulated by false experts.² Thus any non-manipulable test that passes true experts necessarily violates the prequential principle (Dawid (1984)), which states that tests of forecasts should be based only on the forecast probabilities along the realized path of data, not on forecasts following hypothetical events that did not occur. In this paper, I restrict attention to this sequential setting to highlight the contrast with these negative results.

Olszewski and Sandroni (2009b) were the first to consider tests that fail true experts on some distributions, referring to the set on which they pass as a *paradigm*. They prove negative results when the paradigm is closed and convex, and identify a nonmanipulable test that fails true experts only if the single-period probabilities are close to 50% too often. Al-Najjar, Sandroni, Smorodinsky, and Weinstein (2008) propose an alternative paradigm in which effective testing is possible. Fortnow and Vohra (2009) take a different approach, obtaining positive results by restricting the class of strategies available to the false expert.

The likelihood test proposed here bears a close resemblance to the comparative test suggested by Al-Najjar and Weinstein (2008). When the expert is uninformed, the tester's forecasts play a role akin to the informed expert's forecasts in their setting. However, there is an essential difference. In our setting, the uninformed expert knows the tester's prior, and hence can predict the tester's forecasts, but is prevented from following these forecasts by the definition of the test. In Al-Najjar and Weinstein's setting, the uninformed expert

²These results represent the culmination of a line of research considering various classes of tests, beginning with Foster and Vohra (1998). See also Lehrer (2001), Olszewski and Sandroni (2009b), Sandroni (2003), Sandroni et al. (2003), and Vovk and Shafer (2005).

fails to manipulate the test because she is unable to predict the forecasts that the informed expert will make and can only pass by making forecasts close to those of the informed expert. Section 7 below elaborates on the relationship between the present setting and tests of multiple experts.

Like the present paper, Olszewski and Sandroni (2009b) also consider the probability of manipulation with respect to some prior measure, but in their case the prior is defined directly over the set of realizations. Proposition 3 of their paper shows that a false expert can pass on a set of realizations having large measure with respect to any fixed prior as long as the test does not reject correct forecasts based on the data-generating process. Thus it is crucial that our test does not pass true experts on every data-generating process. Otherwise, taking the reduction of the tester's prior μ to be the measure in Olszewski and Sandroni's result would show that the false expert can manipulate with high μ -probability.

3 The test

An outcome is realized in each period $t = 1, 2, \dots$. For simplicity, the set of possible outcomes in each period is taken to be $\{0, 1\}$. Let $\Omega = \{0, 1\}^{\mathbb{N}}$ denote the set of realizations endowed with the product topology, where $\mathbb{N} = \{1, 2, \dots\}$. We denote by $\omega = (\omega_t)_{t=1}^{\infty}$ a generic element of Ω . The realization ω is generated according to a probability distribution $P \in \Delta(\Omega)$, where $\Delta(X)$ denotes the set of probability distributions defined on the Borel subsets of a set X . The tester forms a prior belief about the distribution P according to some $\mu \in \Delta\Delta(\Omega)$, where $\Delta(\Omega)$ is endowed with the weak* topology. Given any $Q \in \Delta(\Omega)$ and any finite history $\omega^t = (\omega_1, \dots, \omega_t) \in \{0, 1\}^t$ of outcomes, we will write $Q(\omega^t) \in \Delta\{0, 1\}$ for the probability distribution over ω_{t+1} conditional on ω^t having occurred.

In each period $t = 1, 2, \dots$, the expert reports a *forecast* (or *prediction*) $p_{t+1} \in \Delta\{0, 1\}$, interpreted as the probability distribution over ω_{t+1} given the realized outcome so far.

Before reporting p_{t+1} , the expert perfectly observes the history $\omega^t = (\omega_1, \dots, \omega_t)$ of realized outcomes to date; thus each forecast p_{t+1} should be interpreted as a statement about probabilities for outcomes in period $t + 1$ *conditional* on the history of previous outcomes. The tester also observes each realized outcome ω_t .

There are two kinds of expert: true (or informed) and false (or uninformed). True experts observe the distribution P before choosing their forecasts and simply report the true one-step-ahead probabilities given the history of realized outcomes to date. That is, following each finite history of realizations ω^t , true experts report the forecast $p_{t+1} = P(\omega^t)$.^{3,4} False experts know only the tester's prior μ and may choose their forecasts according to any mixed strategy. Fixing μ , a (behavioral) strategy for the false expert is therefore a collection $\sigma = \{\sigma_t\}_{t=1}^\infty$ of functions

$$\sigma_t : \mathcal{H}_{t-1} \longrightarrow \Delta\Delta\{0, 1\},$$

where $\mathcal{H}_{t-1} = \{0, 1\}^{t-1} \times (\Delta\{0, 1\})^{t-1}$ is the set of histories $(\omega^{t-1}, (p_1, \dots, p_{t-1}))$ of realized outcomes and forecasts up to period $t - 1$. Thus the false expert's forecast in each period may depend on both the realized outcomes to date and the realizations of her own previous randomized choices of forecasts.

Note that, by the Kolmogorov Extension Theorem, distributions $P \in \Delta(\Omega)$ are in direct correspondence (up to sets of measure zero) with complete families of one-step-ahead forecasts $P(\omega^t) \in \Delta(\{0, 1\})$, one for each finite history ω^t of outcomes. Thus a distribution $P \in \Delta(\Omega)$ corresponds to a pure forecasting strategy for a false expert.

³Note that, in order to employ this strategy, true experts need not know the full distribution P . It is enough for them to know only the one-step-ahead conditional probabilities along the realized outcome path.

⁴In light of this definition, the distribution P captures exactly the knowledge that an informed expert has about the data-generating process. For instance, if a clairvoyant true expert can foresee the outcome of each toss of a fair coin, then P must assign probability 1 to the actual sequence of outcomes, not equal probabilities to the two outcomes in each period.

The present paper focuses on a likelihood ratio test that, given the tester's prior μ , passes or fails the expert depending only on the realized outcome $\omega \in \Omega$ and the sequence (p_1, p_2, \dots) of forecasts reported by the expert. The test compares the likelihood the expert's predictions assign to the realized data to that of the tester's predictions (given by expectations based on the tester's posterior beliefs in each period). The expert passes the test only if her predictions assign a higher likelihood to the data *and* are sufficiently different from the tester's forecasts. Formally, given μ , define the mean distribution $\bar{P} = \int_{\Delta(\Omega)} P d\mu(P)$. The test score following any finite history $h^t \in \mathcal{H}_t$ of outcomes and expert forecasts is defined to be

$$S(h^t) = \frac{p_1(\omega_1) \cdots p_t(\omega_t)}{\bar{p}_1(\omega_1) \cdots \bar{p}_t(\omega_t)},$$

where $p_s(\omega_s)$ denotes the probability that the expert's forecast for period s assigned to the realized outcome ω_s , and $\bar{p}_s(\omega_s) = \bar{P}(\omega^{s-1})(\omega_s)$ denotes the one-step-ahead conditional probability assigned to the realized outcome ω_s by the mean distribution conditional on the realized history ω^{s-1} to date. Note that the event that $\bar{p}_t(\omega_t) = 0$ for some t occurs with μ -probability 0; in that case, the expert passes the test. Otherwise, the expert passes the test if

1. $\liminf_{t \rightarrow \infty} S(h^t) > 1$, and
2. $\sum_t (p_t(1) - \bar{p}_t(1))^2$ diverges.

This test fails true experts if the tester's forecasts converge to the correct probabilities sufficiently quickly. If the tester uses the forecasts to choose an actions in some decision problem in each period, then when this convergence occurs, the additional value of the expert's forecasts becomes small even when they are correct, so failing to pass a true expert causes little harm to the tester (see Echenique and Shmaya (2008), who show that, as a tester becomes patient, using an incorrect theory is harmful only if the likelihood ratio

is unbounded).

If the test was based only on the likelihood ratio without the additional condition that forecasts differ from those of the tester, then the test could be easily manipulated with positive probability simply by choosing forecasts different from the tester's until the likelihood ratio exceeds 1, then choosing forecasts identical to the tester's forever after. Some such strategy succeeds with positive probability as long as the tester does not assign probability 1 to some Dirac measure.

It is worth noting that, while the test has properties that are desirable from the perspective of a Bayesian tester, the test itself is not Bayesian insofar as it does not depend on the tester's belief about the expert's knowledge. An alternative approach would be for the tester simply to update the probability that the expert knows the true distribution based on the forecasts and the realized data. In order to do so, however, the tester would have to form beliefs about the false expert's forecasting strategy. Instead, the likelihood ratio test proposed here is independent of these beliefs and exhibits the desired properties for *every* strategy a false expert might use (and hence with respect to any beliefs the tester might form about the false expert's strategy).

3.1 The tester's decision problem

The typical requirement in the literature that true experts must pass with high probability for *every* true distribution is very strong and difficult to justify in terms of the tester's decision problem. For natural payoffs, this requirement arises only if the tester has an extreme form of ambiguity aversion. To see this, consider a tester whose preferences are represented by the maxmin expected utility of Gilboa and Schmeidler (1989). For simplicity, suppose that the tester's belief about the type of the expert is constant and assigns positive probability to each of the two types. The tester forms a set S of priors over the set

$\Delta(\Omega)$ of possible true distributions. Suppose moreover that the tester receives a positive payoff for passing a true expert, a positive payoff for failing a false expert, and a payoff of 0 for failing a true expert or passing a false expert.

Consider a test T with no type I error that requires the expert to report a full distribution $P \in \Delta(\Omega)$ (as opposed to a sequence of one-step-ahead forecasts). For any fixed distribution Q , let T_Q be the test that delivers the same verdict as T whenever the expert forecasts a distribution other than Q but fails the expert if she forecasts Q (regardless of the realized data). In order for the tester to strictly prefer T to T_Q , Q must be an atom of some prior in S . The requirement that there be no type I error therefore corresponds to this setting if for every $Q \in \Delta(\Omega)$, there exists some prior in S having Q as an atom. This property assumes that the tester has a strong form of ambiguity aversion.

In this paper, I (implicitly) consider a tester who is a classical expected utility maximizer. As above, the tester's payoff depends only on the type of the expert and whether the expert passes or fails the test. I assume that the tester prefers to pass the true expert and fail the false expert. The tester has a prior belief over the set of possible distributions and a fixed prior belief over the type of the expert that is independent of the true distribution. In this setting, the tester's goal is to pass true experts and fail false experts, both with high probability with respect to her prior.

4 Properties of the test

A Bayesian tester with prior μ should desire the following properties of a test:

1. A true expert should pass the test with high P -probability for a set of distributions P having large μ -probability.
2. No matter what strategy σ she uses, a false expert should fail the test with high

(P, σ) -probability for a set of distributions P having large μ -probability.

As noted above, for some priors μ it is not possible to satisfy both of these criteria simultaneously. The following result indicates that these properties can be satisfied under a general condition on μ , namely that the μ -probability that the tester learns the true probabilities sufficiently quickly is small. Under this condition, the second property holds in a strong form: the false expert almost surely fails to manipulate the test.

Let $\varepsilon = \Pr_\mu \left(\sum_t (p_t(1) - \bar{p}_t(1))^2 \text{ converges} \right)$, where $p_t = P(\omega^{t-1})$ denotes the true probability density in period t conditional on ω^{t-1} . The following proposition shows that, when ε is small, the likelihood ratio test from Section 3 satisfies both of the above desiderata.⁵

Proposition 1. *For the likelihood ratio test described above, (i) a true expert passes the test with μ -probability $1 - \varepsilon$, and (ii) for any forecasting strategy σ , the false expert passes the test with (μ, σ) -probability 0 (regardless of the value of ε).*

Proof. The proof begins by extending one direction of Kakutani's Theorem for Product Martingales (see, e.g., Williams (1991)) to allow for dependence among the factors. First consider the case of a false expert using a pure forecasting strategy. Given any sequence of forecasts q_t , the single-period ratio $Y_t = q_t(\omega_t) / \bar{p}_t(\omega_t)$ satisfies $E_\mu(Y_t | Y_1, \dots, Y_{t-1}) = 1$ for all t . The product $X_t = \prod_{k \leq t} Y_k$ is a non-negative martingale with respect to \bar{P} that is bounded in L^1 . By the martingale convergence theorem, $(X_t)_{t=1}^\infty$ converges almost surely to some random variable X_∞ .

⁵For many priors we have $\varepsilon = 0$, in which case Proposition 1 implies that the test almost surely gives a correct verdict on whether or not the expert is informed. Section 6 provides a natural example in which $\varepsilon = 0$ even though the tester learns the true probabilities in the limit as $t \rightarrow \infty$.

Let $a_t(Y_1, \dots, Y_{t-1}) = E_\mu(\sqrt{Y_t} | Y_1, \dots, Y_{t-1})$ and let

$$Z_t = \prod_{k \leq t} \frac{\sqrt{Y_k}}{a_k(Y_1, \dots, Y_{k-1})}.$$

Then Z_t is a nonnegative martingale. Moreover, by inductively applying the law of iterated expectations, one sees that $E_\mu(Z_t) = 1$ for all t . In particular, $(Z_t)_{t=1}^\infty$ is bounded in L^1 , and hence converges almost surely to some Z_∞ by the martingale convergence theorem. It follows that if $\prod_t a_t(Y_1, \dots, Y_{t-1}) = 0$ then $\prod_t \sqrt{Y_t} = 0$ almost surely, and hence $\prod_t Y_t = 0$ almost surely. Therefore, X_t almost surely converges to zero whenever $\prod_t a_t(Y_1, \dots, Y_{t-1})$ converges to zero.

It remains to show that $\prod_t a_t(Y_1, \dots, Y_{t-1}) = 0$ whenever $\sum_t (q_t(1) - \bar{p}_t(1))^2$ diverges.

Note that

$$\begin{aligned} a_t(Y_1, \dots, Y_{t-1}) &= E_\mu(\sqrt{Y_t} | Y_1, \dots, Y_{t-1}) = \bar{p}_t(1) \sqrt{\frac{q_t(1)}{\bar{p}_t(1)}} + (1 - \bar{p}_t(1)) \sqrt{\frac{1 - q_t(1)}{1 - \bar{p}_t(1)}} \\ &= \sqrt{\bar{p}_t(1)q_t(1)} + \sqrt{(1 - \bar{p}_t(1))(1 - q_t(1))}. \end{aligned}$$

Letting $\delta_t = q_t(1) - \bar{p}_t(1)$, we have

$$a_t(Y_1, \dots, Y_{t-1}) = \sqrt{\bar{p}_t(1)^2 + \delta_t \bar{p}_t(1)} + \sqrt{(1 - \bar{p}_t(1))^2 - \delta_t(1 - \bar{p}_t(1))}.$$

For fixed δ , the function

$$f(p) = \sqrt{p^2 + \delta p} + \sqrt{(1 - p)^2 - \delta(1 - p)}$$

is maximized when $p = \frac{1-\delta}{2}$, with maximum value $\sqrt{1-\delta^2}$. Hence we have

$$\prod_t a_t(Y_1, \dots, Y_{t-1}) \leq \prod_t \sqrt{1-\delta_t^2},$$

so it suffices to show that $\prod_t \sqrt{1-\delta_t^2} = 0$ whenever $\sum_t (q_t(1) - \bar{p}_t(1))^2 = \sum_t \delta_t^2$ diverges.

Note that $\prod_t \sqrt{1-\delta_t^2} = 0$ if and only if $\sum_t \log(1-\delta_t^2)$ diverges. Since

$$\lim_{x \rightarrow 0} \frac{-\log(1-x)}{x} = \lim_{x \rightarrow 0} \frac{1}{1-x} = 1,$$

the limit comparison test implies that $\sum_t \log(1-\delta_t^2)$ converges if and only if $\sum_t \delta_t^2$ does, as needed.

We have shown that for any pure forecasting strategy, a false expert μ -almost surely fails the test whenever $\sum_t (q_t(1) - \bar{p}_t(1))^2$ diverges. Since the test always fails the expert if this sum converges, it follows that the expert almost surely fails the test for every pure forecasting strategy. Therefore, for any mixed strategy, taking expectations with respect to the mixing distribution implies that the expert almost surely fails the test.

The same argument with the roles of the numerator and denominator reversed shows that, conditional on $\sum_t (p_t(1) - \bar{p}_t(1))^2$ diverging, a true expert passes the test P -almost surely. \square

Proposition 1 shows that there exists a test that effectively distinguishes between true and false experts as long as ε is small. Moreover, the test perfectly satisfies the desired criteria when $\varepsilon = 0$. In order for ε to be positive, the tester must possess considerable knowledge of the data-generating process: not only must the tester's forecasts approach the true probabilities (with positive probability), they must do so quickly. Indeed, the condition that $\sum_t (p_t(1) - \bar{p}_t(1))^2$ converges implies that $t^{1/2} |p_t(1) - \bar{p}_t(1)|$ vanishes as

$t \rightarrow \infty$ on some density one sequence,⁶ which is the usual standard for fast convergence of forecasts (see, e.g., Sandroni and Smorodinsky (1999)).

As an alternative to the decision problem described in Section 3.1, in which the tester simply wants to know whether or not the expert is informed, Echenique and Shmaya (2008) and Olszewski and Pęski (2008) consider a tester who is directly interested in predicting outcomes in each period. In their settings, in each period t , the tester chooses an action a_t before observing the realized outcome ω_t and receives a flow payoff depending on the pair (a_t, ω_t) . Accurate forecasts may help the tester to choose actions generating higher payoffs. One can show that if $\sum_t (p_t(1) - \bar{p}_t(1))^2$ converges, then the payoff loss to the tester from following her own predictions instead of those of a true expert vanish as the tester becomes very patient. Proposition 1 thus offers a counterpoint to the “you won’t harm me if you fool me” result of Echenique and Shmaya (2008), which shows the existence of a test that passes true experts and creates little loss for a patient tester if a false expert passes. Here true experts may fail, but again there is little loss for the tester.

5 Robustness

The tester’s prior places more structure on the testing problem than is the norm in the literature. In particular, every prior assigns probability 1 to some topologically small set (in the sense of Baire category). One might worry, then, that the test rejects true experts on many distributions. However, this is not the case; the set of distributions on which true experts fail the test is topologically small. In order to prove this, we begin with a general result on merging of opinions. This result may be of independent interest in light of the central role played by merging in the literature on learning and reputation in repeated games (see, e.g., Kalai and Lehrer (1993) and Sorin (1999)).

⁶The density of an increasing sequence $(t_k)_{k=1}^{\infty}$ with $t_k \in \mathbb{N}$ is defined to be $\limsup_{k \rightarrow \infty} k/t_k$.

Given distributions $P, Q \in \Delta(\Omega)$, we say that Q *weakly merges with P at ω* if for every $\delta > 0$ there exists some T such that

$$|p_t(1) - q_t(1)| < \delta \text{ for all } t > T,$$

where p_t and q_t denote the one-step-ahead distributions along ω associated with P and Q respectively. The distribution Q is said to weakly merge with P with probability π if

$$P(\{\omega \mid Q \text{ weakly merges with } P \text{ at } \omega\}) = \pi.$$

Most work on merging has focused on global concepts that require merging to occur with probability one.⁷ The notion of weak merging introduced by Kalai and Lehrer (1994) corresponds to almost sure weak merging in our terminology.

Proposition 2. *For any distribution $Q \in \Delta(\Omega)$, the set of distributions P such that Q weakly merges with P with positive probability is category I.*

Proof. The proof follows an approach similar to that of Feinberg and Stewart (2008), Proposition 2. Choose any $\delta \in (0, 1/4)$. For each finite history h and $\varepsilon \in (0, 1)$, let $S(h, \varepsilon) \subset \Delta(\Omega)$ be the set of distributions P such that $P(h) \geq \varepsilon$ and

$$\Pr_P(|p_t(1) - q_t(1)| < \delta \text{ for all } t \geq \tau \mid h) \geq \varepsilon, \tag{1}$$

where τ is the length of h and p_t and q_t denote one-step-ahead distributions as above. The set of distributions to which Q weakly merges with positive probability is contained in the countable union

$$\bigcup_{\text{finite histories } h} \bigcup_n S(h, \varepsilon^n),$$

⁷Lehrer and Smorodinsky (2000) is an exception. They study a weaker notion of pointwise merging.

so it suffices to show that each $S(h, \varepsilon)$ is nowhere dense.

First we will prove that $S(h, \varepsilon)$ is closed by showing that its complement is open. Consider $P \notin S(h, \varepsilon)$. Either $P(h) < \varepsilon$ or (1) fails. If $P(h) < \varepsilon$, then the set $\{P' \mid P'(h) < \varepsilon\}$ contains P , is open in $\Delta(\Omega)$, and is disjoint from $S(h, \varepsilon)$. If (1) fails, then there exists some $T > \tau$ such that

$$\Pr_P (|p_t(1) - q_t(1)| < \delta \text{ for all } t = \tau, \dots, T \mid h) = \varepsilon' < \varepsilon.$$

The set

$$\{P' \mid |P'(E) - P(E)| < \varepsilon - \varepsilon' \text{ for all events } E \text{ determined by time } T\}$$

contains P , is open, and is disjoint from $S(h, \varepsilon)$.

All that remains is to show that $S(h, \varepsilon)$ has empty interior. Again let τ denote the length of h . Fixing $P \in S(h, \varepsilon)$, we want to construct a sequence of distributions $P_n \notin S(h, \varepsilon)$ converging to P . Choosing P_n so that the one-step-ahead forecasts agree with P in every period $t \leq \tau + n$ and differ from P by at least 2δ following every history of length $\tau + n$ does the job. \square

Corollary 1. *The set of distributions $P \in \Delta(\Omega)$ for which a true expert fails the test with positive P -probability is category I.*

Proof. Suppose that a true expert fails with positive P -probability when the distribution is P . Then the sum $\sum_t (p_t(1) - \bar{p}_t(1))^2$ converges with positive P -probability. In particular, for any $\delta > 0$, there exists some finite history h_τ of length τ such that $P(h_\tau) > 0$ and, with positive probability, $|p_t(1) - \bar{p}_t(1)| < \delta$ for all $t \geq \tau$. Therefore, \bar{P} weakly merges with P with positive probability. \square

Proposition 2 improves upon the infinite horizon analogue of the test studied by Olszewski and Sandroni (2009b) that rejects experts who forecast close to 50% too often. Their test cannot be manipulated but rejects true experts on a set of distributions that is topologically small in a weaker sense than category I. One might think that forecasts close to 50% are relatively uninformative, and hence rejecting experts who make such forecasts is reasonable. Relative to a given prior, however, a forecast close to 50% can be very informative. The above test accounts for this feature by only failing true experts if their forecasts are uninformative relative to the tester's prior.

6 Example

The following example illustrates the power of the test even when the tester learns the true probabilities (but not too quickly). Suppose that the tester's prior μ can be parameterized by a uniform distribution over $\pi \in [0, 1]$, and for each π the true process is i.i.d. with probability π in each period. In this case, one can easily devise non-manipulable tests that pass true experts with μ -probability 1. One example is the test that passes the expert if and only if the empirical frequency converges to her first-period forecast p_1 . However, this test is not robust to incorrect priors in the sense of Proposition 2; true experts almost surely fail the test on a large set of distributions. The likelihood ratio test proposed above is robust in this sense and passes true experts with μ -probability equal to the probability with which the sum $\sum_t (\bar{p}_t - \pi)^2$ diverges. Even though the tester almost surely asymptotically learns the true probabilities in this setting (see, e.g., Doob (1949)), it turns out that this learning occurs sufficiently slowly for the test to be effective. The following claim, together with Proposition 1, shows that the test passes true experts with P -probability 1 for μ -almost every P , and false experts with probability 0.

Claim. *For each $\pi \in (0, 1)$, the sum $\sum_t (\bar{p}_t - \pi)^2$ diverges with probability 1.*

Proof. Following any finite history $\omega^t = (\omega_1, \dots, \omega_t)$, let $n_t = \sum_{s=1}^t \omega_s$. It is straightforward to show that $\bar{p}_{t+1}(\omega^t) = \frac{n_t+1}{t+2}$. Accordingly, consider

$$\sum_t \left(\frac{n_t+1}{t+2} - \pi \right)^2.$$

By the limit comparison test, this sum converges if and only if

$$\sum_t \left(\frac{n_t}{t} - \pi \right)^2$$

does.

Define the density of a sequence $(x_t) \in \{0, 1\}^{\mathbb{N}}$ to be $\limsup_T \frac{\sum_{t=1}^T x_t}{T}$. For $\varepsilon > 0$, define a random variable x_t^ε by

$$x_t^\varepsilon = \begin{cases} 1 & \text{if } t \left(\frac{n_t}{t} - \pi \right)^2 > \varepsilon \\ 0 & \text{otherwise.} \end{cases}$$

Note that since $\sum_t \left(\frac{n_t}{t} - \pi \right)^2 = \sum_t \frac{1}{t} t \left(\frac{n_t}{t} - \pi \right)^2$ and $\sum_t \frac{1}{t}$ diverges, the sum $\sum_t \left(\frac{n_t}{t} - \pi \right)^2$ diverges whenever the sequence $(x_t^\varepsilon)_t$ has positive density. Thus it suffices to show that, as ε vanishes, the probability that the sequence $(x_t^\varepsilon)_t$ has positive density tends to 1.

Let E_t^ε denote the event that $x_t^\varepsilon = 0$. For each k and T , we have

$$\Pr \left(\sum_{t=1}^T (1 - x_t^\varepsilon) \geq k \right) \leq \frac{1}{k} \sum_{t=1}^T \Pr(E_t^\varepsilon)$$

since the left-hand side is the probability that E_t^ε occurs at least k times up to time T , and whenever E_t^ε occurs k times, the right-hand side counts (upper bounds on) the associated probabilities at least k times. Note that, by the Central Limit Theorem, given $\delta > 0$, there exist T' large enough and $\varepsilon > 0$ small enough so that $\Pr(E_t^\varepsilon) < \delta/2$ for all $t > T'$. Thus in

the limit as $T \rightarrow \infty$, we have

$$\Pr\left(\sum_{t=1}^T(1-x_t^\varepsilon)\geq\eta T\right)\leq\frac{1}{\eta T}T\frac{\delta}{2}$$

for $\eta\in(0,1)$. In particular, for $\eta>1/2$, we have

$$\Pr\left(\sum_{t=1}^Tx_t^\varepsilon\leq(1-\eta)T\right)\leq\delta\tag{2}$$

as T tends to ∞ .

I claim that (2) implies that $(x_t^\varepsilon)_t$ has density at least $1-\eta$ with probability at least $1-\delta$. Since the choice of δ was arbitrary, this claim completes the proof. To prove the claim, suppose that it does not hold. Then there exists a set of sequences $(x_t^\varepsilon)_t$ with probability greater than δ each with density less than $1-\eta$. Hence there must exist some T'' such that for a subset of these sequences $(x_t^\varepsilon)_t$ of measure greater than δ , the frequency of ones in $(x_t^\varepsilon)_t$ up to time T is less than $1-\eta$ for all $T>T''$, violating (2). \square

7 Discussion

The test studied here may be thought of as a test of two experts in which one of the experts forecasts according to the tester’s prior μ . The test adds to a simple likelihood ratio test a condition that the expert passes only if her forecasts are sufficiently different from the tester’s. More generally, one could consider modifying any test of two experts so as to make the experts fail if their forecasts converge sufficiently quickly. The positive results above rely on the property of the original likelihood ratio test (without this modification) that a Bayesian who reports truthfully expects to pass the test against any fixed strategy of the other forecaster—in other words, against any false expert—unless the forecasts of the two

experts converge sufficiently quickly. In our setting, once the closeness condition is added, this property ensures that a true expert passes unless the tester's forecasts converge quickly to the true probabilities. This property also ensures that, for distributions generated by the tester's prior, a false expert cannot outperform the tester in terms of the likelihood ratio unless her forecasts converge quickly to those of the tester, in which case she fails the modified test.

Al-Najjar and Weinstein (2008) and Feinberg and Stewart (2008) propose multiple expert tests that cannot be manipulated. Both tests derive their power from properties similar to those described above. More precisely, true experts pass both tests, and in the presence of a true expert, a false expert can pass only if her forecasts are eventually close to those of the true expert. By modifying these tests to reject experts whose forecasts become close, and treating the tester as one of the two forecasters, positive results similar to Proposition 1 and Corollary 1 above could be obtained beginning from either test, although the precise closeness condition that must be added may differ.

It may seem counterintuitive that the proposed test encounters problems when the tester learns the truth. After all, a tester who knows the true probabilities can very easily check the validity of the expert's forecasts. The reason is that in this case the false expert also learns the truth since she knows the tester's prior. If the false expert makes forecasts that converge rapidly to the tester's forecasts then the data cannot reliably distinguish which of the two is closer to the truth.

To gain a partial intuition for the contrast between the positive results presented here and the negative results in the literature, consider the problem of designing a non-manipulable test with no type I error. The negative results can be understood by reformulating this testing problem as a zero-sum game between a forecaster who wants to choose forecasts to pass a given test and malevolent Nature who wants to choose the realization to

make the forecaster fail. If the Minmax Theorem holds in this setting, then the worst-case expected payoffs for true and false experts must be equal. It follows that if a true expert passes almost surely for any distribution chosen by Nature, then a false expert must also pass almost surely. For the above likelihood ratio test, given any prior for the tester, true experts fail on some distributions. This feature makes the worst-case payoff for a true expert equal to that of failure, which suffices to prevent manipulation by a false expert even if the set of distributions on which a true expert fails has small prior probability.

References

- Al-Najjar, N., A. Sandroni, R. Smorodinsky, and J. Weinstein (2008). Testing theories with learnable and predictive representations. Mimeo.
- Al-Najjar, N. and J. Weinstein (2008). Comparative testing of experts. *Econometrica* 76, 541–559.
- Dawid, P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A* 147, 278–292.
- Dekel, E. and Y. Feinberg (2006). Non-bayesian testing of a stochastic prediction. *Review of Economic Studies* 73, 893–906.
- Doob, J. (1949). *Colloques Internationaux du Centre National de la Recherche Scientifique*, Chapter Application of the theory of martingales, pp. 23–27.
- Echenique, F. and E. Shmaya (2008). You won’t harm me if you fool me. Mimeo.
- Feinberg, Y. and C. Stewart (2008). Testing multiple forecasters. *Econometrica* 76, 561–582.
- Fortnow, L. and R. Vohra (2009). The complexity of testing forecasts. *Econometrica* 77, 93–105.

- Foster, D. and R. Vohra (1998). Asymptotic calibration. *Biometrika* 85, 379–390.
- Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18, 141–153.
- Kalai, E. and E. Lehrer (1993). Rational learning leads to Nash equilibrium. *Econometrica* 61, 1019–1045.
- Kalai, E. and E. Lehrer (1994). Weak and strong merging of opinions. *Journal of Mathematical Economics* 23, 73–86.
- Lehrer, E. (2001). Any inspection is manipulable. *Econometrica* 69, 1333–1347.
- Lehrer, E. and R. Smorodinsky (2000). Relative entropy in sequential decision problems. *Journal of Mathematical Economics* 33, 425–439.
- Olszewski, W. and M. Pęski (2008). The principal-agent approach to testing experts. Mimeo.
- Olszewski, W. and A. Sandroni (2008). Manipulability of future-independent tests. *Econometrica* 76, 1437–1466.
- Olszewski, W. and A. Sandroni (2009a). A nonmanipulable test.
- Olszewski, W. and A. Sandroni (2009b). Strategic manipulation of empirical tests. *Mathematics of Operations Research* 34, 57–70.
- Sandroni, A. (2003). The reproducible properties of correct forecasts. *International Journal of Game Theory* 32, 151–159.
- Sandroni, A. and R. Smorodinsky (1999). The speed of rational learning. *International Journal of Game Theory* 28, 199–210.
- Sandroni, A., R. Smorodinsky, and R. Vohra (2003). Calibration with many checking rules. *Mathematics of Operations Research* 28, 141–153.

- Shmaya, E. (2008). Many inspections are manipulable. *Theoretical Economics* 3, 393–408.
- Sorin, S. (1999). Merging, reputation, and repeated games with incomplete information. *Games and Economic Behavior* 29, 274–308.
- Vovk, V. and G. Shafer (2005). Good randomized sequential probability forecasting is always possible. *Journal of the Royal Statistical Society Series B* 67, 747–763.
- Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.