Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 26 (2010) 498–510
international journal of forecasting
www.elsevier.com/locate/ijforecast

Finding profitable forecast combinations using probability scoring rules

Andrew Grant, David Johnstone*

School of Business, University of Sydney, NSW 2006, Australia

Abstract

This study examines the success of bets on Australian Football League (AFL) matches made by identifying panels of highly proficient forecasters and betting on the basis of their pooled opinions. The data set is unusual, in that all forecasts are in the form of probabilities. Bets are made "on paper" against quoted market betting odds according to the (fractional) Kelly criterion. To identify expertise, individual forecasters are scored using conventional probability scoring rules, a "Kelly score" representing the forecaster's historical paper profits from Kelly-betting, and the more simplistic "categorical score" (number of misclassifications). Despite implicitly truncating all probabilities to either 0 or 1 before evaluation, and thus losing a lot of information, the categorical scoring rule appears to be a propitious way of ranking probability forecasters. Bootstrap significance tests indicate that this improvement is not attributable to chance.

Keywords: Probability scoring rule; Kelly betting; Kelly probability score; Combining probability forecasts; Economic forecast evaluation; Probability football

1. Introduction

A fundamental question in any field is whether experts in that field are expert enough that they can predict the future rather than merely recounting the past. A more pointed question is whether expert forecasts are sufficiently accurate that the forecaster's principal can use them to make money. To an economist, this is the litmus test of forecasting expertise (e.g. Clements, 2004; Granger & Pesaran, 2000; Lopez, 2001; Stekler, 1994). Evaluating forecasts based on their economic consequences dates back at least to Thompson and Brier (1955), and is a long held standard in the theory of probability forecasting, particularly in applications such as meteorology, where inaccurate forecasts are costly and accurate forecasts offer enormous economic opportunities.

Winkler and Murphy (1979) demonstrate the economic utility of probability forecasts ("Rain tomorrow with probability $p = 0.65$") over their less informative categorical forms ("Rain tomorrow"). The simplest form of this argument is that if a forecaster announces "Rain" whenever $p > 0.5$, say, then it is more informative to know the exact value of his subjective p (e.g. $p = 0.95$) than merely that he predicts

* Corresponding author.
E-mail address: d.johnstone@econ.usyd.edu.au (D. Johnstone).

0169-2070/$ - see front matter
doi:10.1016/j.ijforecast.2010.01.002

"rain", signifying $p \in (0.5, 1]$. Once users observe the relevant likelihood functions $f(p|Rain)$ and $f(p|NotRain)$ of the forecaster's different possible probability announcements p, it is easy to show that the expected value (EVII) of a probability forecast generally exceeds the expected value of the same forecast truncated to a category before announcement. This extra resolution comes from the likelihood ratio $f(p|Rain)/f(p|NotRain)$ generally being more pronounced than the equivalent ratio based on the duller signal "Rain" or "Not Rain".

The superiority (in principle) of probability forecasting over categorical forecasting is widely advocated (e.g. de Finetti, 1970; Lindley, 1982; Savage, 1971, p. 795; Winkler & Murphy, 1968, p. 752, 1979, p. 138), but is not essential to this paper. Our objective is to demonstrate that even if we adopt a probability forecasting framework, there may still be something to be gained by transforming each forecaster's probabilities into categorical forecasts and merely counting misclassifications. There are precedents in the literature for selecting experts or expert forecasting systems according to multiple different performance criteria (Clemen, Murphy, & Winkler, 1995, p. 135; Wallsten, Budescu, Erev, & Diederich, 1997, p. 249); however, categorical forecast evaluation is not conspicuous as a natural way of scoring forecasts which are expressed as probabilities. The empirical evidence detailed in this paper suggests that the marginal economic value of a "categorically most expert" forecaster can be material, even after the most expert forecasters have been identified by the standards of conventional probability scoring rules.

2. Experimental design

The experiment is based on a sample of probability forecasts made by contestants in a web-based football "tipping" competition, which has been conducted by the computer science faculty at Monash University in Melbourne, Australia, since 1995.[1] Each year, the competition attracts a highly competitive cohort of forecasters. The football code is known as AFL, short for Australian Football League, and has a quasi-religious following among its supporters.

The experience, commitment and depth of technical understanding typical of the competition's most devoted forecasters offers the possibility of a "free ride" for those who follow them with bets. A further aspect of AFL that may open opportunities for profitable betting is that there is fiercely parochial support for most of the teams in the competition, and hence there is the possibility that the betting odds are distorted by a significant flow of emotionally inspired betting (Levitt, 2004, p. 240; Wolfers & Zitzewitz, 2004, p. 118). If market odds tend to fall away from the subjective odds of highly expert forecasters, as may occur when there is a large volume of partisan betting relative to the volume of informed or "smart" money, then there may be room for systematic profit-taking.

To test for such tradeable market inefficiencies, we devise a betting strategy using the forecasters' predictions as if they were available to us before the game. The essential features of our experiment, including the data, the forecast attributes by which forecasters' past performances are evaluated and the decision rule by which bets are selected, are described briefly below. We make no attempt to allow for gaming that may cause participants to lodge probabilities other than their true beliefs (cf. Lichtendahl & Winkler, 2007; Lindley, 1982, p. 8; Kadane & Winkler, 1988; Winkler, 1967, p. 1115, 1971, p. 680). In taking this approach, we test whether the forecasting expertise on show, particularly among the most successful forecasters within the competition, is substantial enough that the chosen "best" forecasters have a residual economic value even after (or perhaps because of) the effects of gaming by some or all individuals.

2.1. The data

Our data set contains the probability predictions of the outcomes of AFL matches made by forecasters participating in the Monash University Probabilistic Football Tipping Competition over the seven seasons from 1999 to 2005.[2] Each season runs for 22

[1] Details of the Monash University competition, designed and operated by David Dowe and Torsten Seemann, can be found at: http://www.csse.monash.edu.au/~footy/.

[2] The earliest reported studies of probabilistic football prediction competitions appear to be those of de Finetti (1982, pp. 4–6) and Winkler (1971), who conducted such competitions on American football in 1966 and Italian football in 1971, respectively. Chen, Chu, Mullen, and Pennock (2005) and Servan-Schreiber, Wolfers, Pennock, and Galebach (2004) provide the most recent studies, both using data obtained from the online American football forecasting competition found at http://probabilityfootball.com.

weeks or "rounds", where there are 8 separate games in each pre-finals round, making a total of 176 games per season. The competition is open to everyone, with no entry fees and no prizes. In each season, the participants' aliases and cumulative scores are all shown on the competition website after each "round" or week's matches, meaning that everyone knows their own position in the rankings and how far behind the leaders they are.

The data set includes the names of the two teams playing, the ground at which the game is played, and the bookmaker's odds on each team winning. The odds are those offered before the game by TABSPORTSBET, the largest (by volume) of the several Australian bookmakers taking bets on AFL.

In each game of each round, participant i states a personal probability $p_i = \Pr(x = 1)$ of the "home" team winning that game. His competition score for that game is calculated by a form of the logarithmic probability score[3]

$$s_i = \begin{cases} 1 + \log_2(p_i) & \text{if } \textit{Home } (x = 1) \\ 1 + \log_2(1 - p_i) & \text{if } \textit{Away } (x = 0) \\ 1 + \dfrac{1}{2}\log_2(p_i\{1 - p_i\}) & \text{if the game was tied.} \end{cases}$$

The rules of AFL and the way in which it is played mean that tied games are extremely uncommon (about 1% overall). If there were more ties, it would be necessary to obtain probabilities in each game for at least two of the three possible outcomes. However, given how seldom ties occur, it is practical to only require participants to lodge a probability p_i of the home team winning. The contestant's probability of an "away" win, $\Pr(x = 0)$, is then taken as $(1 - p_i)$. In each season there are several hundred participants, but not all are seriously engaged. We retain in the sample only those individual forecasters who provide probabilities for every game in the season. The resulting sample sizes are between 41 and 50 forecasters per season.

2.2. Selecting the "best" forecasters

To identify those forecasters whose opinions are likely to be of the greatest economic worth, we score each of the forecasters using four distinct criteria. The first is the conventional $\log_e$ probability score, $s(p_i, x) = \ln(p_i^x(1 - p_i)^{1-x})$. The second is a raw form of the Brier score, $s(p_i, x) = -x(p_i)^2 - (1 - x)(1 - p_i)^2$. The third is the "old fashioned" categorical score $s(p_i, x) = xRnd(p_i) + (1-x)(1-Rnd(p_i))$, where $Rnd(p_i)$ equals 1 for $p_i > 0.5$ and 0 for $p_i < 0.5$. When $p_i = 0.5$ exactly, which in principle should never occur, since an expert forecaster should always favor one team (if only at the fourth decimal place), the forecaster is given a score of either 1 or 0 at random (effectively by "tossing a coin"). The forecaster is therefore either "lucky" or "unlucky", as occurs whenever someone with a probability of 0.5 is obliged to make a categorical forecast. This is one of the known defects of categorical forecasting. However, its effect in this study is not likely to be severe, given that forecast probabilities of exactly 0.5 are not very frequent among the most proficient forecasters, and where they occur, each forecaster's luck will probably balance out with that of his competitors. Finally, in the rare event of a tied game, both the log score and the categorical scores are set equal to zero for all forecasters. Savage (1971, pp. 797–798) provides an interesting justification for this.

The fourth method of scoring is called the "Kelly score", and captures the paper profit earned in the game in question by betting against the bookmaker's advertised odds according to (i) the "Kelly criterion" and (ii) the subjective probability p_i reported by the forecaster. This economic scoring rule measures the forecaster's hypothetical success or potential as a Kelly bettor. Its advantage is that it reveals the forecasters who are historically the most profitable, rather than merely the most accurate, in their beliefs. For example, imagine a forecaster who is generally accurate but whose "most accurate" probabilities tend to lie within the bookmaker's bid-ask spread. This forecaster makes few bets and tends to bet only when his probability judgments are least reliable, meaning that overall he makes little or no profit (Johnstone, 2007a).

2.2.1. The Kelly score

Consider an investor whose decision rule is to maximize the expected log utility of wealth at some future time T, $E[\ln(wealth_T)]$. This is a myopic decision rule, since $E[\ln(wealth_T)]$ is maximized by maximizing $E[\ln(wealth_t)]$ at each decision point $t-1$ prior

[3] The literature on "scoring rules" is summarized extensively by Winkler (1996).

to $t = T$ (Luenberger, 1998, p. 426). A bet or portfolio of bets designed to maximize the expected log utility is known to professional gamblers as "Kelly-betting", after Kelly (1956), and is much celebrated (Poundstone, 2005). First among its many interesting properties (MacLean & Ziemba, 1999; MacLean, Ziemba, & Blazenko, 1992), Kelly (1956) found that maximizing $\bar{U} = E[\ln(wealth_t)]$ implicitly maximizes the investor's expected exponential capital growth over repeated trials $t = 1, 2, \ldots, T$, as $T \to \infty$.

Following Kelly (1956, pp. 923–925), the gambler's decision rule is to bet amount $\rho_H W$ on *Home* and amount $\rho_A W$ on *Away* (with $\rho_H + \rho_A \leq 1$) so as to maximize his expected log of fwealth

$$p_i \ln[W\{\rho_H \alpha_H + (1 - \rho_H - \rho_A)\}] + (1 - p_i) \ln[W\{\rho_A \alpha_A + (1 - \rho_H - \rho_A)\}], \quad (1)$$

where p_i is his subjective probability of *Home*, and $\alpha_H > 1$ and $\alpha_A > 1$ are the bookmaker's advertised gross payouts per \$1 bet on *Home* and *Away* (respectively). These payouts include the return of the \$1 bet. In the traditions of British bookmaking, the winning payouts are $\alpha_H = (1 + \frac{1}{\Omega_H})$ and $\alpha_A = (1 + \frac{1}{\Omega_A})$, where Ω_H and Ω_A are called the "odds on", or odds in favor of *Home* and *Away*, respectively. Ideally, Ω_H and Ω_A will be reciprocals and there will be no "vig" or over-round. More typically, the bookmaker's odds-implied "probabilities" of *Home* and *Away*, $1/\alpha_H$ and $1/\alpha_A$ respectively, sum to $(1 + \nu) > 1$, and the bookmaker's over-round is ν (for our bookmaker, $\nu \approx 8\%$).

It is easily shown that in order to maximize Eq. (1), the gambler, who puts probability p_i on *Home* and $(1 - p_i)$ on *Away*, should bet such that

$$\begin{cases} \rho_H = \dfrac{1 - \alpha_H p_i}{1 - \alpha_H} \quad \text{and} \quad \rho_A = 0 \\ \qquad \text{when } p_i > 1/\alpha_H \\ \rho_A = \dfrac{1 - \alpha_A(1 - p_i)}{1 - \alpha_A} \quad \text{and} \quad \rho_H = 0 \\ \qquad \text{when } (1 - p_i) > 1/\alpha_A \\ \rho_H = 0 \quad \text{and} \quad \rho_A = 0 \\ \qquad \text{when } 1 - \dfrac{1}{\alpha_A} < p_i < \dfrac{1}{\alpha_H}. \end{cases}$$

The gambler's log-optimal or Kelly bet is therefore either a fixed fraction of wealth W (independent of the amount of W) placed on just one of the two possible game outcomes, or, in the unfortunate case where his subjective probability p_i lies within the bookmaker's probability spread $\{1 - \frac{1}{\alpha_A}, \frac{1}{\alpha_H}\}$, no bet at all.

After betting according to this rule, the gambler's terminal wealth equals his initial wealth W multiplied by a return factor F equal to

$$\begin{cases} \left(\dfrac{1 - \alpha_H p_i}{1 - \alpha_H}\right)\alpha_H + \left\{1 - \left(\dfrac{1 - \alpha_H p_i}{1 - \alpha_H}\right)\right\} = p\alpha_H \\ \qquad \text{if } p_i > \dfrac{1}{\alpha_H} \text{ and } \textit{Home} \\ 1 - \left(\dfrac{1 - \alpha_A(1 - p_i)}{1 - \alpha_A}\right) = \dfrac{\alpha_A p_i}{\alpha_A - 1} \\ \qquad \text{if } (1 - p_i) > \dfrac{1}{\alpha_A} \text{ and } \textit{Home} \\ \left(\dfrac{1 - \alpha_A(1 - p_i)}{1 - \alpha_A}\right)\alpha_A \\ \qquad + \left\{1 - \left(\dfrac{1 - \alpha_A(1 - p_i)}{1 - \alpha_A}\right)\right\} = (1 - p_i)\alpha_A \\ \qquad \text{if } (1 - p_i) > \dfrac{1}{\alpha_A} \text{ and } \textit{Away} \\ 1 - \left(\dfrac{1 - \alpha_H p_i}{1 - \alpha_H}\right) = \dfrac{\alpha_H(1 - p_i)}{\alpha_H - 1} \\ \qquad \text{if } p_i > \dfrac{1}{\alpha_H} \text{ and } \textit{Away} \\ 1 \quad \text{if } 1 - \dfrac{1}{\alpha_A} < p_i < \dfrac{1}{\alpha_H} \text{ or } \textit{Tie}. \end{cases}$$

In the case of a tied game, all bets are returned, and hence the gambler's initial wealth is restored, implying a Kelly return of $F = 1$. In the same way, the forecaster's return in the event that he makes no bet is $F = 1$ (his wealth is unchanged). After a sequence of T trials, forecaster i's cumulative or overall Kelly return equals $\prod_{t=1}^{T} F_i^t$, where F_i^t represents his Kelly return in trial t. This can be written as $\exp\left\{\sum_{t=1}^{T} \ln(F_i^t)\right\}$, where $\ln(F_i^t)$ represents forecaster i's market-based "Kelly score" in trial t. Note that $s_i^t = \ln(F_i^t)$ is a strictly proper probability scoring rule, since forecaster i maximizes $E[\ln(F_i^t)]$ in trial t if and only if he reports his honest probability p_i^t. The advantage of the Kelly score $\ln(F_i^t)$ is that it captures the same economic forecast quality (or qualities) as the conventional log score, but after making allowance for the existence in real-world prediction markets of a bid-ask spread (cf. Johnstone, 2007a,b; Roulston & Smith, 2002).

2.3. How the bets are made

The betting mechanism is operational in the sense that it uses only information which is available at the time when the bets are made. We imagine betting in each competition round r based on the forecasts of those participants whose forecasts have performed best during the season over rounds 1 through $(r-1)$. Rather than commence betting as soon as the season begins, we allow an observation period at the start of each new season to allow information to accumulate relating to which seem to be the "best" (highest scoring) probability forecasters. Betting starts in round 9, and is based on the forecasts of the forecasters with the highest cumulative scores over an initial 8 week observation period. Test periods of 4, 6, 8, and 10 weeks were trialled, and 8 weeks proved to be the best compromise between waiting for more information on the forecasters' respective performances and getting started "making money".

Before each competition round r, we calculate the cumulative score (the log score, categorical score or Kelly score) of each forecaster i over all competition rounds to date. The cumulative score after the completion of round $(r-1)$ is calculated as

$$S_i^{r-1} = \sum_{r=1}^{r-1} \sum_{g=1}^{8} s_i^{r,g},$$

where $s_i^{r,g}$ represents the score achieved by forecaster i in game $g = \{1, 2, \ldots, 8\}$ of round r. Bets are then placed in round r based on the probability forecasts of the participants with the highest aggregate scores S_i^{r-1} to date. The idea is to place bets in each round by following the opinions of the current competition leaders, rather than sticking with forecasters who performed well early in the competition but were overtaken in more recent rounds (cf. Chen et al., 2005, p. 61).

At the time of each competition round r, we know the individual forecasters with the highest cumulated log, Brier, categorical and Kelly scores, S_i^{r-1}. These are generally different individuals, since each different score function captures or emphasizes a different forecast attribute or blend of attributes (Winkler & Murphy, 1968, p. 756). Consistent with a large body of existing evidence, we find that betting is more profitable when it is based on combined forecasts rather than on a single opinion.

2.3.1. Combining personal probabilities

Different forecasters are bound to have asymmetric expertise and differing sources of information, especially "inside" information, and perhaps even interpret their most common experiences and signals contrarily. Provided that their individual forecasts are at least partly independent, neither is statistically redundant, implying that both can contribute to the accuracy of a combined forecast (cf. Clemen et al., 1995, pp. 135–137). The existing theory offers no single "off the shelf" method of combining probabilities (Winkler, 1986, p. 301). Normative methods of combination, particularly Bayesian methods, are highly evolved, but are undermined either by disagreement over the axioms or philosophical criteria that combined probabilities should satisfy, or by the difficulty of modeling individual expert forecasts in terms of their likelihood functions, statistical dependencies and other attributes such as individual psychological biases. There is also the practical issue of estimating all relevant input parameters (e.g. the likelihood of a particular forecaster reporting some given probability conditional on each of the possible states of nature) when these are innately subjective, and often unstable (cf. Jose & Winkler, 2008, p. 163).

The problem of combining subjective probabilities has been studied extensively. Clemen (1989) and Winkler (1986) provide detailed surveys of the relevant literature. At a practical level, there is a lot of experimental evidence supporting simple heuristic methods. It is widely concluded that the simple linear opinion pool or unweighted arithmetic average is not only the easiest and most intuitive method of combining expert probabilities, but also perhaps the most effective, as measured by its accuracy relative to that of the constituent individual experts and that of more complicated combinations. References include Chen et al. (2005), Clemen (1989), Clemen and Winkler (1986, 1999), Servan-Schreiber et al. (2004), Timmermann (2006), Wallsten et al. (1997), Winkler (1971) and Winkler, Murphy, and Katz (1977).[4]

Rather than experiment with different combination algorithms in any detail, all probabilities in this simulation are combined by their simple arithmetic averages. Normalized geometric average probabilities are

[4] Jose and Winkler (2008) suggest that the mean can be further improved upon by robustness methods that correct for its sensitivity to outliers.

bolder (closer to 0 or 1) than simple averages, and thus promote bigger bets, particularly when the gambler adopts a less risk averse decision rule (see below), leading to frequent significant losses and lower profits overall. See Chen et al. (2005, p. 62), for closely related discussion.

2.3.2. Deciding how much to bet

A generally little known and under-appreciated body of literature is the mathematical theory of gambling, starting with the original betting paper by Kelly (1956) and now fragmented across many disciplines, particularly management science, operations research, information science, statistics, finance and economics. One of the most important findings in this body of literature, based in theory and with very extensive applications in casinos and other real-money betting environments, is the importance to professional gamblers of "fractional Kelly" betting. A fractional Kelly bet has the gambler wager just some fraction λ ($0 < \lambda < 1$) of his "full-Kelly" bet (see above). A very common rule of thumb is to bet "half-Kelly" ($\lambda = 0.5$). Thorp (1969, 2000), who, with Ziemba (known famously as *Dr Z* in his racetrack persona), is the most well known proponent of λ-Kelly betting, provides much of the theoretical justification for this decision rule in both practical betting and investment contexts.

Within the mathematical gambling literature, in a series of papers that bridge decision theory and financial economics, Li (1993), MacLean et al. (1992), MacLean, Sanegre, Zhao, and Ziemba (2004) and MacLean and Ziemba (1999) provide a very thorough understanding of the risk-return properties of fractional Kelly betting.[5] These studies, applicable to "favorable" games, wherein the gambler has "an edge", in that his probability assessment is based on better information than the posted market (casino or racetrack) betting odds, demonstrate the effective trade-off between expected rates of return and risk (measured in different ways) achieved by invoking a fixed Kelly fraction $\lambda < 1$. In brief, as λ is reduced from full-Kelly ($\lambda = 1$) toward no bet at all ($\lambda = 0$), there accrues, to begin with, a very worthwhile gain in investment security (measured for example by the probability of an "acceptable or better" outcome, however reasonably defined) relative to what is foregone in expected return or expected capital growth. In the case of continuous lognormal returns, λ-Kelly betting is shown to produce a two-dimensional efficient frontier of betting possibilities, analogous to the mean-variance efficient frontier in portfolio theory, and to proxy for practically any risk-averse individual's utility function at some fixed λ (a smaller λ equates to a greater risk aversion).

To explore the available betting strategies as widely as is practicable, and thereby allow for different levels of risk aversion, we back-test market-based betting outcomes under fractional Kelly-betting over the range of possible fixed λ values, $\lambda \in \{0.1, 0.2, \ldots, 0.9, 1\}$. Consistent with the results of previous theoretical and empirical studies, we find that over our finite betting interval (seven seasons), full-Kelly betting fails to produce the highest capital growth of all λ-Kelly betting rules. While full-Kelly or $\lambda = 1$ betting is "asymptotically optimal", or optimal over the "very long run", assuming of course that the gambler truly has a probabilistic edge, its less appealing property is that, over any shorter run, full-Kelly wagers are almost certain to result in an occasional substantial or even catastrophic capital loss, and will not always recover to the extent that the $\lambda = 1$ gambler's wealth after T trials exceeds that of a matching fractional-Kelly gambler with the same edge (even one with very low fixed λ). In the case of our study, a Kelly betting fraction of $\lambda = 0.3$ or 0.4 produces the highest terminal wealth over seven seasons of capital accumulation (see the results below).

A possible breakdown in any practical application of a λ-Kelly betting rule occurs when the required bet is larger than the bookmaker will accept. All bets in our study are capped at $25,000, an amount that the AFL betting market will absorb without any *ex ante* revision of odds. Note that, on the assumption that the gambler commences betting with $100 and Kelly-bets continuously over seven years, a single bet of more than $25,000 is extremely rare, even at $\lambda = 1$.

3. Results

We look first at betting based on individual forecasters' opinions, then at betting based on merged opinions. The gains from pooling forecasts before betting prove to be substantial and consistent enough that only combined probabilities seem worthy of further investigation.

[5] Other references are listed by MacLean et al. (1992, p. 1574).

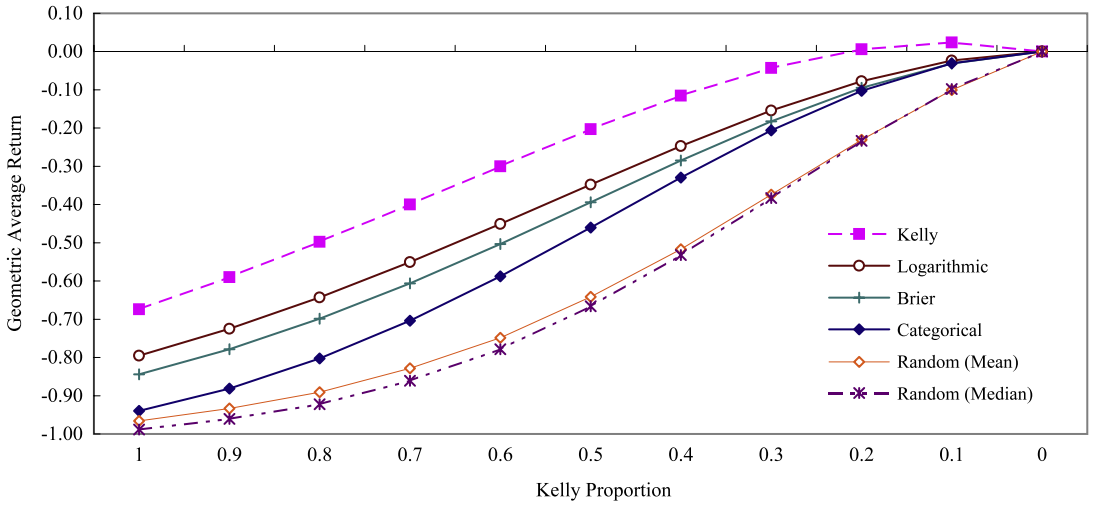

Fig. 1. Geometric average returns (per season) from λ-Kelly betting on individual forecasters' probabilities.

3.1. Individual forecasts

The first set of results, shown in Fig. 1, contains the geometric average returns (for each season) from fractional-Kelly betting based on the probabilities reported by each of the first ranked forecasters under the (i) Kelly score (ii) log score, (iii) Brier score and (iv) categorical score. For convenience, we label these individuals as Kelly#1, Log#1, Brier#1 and Cat#1 respectively. The identities of Kelly#1, Log#1, Brier#1 and Cat#1 are observable at any time during each season based on the historical results to date.

It is immediately evident from Fig. 1 that the geometric average returns (for each season) over the seven seasons of betting are uniformly very poor. Apart from the case of 10% or 20% Kelly betting following the forecasts of Kelly#1 (who may be a different forecaster each week), the average returns are all negative. Full-Kelly ($\lambda = 1$) betting produces the most catastrophic results. The best full-Kelly outcome over the seven seasons is a geometric average loss of 67% per season, achieved by betting each week on the forecasts of Kelly#1.

There is not the slightest hint in these results that a profitable betting strategy can be found. Where returns are positive they are negligible, indicating that the individual forecasting proficiency on show is not sufficient to either beat the bookmaker or overcome the bookmaker's advantage built into the bid-ask spread.

Consistent with previous theoretical arguments suggesting that categorical forecast evaluation is inherently uninformative, the least profitable selection rule is the categorical score. Interestingly, however, this conclusion is qualified once multiple forecasters are chosen and their probability forecasts pooled (see below).

It is important to note that although no profitable individual forecaster selection rule is demonstrated, there is strong evidence that scoring rules are a better way of selecting an individual forecaster than chance. The worst results in Fig. 1 come from choosing forecasters at random rather than using a measure of past performance. Random selection according to a bootstrap procedure (explained below) produces a distribution of geometric average λ-Kelly betting returns for each possible fixed λ, and bootstrap significance levels (not reported here) have been calculated for each strategy for each value of λ. Only the Kelly#1 forecaster consistently outperforms a chance selection at conventional significance levels. The bootstrap p-values for Kelly#1 are between 0.004 and 0.071, and are nearly all very small. Log#1 achieves similar results, particularly for higher values of λ. With a maximum p-value of 0.208, even the categorical rule (the least successful selection method) appears to offer an economically "systematic" rather than an effectively random method of selecting an individual probability forecaster.

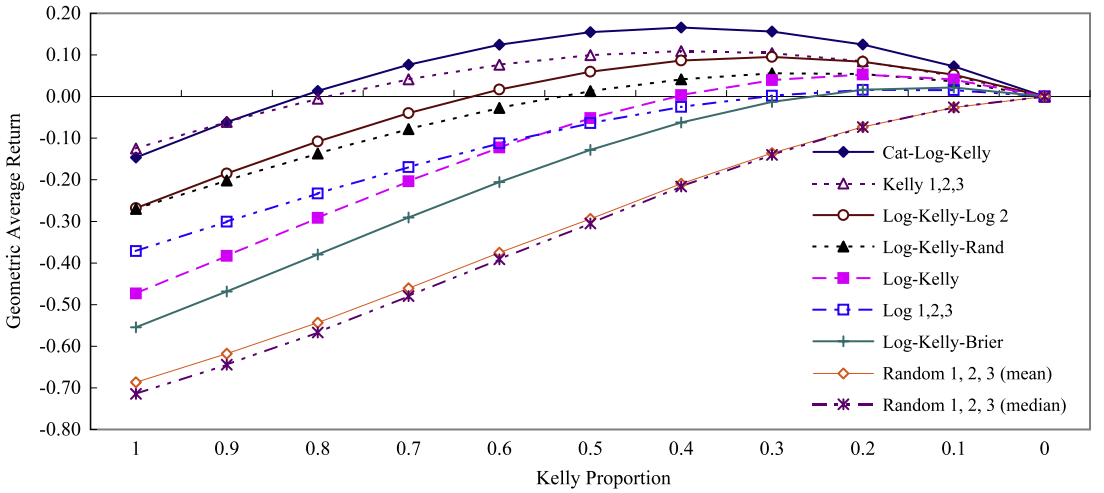

Fig. 2. Geometric average returns (per season) from λ-Kelly betting on pooled forecasters' probabilities.

3.2. *Pooled forecasts*

The results of hypothetical betting over the seven seasons using different three-forecaster opinion pools and Kelly fractions are shown in Fig. 2. The vertical axis shows the geometric average return per season over the full seven seasons. Our first observation is that betting based on the pooled (simple average) probability of a panel of "experts" identified using probability scoring rules is profitable, even after taking into account the probability spread imposed by the betting agency. The highest profits are obtained using 40%-Kelly betting based on the three-forecaster probability opinion pool comprising the first ranked forecasters under the (i) log score, (ii) Kelly score and (iii) categorical score, namely Log#1, Kelly#1 and Cat#1, respectively.[6] Using actual market betting odds, and placing bets according to a fractional Kelly rule, the geometric average return is at best 16.6% per season over the seven seasons (meaning that \$1 becomes $\$1(1.166)^7 = \$2.93$ after seven seasons, not counting any risk-free return earned between seasons or between weekends in each football season). Expressed another way, the geometric average return is 16.6% per 12–13 weeks, since each season covers 22 weeks and betting starts in week 9.

This decidedly attractive-looking return[7] is indicative of latent expertise among the most proficient forecasters in our data set—expertise which is realized only when individual opinions are pooled. The results from betting on individual forecasts are certainly not of the same magnitude.

To compare the best performing opinion pool with objective benchmarks, betting outcomes are calculated for seven control groups. In order of their λ-Kelly betting performances, these are: (i) {Kelly#1, Kelly#2, Kelly #3}; (ii) {Log#1, Kelly#1, Log#2}; (iii) {Log#1, Kelly#1, Rand#1} (Rand#1 is a randomly chosen forecaster); (iv) a two-forecaster pool, {Log#1, Kelly#1}; (v) {Log#1, Kelly#1, Brier#1}; (vi) {Log#1, Log#2, Log#3}; and finally (vii) {Rand#1, Rand#2 and Rand#3} (three randomly selected forecasters).

[6] In some weeks, Log#1, Kelly#1 and Cat#1 could be the same person, or one person might be the ranked first according to two of the scores. Pooling all seven years of data, the average number of different forecasters in this three-forecaster pool varies from 2.71 (week 1) to 2.29 (week 14). Note that as the pre-finals season draws to a close (week 14), there is marginally more agreement between the three rules as to which person is the best forecaster (though there is clearly no strong consensus reached).

[7] From a portfolio theory-CAPM perspective, betting on AFL games has a beta (or covariance with the stock of all other assets, e.g. stocks) equal to zero, and hence only needs to return the risk-free rate to warrant its inclusion in a rational portfolio of risky assets. In principle, therefore, there is scope for expert sports analysts to run a portfolio of sports bets and to sell shares in this portfolio to hedge funds and other funds management institutions. Moreover, if sports betting markets and prediction markets generally exhibit mispricing or demonstrable market inefficiency, then this would seem an inevitable development within the financial services sector.

Interestingly, as can be seen from Fig. 2, there is an obvious order of Kelly betting dominance between the different opinion pools considered. More specifically, for all Kelly fractions λ ($0 < \lambda < 1$), betting profits are increased by adding a third forecaster to the two-forecaster pool {Log#1, Kelly#1}, and the best choice for the third forecaster is Cat#1, followed by Log#2, Rand#1 and Brier#1.

It is not surprising that a randomly chosen forecaster adds to the two-forecaster pool's performance. The curve in Fig. 2 relating to the opinion pool {Log#1, Kelly#1, Rand#1} plots its mean geometric annual return, calculated by drawing Rand#1 each week (for seven years) using a chance mechanism (see below for details), and then repeating this process 10,000 times to produce a bootstrap distribution of seven-year geometric returns. The contribution made by Rand#1 is consistent with the many previous empirical studies showing evidence of the average opinions of some small number of forecasters outperforming either an individual or the average of a smaller group.

The more surprising result, and one for which there appears to be no precedent in the literature, is that the categorical evaluation of probability forecasts assists in finding a "good" probability forecaster, and identifies a more valuable third forecaster than any of the other scores considered.

One possible explanation of the incremental profits contributed by Cat#1 is that a categorical expert's forecasts are much more valuable "in portfolio" than of themselves (cf. Batchelor & Dua, 1995). Imagine that in general Cat#1 states either very high or very low probabilities, and is very often categorically "right" (i.e., on the "right" side of 0.5). Such a forecaster is likely to perform badly according to the log (or Kelly) score. In those trials where he is categorically wrong, he will attract a severe penalty for his apparent overconfidence. At worst he will score $\ln(0) = -\infty$, from which he cannot recover (this is a well known "defect" of the log score; see Roulston & Smith, 2002, p. 1654; Winkler, 1967, p. 1119, 1971, p. 679).[8] However, when combined with other forecasters in a pool, the effect of any such systematic overconfidence will be moderated, and in the frequent trials where Cat#1 is categorically right, his bold or even reckless probability statement will have a very beneficial effect on the pooled (simple average) probability, and may well spawn a highly profitable bet.

Another unexpected result, evident in Fig. 2, is that the addition of Brier#1 to the two-forecaster opinion pool {Log#1, Kelly#1} is highly counterproductive. One possible explanation for this is that the Brier and log scores tend to reward the same profile of probabilistic forecasting expertise, and hence do not add much to each other when it comes to diversifying the opinion pool.[9] To improve the combined (average) probability forecast, an additional forecaster is better to contribute something idiosyncratic rather than generally more of the same. More of the same opinion often means dragging the combined probability toward the market consensus, and hence reducing both the size of the bet and its potential return.

The final inference from Fig. 2, re-confirming the results shown in Fig. 1, is that scoring rules clearly outperform any selection of forecasters by chance. The worst-performing opinion pool in Fig. 2 is the combination of three randomly chosen forecasters. This result would suggest that there is little to be gained in a market context, where prices are already representative of a broad average of opinions, from simply averaging a group of forecasters' opinions without first selecting "good" forecasters. Their average opinion might outperform that of any of the constituent individuals, but our results suggest that the gain from averaging is not enough for the combination to outperform the bookmaker. Note, however, that the combination of three randomly chosen forecasters does better than a single randomly selected forecaster. This can be seen by comparing the mean and median results for {Rand#1, Rand#2, Rand#3} in Fig. 2 with the same results for Rand#1 in Fig. 1. The three-random-forecaster pool accrues much smaller losses.

One aspect of our results that remains unexplained is why the input of one randomly selected forecaster increases the profitability of the two-forecaster pool {Log#1, Kelly#1}, while a pool of three randomly

[8] Perhaps "defect" is the wrong word, since this mathematical property mirrors the complete demise of a Kelly gambler who acts on a personal probability of 0 or 1, and loses.

[9] To further ensure that Brier#1 has nothing to offer, Kelly betting returns were also found for the four-forecaster pool {Log#1, Kelly#1, Cat#1, Brier#1}. The results (which are available from the authors) were clearly worse across all λ levels than for the three-forecaster pool with Brier#1 excluded.

Table 1
Geometric average returns (per season) from λ-Kelly betting on pooled forecasters' probabilities.

Forecast pool	Kelly proportion λ									
	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Log-Kelly-Cat	−0.146	−0.062	0.014	0.077	0.124	0.155	0.166	0.156	0.125	0.073
p-level	0.008	0.008	0.008	0.008	0.008	0.010	0.011	0.012	0.013	0.014
p-level#	0.074	0.065	0.066	0.066	0.067	0.067	0.065	0.068	0.069	0.070
Log-Kelly-Log#2	−0.268	−0.185	−0.108	−0.040	0.017	0.060	0.087	0.095	0.084	0.052
p-level	0.019	0.020	0.020	0.021	0.027	0.024	0.031	0.032	0.037	0.039
p-level#	0.211	0.215	0.209	0.212	0.218	0.216	0.230	0.227	0.237	0.240
Log-Kelly-Rand	−0.027	−0.202	−0.137	−0.078	−0.027	0.013	0.041	0.056	0.054	0.036
p-level	0.020	0.023	0.025	0.030	0.031	0.043	0.052	0.060	0.072	0.080
p-level#	0.214	0.243	0.261	0.293	0.315	0.344	0.388	0.400	0.438	0.472
Log-Kelly	−0.473	−0.383	−0.292	−0.204	−0.122	−0.052	0.003	0.040	0.053	0.041
p-level	0.100	0.029	0.088	0.085	0.085	0.080	0.081	0.078	0.075	0.066
p-level#	0.679	0.667	0.630	0.627	0.585	0.566	0.538	0.486	0.448	0.398
Log-Kelly-Brier	−0.555	−0.469	−0.380	−0.291	−0.206	−0.128	−0.062	−0.013	0.016	0.022
p-level	0.176	0.174	0.166	0.161	0.165	0.155	0.160	0.152	0.150	0.141
p-level#	0.849	0.841	0.818	0.817	0.796	0.782	0.771	0.739	0.714	0.690
Kelly#1,#2,#3	−0.125	−0.061	−0.005	0.041	0.077	0.100	0.109	0.104	0.085	0.050
p-level	0.006	0.008	0.009	0.022	0.014	0.018	0.022	0.028	0.036	0.043
Log#1,#2,#3	−0.371	−0.301	−0.233	−0.170	−0.113	−0.064	−0.025	0.002	0.016	0.016
p-level	0.044	0.049	0.055	0.064	0.087	0.088	0.109	0.125	0.151	0.176
Rand#1,#2,#3 (mean)	−0.687	−0.618	−0.543	−0.461	−0.375	−0.294	−0.209	−0.136	−0.073	−0.026
p-level	0.427	0.437	0.444	0.456	0.464	0.472	0.479	0.486	0.498	0.504
Rand#1,#2,#3 (median)	−0.714	−0.644	−0.567	−0.479	−0.391	−0.305	−0.216	−0.140	−0.073	−0.026
p-level	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500

chosen forecasters makes nothing but losses. It seems odd that one random forecaster would add value to an existing two-forecaster pool, while a pool of three random forecasters loses money relative to making no bets at all.

3.3. Bootstrap significance tests

The betting profits from any given opinion pool can occur by chance. To test for this possibility, we run randomization tests for each of the opinion pools considered at each fixed value of λ. The associated *p*-values are shown in Table 1. These represent the probability (observed frequency over 10,000 repetitions) of a seven-year cumulative betting profit as large as or larger than that observed when the three forecasters in the pool {Rand#1, Rand#2, Rand#3} are drawn at random rather than using a probability scoring rule. Note that the bootstrap significance tests applied throughout this study match the "Monte Carlo reality check *p*-value" methods proposed by Sullivan, Timmermann, and White (1999) and White (2000) very closely.

It appears from the bootstrap *p*-values that the profits achieved by Kelly betting based on the {Log#1, Kelly#1, Cat#1} opinion pool are highly significant for all fixed values of λ. The *p*-values observed are typically about 2%. The inference, therefore, is that the paper betting profits achieved by λ-Kelly-betting are too high to be the result of chance, as would be the explanation if the scoring rules in question really did no more than rank forecasters in a random order (rather than based on inherent forecasting ability or "accuracy").

The mechanism for selecting forecasters randomly is to keep a running score on each forecaster, in the same way as is done with the probability scores, but with each forecaster being awarded a random score each week rather than a score based on their actual performances. This random quantity is drawn from the empirical distribution of all forecasters' log scores over all weeks. Several other methods of randomization were trialled, but the differences in the results were negligible.

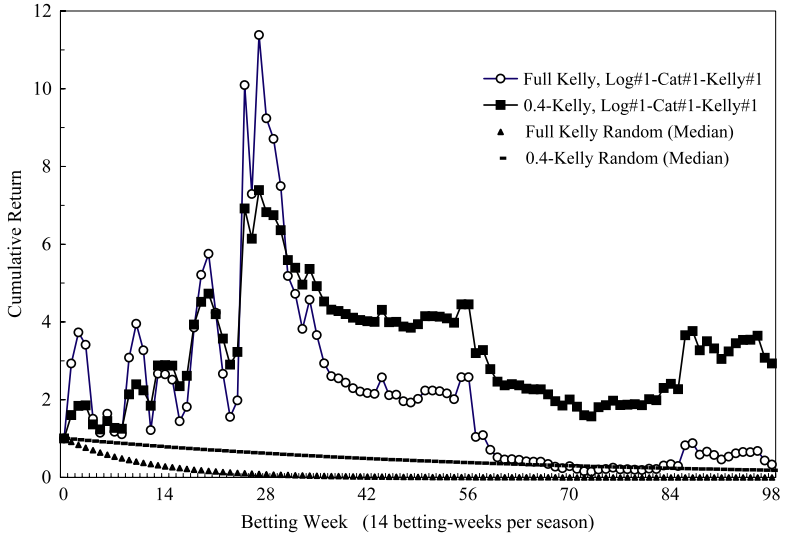

Fig. 3. Cumulative return over the betting period from λ-Kelly betting on pooled forecasters' probabilities, shown on a weekly basis (98 betting-weeks over 7 seasons).

The remaining bootstrap test is designed to test whether the marginal contribution of a third forecaster Ψ to the pool {Log#1, Kelly#1, Ψ} is significant, or can be explained by chance. A bootstrap distribution of profits is calculated for each fixed λ by betting on a forecaster pool containing Log#1 and Kelly#1, together with a randomly selected forecaster, Rand#1. Again, this distribution is generated over 10,000 repeats. The resulting p-value (labeled p-value$^{\#}$) represents the probability (at each value of λ) of the opinion pool {Log#1, Kelly#1, Ψ} producing such a high geometric average betting return (as that which is actually observed), on the condition that the third forecaster Ψ is really only a randomly selected individual (as would be the case if the scoring rule by which Ψ is selected generated only noise rather than a systematic measure of forecasting ability or economic potential).

The bootstrap p-value$^{\#}$ for Ψ = Cat#1 is small (about 7%) across all Kelly fractions λ. It would seem, therefore, that the marginal contribution of the third forecaster Cat#1 to the pool {Log#1, Kelly#1} is unlikely to be the effect of chance alone. By comparison, the same p-value$^{\#}$ for Brier#1 ranges from 0.690 to 0.849, indicating that this scoring rule offers no systematic discrimination of the incremental forecasting value.

3.4. Season-by-season returns volatility

The results presented here are from seven seasons of betting. A long evaluation period is essential for assessing the growth properties of betting strategies, and particularly Kelly-based strategies, which are justified theoretically by their long-run return characteristics. Characteristically of Kelly-betting, there is a large natural volatility in returns over shorter trial sequences. Our data split naturally into the seven seasons, since every football season is different. It is common at an anecdotal level for football tipsters to describe a given season as particularly difficult to predict. A possible explanation for this is that chance events, such as weather and injuries, and the effects of the protocols within the AFL which are designed to make teams "even" in ability and opportunity, combine to make some seasons more "even" (less predictable) than others.

To observe the returns volatility in our sample period, Fig. 3 shows the wealth of a gambler (starting with $1) who follows the recommended strategy of betting on the {Log#1, Kelly#1, Cat#1} opinion pool. Two possible wealth paths are graphed, with the better being that based on "40% Kelly" betting, and the other being based on full-Kelly betting. The remaining two lines on the graph show the median of the cumulative value of a betting fund based on a pool of

three randomly chosen forecasters. Again, the Kelly fractions employed are 40% and 100%.

Clearly, there is only one profitable betting strategy, namely 40% Kelly betting following the {Log#1, Kelly#1, Cat#1} opinion pool. This is apparently the most bankable opinion pool, and yet a gambler who bets on it using full-Kelly is left with little of his endowment after seven seasons of betting. By comparison, and worse still, the median gambler who full-Kelly bets on a pool of random forecasters is nearly bankrupt after just 28 weeks (by the end of the second season).

The key aspect of Fig. 3 is the heightened volatility (risk) of full-Kelly relative to fractional ($\lambda = 0.4$) Kelly betting. While the preferred forecaster pool appears to offer the gambler an edge over the bookmaker, any attempt to milk this advantage too aggressively brings about occasional severe losses, from which the gambler does not recover during the betting period. The more tempered strategy of 40% Kelly betting avoids the worst of these dips, and after seven seasons leaves the gambler with a wealth of $2.93 for every dollar of his initial endowment (a geometric average return of 16.6% per season). While this is a good result, there are lengthy periods of accumulated losses within the seven year betting campaign, and hence some strain on any gambler or institution who commits to such a strategy.

4. Conclusion

The results provided in this paper suggest that a categorical scoring of probability forecasts, although not a "proper" method of scoring, can help to identify useful forecasters. An individual forecaster who does well according to this measure is, by definition, one who is good at getting the estimated probability in the "right" direction from 0.5. This is obviously a very great skill in itself, but it is not necessarily enough to make for profitable betting. Bets are made as a function of the gambler's precise probability assessment relative to the market's assessment, and someone who judges probabilities as being close to 0.5, even if they are generally on the right side of 0.5, will not make many bets, or at least many large bets, and hence are unlikely to be very profitable. It is therefore of great interest that in our experiment the "categorically best" forecaster appears to contribute significantly to the betting profits of a three-forecaster probability pool.

This result appears to be very robust, though of course it carries no guarantee that the categorically superior forecaster will remain useful when the forecaster pool is expanded. We have limited our attention in this study to a three-forecaster pool for several reasons, mainly intuition and convenience. Further trials, not reported here, indicate clearly that when we increase the pool to six forecasters, both the frequency and the size of bets are reduced substantially, and betting profits vanish altogether. Preliminary trials suggest that as more forecasters are added to the pool, the combined probability tends to move toward the "market probability", or toward the probability interval constituting the bookmaker's bid-ask spread, thus reducing or precluding any rational bets.[10] Profitable betting would seem to require a large enough forecaster pool to realize the benefits of combining opinions, but not large enough to swamp the idiosyncratic insights of the individually expert forecasters within the opinion pool.

References

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, *41*, 68–75.

Chen, Y., Chu, C. H., Mullen, T., & Pennock, D. M. (2005). Information markets vs. opinion pools: An empirical comparison. In *Electronic commerce: Proceedings of the 6th ACM conference on electronic commerce* (pp. 58–67). New York: ACM Press.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.

Clemen, R. T., Murphy, A. H., & Winkler, R. L. (1995). Screening probability forecasts: Contrasts between choosing and combining. *International Journal of Forecasting*, *11*, 133–146.

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business and Economic Statistics*, *4*, 39–46.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*, 187–203.

[10] It is hard to say what the ideal number of assessments to include in the opinion pool is (cf. Graham, 1996, pp. 215, 230; Winkler, 1971, p. 683). A more recent study by Winkler and Clemen (2004, p.176) suggests that a small number of assessors, such as three or four, may be most successful.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *The Economic Journal*, *114*, 844–866.

de Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica*, *34*, 129–145.

de Finetti, B. (1982). The proper approach to probability. In G. Koch, & F. Spizzichino (Eds.), *Exchangeability in probability and statistics: Proceedings of the international conference on exchangeability in probability and statistics, 1981* (pp. 1–6). Amsterdam: North-Holland.

Graham, J. (1996). Is a group of economists better than one? Than none? *Journal of Business*, *69*, 193–232.

Granger, C. W. J., & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*(7), 537–560.

Johnstone, D. J. (2007a). Economic Darwinism: Who has the best probabilities? *Theory and Decision*, *62*, 47–96.

Johnstone, D. J. (2007b). The parimutuel Kelly probability scoring rule. *Decision Analysis*, *4*, 66–75.

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*, 163–169.

Kadane, J. B., & Winkler, R. L. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association*, *83*, 357–363.

Kelly, J. L. (1956). A new interpretation of the information rate. *Bell System Technical Journal*, *35*, 917–926.

Levitt, S. D. (2004). Why are gambling markets organized so differently from financial markets? *The Economic Journal*, *114*, 223–246.

Li, Y. (1993). Growth-security investment strategy for long and short runs. *Management Science*, *39*, 915–924.

Lichtendahl, K. C., & Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, *53*, 1745–1755.

Lindley, D. V. (1982). Scoring rules and the inevitability of probability. *International Statistical Review*, *50*, 1–26.

Lopez, J. A. (2001). Evaluating the predictive accuracy of models. *Journal of Forecasting*, *20*, 87–109.

Luenberger, D. G. (1998). *Investment science*. Oxford: Oxford University Press.

MacLean, L. C., Sanegre, R., Zhao, Y., & Ziemba, W. T. (2004). Growth with security. *Journal of Economic Dynamics and Control*, *28*, 937–954.

MacLean, L. C., & Ziemba, W. T. (1999). Growth versus security tradeoffs in dynamic investment analysis. *Annals of Operations Research*, *85*, 193–225.

MacLean, L. C., Ziemba, W. T., & Blazenko, G. (1992). Growth versus security in dynamic investment analysis. *Management Science*, *38*, 1562–1585.

Poundstone, W. (2005). *Fortune's formula: The untold story of the scientific betting system that beat the casinos and Wall Street*. New York: Farrar, Straus and Giroux.

Roulston, M. S., & Smith, L. A. (2002). Evaluating probability forecasts using information theory. *Monthly Weather Review*, *130*, 1653–1660.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Servan-Schreiber, E., Wolfers, J., Pennock, D. M., & Galebach, B. (2004). Prediction markets: Does money matter? *Electronic Markets*, *14*, 243–251.

Stekler, H. O. (1994). Are economic forecasts valuable? *Journal of Forecasting*, *13*, 405–505.

Sullivan, R., Timmermann, A., & White, H. (1999). Data snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, *54*, 1647–1692.

Thompson, J. C., & Brier, G. W. (1955). The economic utility of weather forecasts. *Monthly Weather Review*, *83*, 249–254.

Thorp, E. (1969). Optimal gambling systems for favorable games. *International Statistical Review*, *37*, 273–293.

Thorp, E. (2000). The Kelly criterion in blackjack, sports betting and the stock market. In O. Vancura, J. Cornelius, & W. R. Eadington (Eds.), *Finding the edge: Mathematical analysis of casino games* (pp. 163–213). Reno, NV: Institute for the Study of Gambling and Commercial Gaming.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Amsterdam: North-Holland.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*, 243–268.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*, 1097–1126.

Winkler, R. L. (1967). The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, *62*, 1105–1120.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, *66*, 675–685.

Winkler, R. L. (1986). Expert resolution. *Management Science*, *32*, 298–303.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities (with discussion). *Test*, *5*, 1–60.

Winkler, R. L., & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, *1*, 167–176.

Winkler, R. L., & Murphy, A. H. (1968). Good probability assessors. *Journal of Applied Meteorology*, *7*, 751–758.

Winkler, R. L., & Murphy, A. H. (1979). The value of weather forecasts in the cost-loss ratio situation: An ex ante approach. In *Preprints of the sixth conference on probability and statistics in atmospheric sciences* (pp. 134–138). Boston: American Meteorological Society.

Winkler, R. L., Murphy, A. H., & Katz, R. W. (1977). The consensus of subjective probability forecasts: Are two, three,... heads better than one? In *Preprints of the fifth conference on probability and statistics in atmospheric sciences* (pp. 57–62). Boston: American Meteorological Society.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*, 107–126.