THESIS IN FINANCIAL ECONOMETRICS

Forecasting FIFA World Cup Outcomes via Regularized Mixtures of Predictive Densities

Sergey Marchenko (507560)

July 4, 2021

Supervisor: A. Tetereva

Abstract

In this paper regularized mixtures of predictive densities are used to improve on the density forecasts, when predicting the outcomes of the FIFA World Cup football matches. Opinion pools are used to accumulate all available information, and assign optimal weights to each forecaster. Forecasters in our opinion pool include bookmakers, betting exchanges and FIFA World Ranking. Although, statistically we fail to significantly improve on the individual forecasters, economically we obtain promising results, where regularizations give between 10% and 16% expected profits from betting. Thus, regularizations outperform simple averaging and majority of individual predictors during the FIFA World Cups of 2014 and 2018.

Contents

1 Introduction

Predicting the outcomes of sports events has been of interest for a long time now, particularly, because of the enormous financial turnovers that are transacted on daily basis, Wunderlich and Memmert (2016). Similar to stock markets there is a significant element of unpredictability within the outcome of each event. Particularly in football, experts, such as bookmakers, have tried to perfect the estimation process of the game outcomes as it directly impacts their profits. Sports magazines also seek optimal forecasts to attract football fans who are willing to participate in the open betting markets. Therefore, over time, many alternative forecasting methods have been discussed in the literature.

Stekler et al. (2010) outlines three main forecasting methods; betting market forecasts, model forecasts and expert forecasts. The first method is based on the market predictions of the outcomes, via the odds, which are computed based on participants' predictions. The second method uses the (factor) models, which try and capture all possible factors affecting the game outcome. The third method, is based on the experts' opinion; experts usually being the bookmakers, head coaches, commentators, professional players and so on. Particularly bookmakers would also make use of the statistical models to perfect their forecast accuracy. In addition to those, Frick and Wicker (2016) consider economic forecasts, which they call 'naive' forecasters, who base their predictions not on football expertise but on the team's wage bills, average age and occupancy of the stadium. Last but not least, Stekler et al. (2010) gives much credit to the FIFA World Ranking, which empirically performs as well if not better than the experts' forecasts. The list does not end here, however the above forecasts received most credit in the literature.

Our research aims to make use of different forecasting methods, to come up with an optimal density forecast for the outcome of each game. For this we will consider the work of Diebold et al. (2021), who constructs regularized mixtures of density forecasts, thus extending the work of Diebold and Shin (2019). The key idea of Diebold and Shin (2019) is to transform a set of forecasts of y, $f = (f_1, \ldots, f_K)'$, into a "combined" superior forecast $c(f, w)$, where the weight, w_i $\forall i \in \{1, \ldots, K\}$, on each forecast optimally solves a penalized estimation problem. While this setting indeed gives optimal weights to univariate point forecasts, thus outperforming typical averaging for example, in the context of density forecasts further regularized mixtures yield better probabilistic forecasts, as discussed in Diebold et al. (2021). When predicting the outcomes of the football matches, we are interested in forecasting the probability of each possible outcome (win team 1, lose team 1, draw), thus this research makes use of the relevant regularization settings outlined in Diebold et al. (2021). Specifically, our research question reads: "Can regularized mixtures of predictive densities outperform the experts' forecasts for

the outcomes of the FIFA World Cup matches, when different sources of forecasts are combined?"

To further motivate the use of regularized mixtures of density forecasts, one can again refer to work of Wunderlich and Memmert (2016), who proposes combining both, open betting market and FIFA ranking-based forecasts, to predict the outcome. Furthermore, Frick and Wicker (2016) suggests that economic and expert forecasts compliment each other, and Štrumbelj (2014) concludes that it makes a difference as to which bookmaker or betting exchange we choose, when two or more are available. The above insights indicate that alternative forecasters base their predictions on different factors, and use different sources of information. Hence, by optimally combining several forecasters we can make use of the alternative sources of information, and improve on the accuracy of density forecasts.

We are going to apply the methods proposed by Diebold et al. (2021) to predict the outcomes of the individual FIFA World Cup matches. Thus, for each game several forecasters are consider, each of which should provide us with a density forecast. The latter causes difficulties, as the economic forecasts proposed by Frick and Wicker (2016), for example, do not give density forecasts, instead they rank teams based on their probability of winning the championship. Instead, we will refer to the odds from the open betting market, which should include some of the 'naive' forecasts, as not everyone participating in those bets will have the expertise bookmakers will have. Furthermore, we will include the density forecasts based on the FIFA World Ranking, as suggested by Wunderlich and Memmert (2016), which are based solely on the past teams' performance, thus ignoring the current condition (injuries, home advantage, morale, etc.) of each participating team. These should compliment the forecasts derived from the open betting market odds, which, according to Wunderlich and Memmert (2016), are biased towards the most recent events. Finally, experts will also be treated as separate forecasters, as they are known to make use of the information the above forecasters ignore or do not have access to. Štrumbelj (2014) argues for the presence of the insider information, particularly among bookmakers, and Zeileis et al. (2018) talks about advanced models bookmakers use to maximise their profits. Again, since bookmakers' profits are based on the quality of their forecasts, we expect their predictions to be at the cutting edge amongst others. Note that different experts will also be biased towards particular teams because of their access to insider and/or available information; and the information they consider to be relevant, when constructing density forecasts, will also differ amongst bookmakers, according to Štrumbelj (2014). Therefore, in a similar fashion to Zeileis et al. (2018), we will consider multiple bookmakers, where each bookmaker will be treated as a separate forecaster. Therefore, our opinions pool will include multiple bookmakers, FIFA world Ranking and a couple of open betting exchanges. Bookmakers and betting exchanges will

provide the odds, which, together with the FIFA World Ranking, can be converted into density forecasts, as will be shown in section 3.

Finally, several density forecast evaluation methods are proposed by existing literature. In particular, when forecasting outcomes of football matches, Wheatcroft (2019) considers three evaluation methods, ranked probability score (RPS), Brier score and the ignorance score. Each of these scoring rules favours different desired properties of our forecasts. For example the RPS has a so-called "sensitive to distance" property, where it makes use of the ordinal outcomes, and accounts for the fact that a home win is closer to a draw than it is to an away win. For that reason it is commonly used when evaluating density forecasts of the football games. Brier and ignorance scores do not have this property. RPS and Brier scores are non-local[1], while ignorance score is. However, following the result of Wheatcroft (2019), the non-locality and sensitivity to distance as properties of scoring rules can be questioned, as the ignorance score outperforms both, in the context of football matches. Furthermore, sensitivity to distance only applies to 1 of the 32 teams during the World Cup, as all the other teams are playing away, and even then no extra points are rewarded for away wins, as they are, for example, in Champions League. Hence, in the context of our research, using the ignorance score to evaluate the density forecasts should suffice.

Another approach to measuring forecast accuracy, proposed by Wunderlich and Memmert (2020), is the so called *economic* approach, where the profitability of our models is evaluated. Goddard and Asimakopoulos (2004) point towards a possibility of systematically generating positive betting returns in the absence of a superior model accuracy. While Wunderlich and Memmert (2020) also mentions that "*betting returns should not be treated as a valid measure of model accuracy*", Lessmann et al. (2010) argues that "*a model's profitability is the primary indicator of forecasting accuracy*" in the context of horse racing. Therefore, given that the football betting market is much larger than the one of horse racing,[2] motivated by the results of Wunderlich and Memmert (2020), Goddard and Asimakopoulos (2004), Leitch and Tanner (1991), and Lessmann et al. (2010) we will consider the profitability of our models, in addition to their accuracy.

After having introduced the problem and explained the relevant literature, we talk about the data set used for the empirical study. Then all the required methods will be outlined, which includes obtaining probability forecasts, measuring objectives and introducing different regularizations. Next, the set up and the results of a simulation study will be presented, followed

[1] A non-local score takes at least some of the rest of the forecast distribution into account. A local score only considers the probability at the outcome and disregards the rest of the distribution.

[2] https://www.pledgesports.org/2020/05/most-popular-sports-to-bet-on/

by empirical results evaluated at two different objectives. Finally, we will conclude the paper.

2 Data

For the purpose of this research two most recent FIFA World Cups (2014, 2018) are considered, consisting of 64 games each. Hence, we evaluate density forecasts for 128 games. Each forecast will be a regularized mixture of 23 predictive densities, thus we consider 23 different forecasters. This includes 20 expert forecasters, being the bookmakers, 2 betting exchanges, corresponding to the open betting market, and two sets of FIFA World Rankings, being the most recent one prior to the 2014 and 2018 World Cups.

The 23 forecasters are somewhat in line with the ones considered by Zeileis et al. (2018), who took the most well-known and reputable bookmakers. Based on the work of Diebold et al. (2021), who use a total of 19 forecasters, forming a pool with 23 forecasters seems sufficient. The data summarizing the relevant bookmakers' odds is taken from https://www.oddsportal.com/. Wunderlich and Memmert (2016) suggests the use of many betting exchanges is unnecessary due to the openness of the online betting markets. Particularly, using the Betfair[3] odds should sufficiently capture the information available in the open markets. In addition to Betfair I will also consider Matchbook betting exchange, which is more recent. The data summarizing the relevant odds from the open market is taken from https://historicdata.betfair.com/. Finally, the relevant World Rankings are obtained from the FIFA website, https://www.fifa.com/fifa-world-ranking/ranking-table/men/rank/id10887/.

3 Methodology

In this section we explain how the density forecasts are obtained from the betting odds as well as the FIFA World Ranking. Then the ignorance score is outlined, which will be the objective function that is optimized in the context of the penalized estimation problem. Finally, using the work of Diebold et al. (2021), we outline different sets of penalties, which produce regularized mixtures of predictive densities. Remember that each football match consists of 3 outcomes: win team 1, lose team 1 and draw.

[3]Oldest and the most well know betting exchange. Betfair has one of largest trading volumes, providing a good set of market-based odds

3.1 Odds to Probabilities

3.1.1 Basic normalization

Basic normalization is considered by Štrumbelj (2014) and Wunderlich and Memmert (2016), where $\mathbf{o} = (o_1, \ldots, o_n)$ are the odds for each match outcome, for $n \geq 2$. Odds can be (roughly) interpreted as the inverse probability of winning plus the bookmaker's profit margin. Note that $o_i > 1$ for all $i = 1, \ldots, n$. Then we define $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ to be the inverse odds, where $\pi_i = \frac{1}{o_i}$. Note that the inverse odds will sum to more than 1, and hence have to be normalized, removing the bookmaker's profit margin. The following is done by dividing by the sum of inverse odds, thus obtaining $p_i = \frac{\pi_i}{\sum_{i=1}^{i=n} \pi_i}$ to be the set of values adding up to 1, which can be interpreted as outcome probabilities. Note that during normalization we assume that the bookmaker's profit margin is constant across different game outcomes.

3.1.2 FIFA World Ranking

We derive the probabilistic forecasts from the FIFA World Ranking using the model-based approach proposed by Wunderlich and Memmert (2016). First, we transfer the ranking into the expected number of goals scored by each team. Second, we transfer the expected goals into probabilities of each outcome using the bivariate Poisson distribution suggested by Karlis et al. (2005). Below is the detailed summary of the methodology.

1. For each game we expect the higher ranked team to score more goals. We define the ranking points of a higher ranked team as pts_{max}, and the ranking points of a lower ranked team as pts_{min}. Then the expected number of goals scored by each team can be defined as exp_{max} and exp_{min} respectively. The following equation should hold:

$$\frac{exp_{max}}{exp_{max} + exp_{min}} = \frac{pts_{max}}{pts_{max} + pts_{min}}. \tag{1}$$

   Now we need to define the overall expected number of goals scored in each game. Due to the limited information about special offensive or defensive qualities of teams contained in the ranking, we treat all teams equally. Therefore, we assumed the expected goals scored by both teams to be equal in each match and estimated this value by using the average number of goals scored in the previous World Cup ($\hat{g}$). The following equation should hold:

$$exp_{max} + exp_{min} = \hat{g}. \tag{2}$$

   By solving the above system of equations we derive the expected number of goals scored by each team.

2. Given that exp_{max} and exp_{min} is the expected number of goals scored by each team, the probability of the match ending with a result of $X : Y$ is

$$P[X : Y|\lambda_1, \lambda_2, \lambda_3] = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1,^x}{x!} \frac{\lambda_2,^y}{y!} \sum_{i=0}^{min(x,y)} \binom{x}{i} \binom{y}{i} i! \binom{\lambda_3}{\lambda_1\lambda_2}^i, \text{ where} \tag{3}$$

$$\lambda_3 = 0.05(exp_{max} + exp_{min}),$$

$$\lambda_1 = exp_{max} - \lambda_3,$$

$$\lambda_2 = exp_{min} - \lambda_3.$$

The probability of each outcome can then be derived from the above probability density function.

3.2 Objective functions

3.2.1 Log Score

Given the discrete nature of the density forecasts for a scalar variable y, we define $m = 1, \ldots, M$ bins, in which the value of y can be placed. The forecasts are denoted by $\mathbf{p} = (p_1, \ldots, p_M)'$.

As outlined in section 1, the ignorance score will be used as an objective function, primarily based on the results of Wheatcroft (2019), and the fact that sensitivity to distance can be neglected when considering the outcomes of the Word Cup matches. There is no reason to treat away games any differently to home games, as they are not rewarded with a so called 'away' bonus.

The ignorance score is equivalent to the log score, defined by Diebold et al. (2021), as

$$L(p, y) = -\log\left(\sum_{m=1}^{M} p_m \mathbb{1}(y \in b_m)\right), \tag{4}$$

where p_m is the probability assigned to bin b_m, and $\mathbb{1}(y \in b_m) = 1$ if $y \in b_m$ and 0 otherwise. When optimising the objective function the smallest value of L is desired, as it maximises the probability of the true outcome. Note that the above ignorance score is defined for a single forecaster in a single period.

We now modify our notation to identify the specific forecaster, $k = 1, \ldots, K$. Additionally, we want to compute the scores for a set of football matches, each identified in its own period $t = 1, \ldots, T$. These additions simply involve summing over time and inserting "k" subscripts in relevant places. Hence, we obtain

$$L_k(\boldsymbol{p_k}, \mathbf{y}) = \sum_{t=1}^{T} \left(-\log\left(\sum_{m=1}^{M} p_{mkt} \mathbb{1}(y_t \in b_m)\right)\right), \; k = 1, \ldots, K, \tag{5}$$

where $\boldsymbol{p_k} = (p_{k1}, \ldots, p_{kT})$ is the sequence of density forecasts over time for forecaster k, and $\mathbf{y} = (y_1, \ldots, y_T)$ is the sequence of realizations over time.

3.2.2 Expected Betting Returns

Another way to examine the quality of our predictions is to look at the profitability of our models. The following objective is justified from its practical relevance - at the end of the day we are not solely interested in improving the density forecasts, but primarily in maximising our expected profits, based on the density forecasts. The idea is that the game outcomes make it possible to calculate the betting returns that would have been realised if these bets had been placed prior to the events, as summarized by Wunderlich and Memmert (2020). Betting returns are used as proxy for the model profitability.

We will maximise the expected betting returns. The objective function is thus defined as $e_i = o_i\hat{p}_i - 1$ for all $i = 1, \ldots, n$, where e_i is the expected value of a bet, o_i is the betting odd, 1 is the amount that we bet, and $\hat{p}_i$ is the forecasted probability of the outcome we are betting on. Note that e_i, unlike log loss, is not a straightforward function to optimise in the penalised estimation problem setting, due to its linear and non-smooth qualities. Thus for the purpose of this research the optimal weights are computed using the log loss objective, and then evaluated via the betting returns. To maximise the expected betting returns we iterate through every odd given by different bookmakers for each game outcome. Therefore, for every bet the bookmaker with the highest available odd for the game outcome, corresponding to the highest expected value of a bet (e_i), will be chosen. The probability distribution used to compute the expected value of a bet will indeed be computed by our models.

3.3 Penalties

Granger and Ramanathan (1984) have recognized that without the imposition of any restrictions on the mixture weights $\mathbf{w} = (w_1, \ldots, w_K)$ when optimizing point forecasts, the optimal set of weights will be obtained. However, according to Brodie et al. (2009) that is not the case when constructing density forecasts, and essential regularization in the form of simplex constraint is required. Diebold et al. (2021) then provides insights into simultaneously imposing other regularization constraints, which further improve density forecasts.

3.3.1 Simplex

Simplex constraint imposes the non-negativity ($w_i \geq 0 \ \forall i$) and sum-to-one ($\sum_{i=1}^{K} w_i = 1$) of the mixture weights. The non-negativity constraint avoids pathological results, where, for example, mixture density can take negative values. While sum-to-one constraint is required to ensure for the mixture combination to be a valid probability density. These are extensively dis-

cussed by Diebold et al. (2021). Therefore, simplex constraint provides essential regularization, which ensures the feasibility of our solution.

Diebold et al. (2021) outlines that the imposition of the simplex constraint is *"not only necessary to eliminate pathologies, but also desirable to provide regularization"*. Adding this constraint to the ignorance score, we obtain the following optimization problem:

$$\underset{w}{\operatorname{argmin}} \left\{ -\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} w_k \left(\sum_{m=1}^{M} p_{mkt} \mathbb{1}(y_t \in b_m) \right) \right) \right\} \tag{6}$$
$$\text{s.t. } w_k \in (0,1),\ \sum_{i=1}^{K} w_i = 1.$$

As described in Diebold et al. (2021) the L^1 simplex regularization formulates a special case of L^1 LASSO regularization, corresponding to a specific choice of LASSO regularization parameter.

3.3.2 Simplex+Divergence

Next we impose a general penalty based on the divergence between two discrete probability measures. The estimator can be written as

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left\{ -\sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} w_k \left(\sum_{m=1}^{M} p_{mkt} \mathbb{1}(y_t \in b_m) \right) \right) + \lambda D(w, w*) \right\} \tag{7}$$
$$\text{s.t. } w_k \in [0,1],\ \sum_{i=1}^{K} w_i = 1,$$

where $D(w, w*)$ is a measure of divergence between w and $w*$. Note that once the simplex restriction is imposed, w can be interpreted as a discrete probability measure on $1, 2, \ldots, K$. We will consider two different divergence measures $D(w, w*)$ to obtain new regularized estimators.

1. The L^2 norm,
$$D(w, w*) = \sum_{k=1}^{K} \left(w_k - \frac{1}{K} \right)^2, \tag{8}$$
where $w* = \frac{1}{K}$, shrinking the solution towards equal weights. Diebold et al. (2021) defines this regularization setting as the simplex plus egalitarian ridge penalty (simplex+ridge).

   Due to the nature of the simplex constraint, we obtain sparse models, which only consider a limited number of forecasters, giving the rest a 0 weight. Therefore, as proposed by Diebold et al. (2021), *"we may want to shrink all K mixture weights away from 0, thereby 'undoing' the selection implicit in the LASSO-style L^1 penalty"*. Thus we will allow for non-zero weights on all forecasts, which indeed introduces L^2 regularization.

2. Kullback-Leibler divergence (entropy) from w to $w*$,

$$D(w, w*) = -\log K - \sum_{k=1}^{K} \log w_k, \tag{9}$$

produces a "simplex+entropy" penalty, $-\sum_{k=1}^{K} \log w_k$. The simplex+entropy regularized estimator is derived from the posterior mode in a Bayesian analysis with a log score, substituting the log likelihood, and a Dirichlet prior, see Diebold et al. (2021) for a detailed derivation. Note that such formulation puts positive probability only on the unit simplex and shrinks weights toward equality for a certain configuration of hyperparameters K, thus resulting in the L^2 norm.

The motivation behind the choice of these particular divergence penalties lies in the results of Diebold et al. (2021). Particularly the two divergence penalties together with the unit simplex produce the lowest log scores. Moreover, these methods allow for a variety of forecasters to be included in the set of regularized mixtures, particularly simplex+entropy includes all available forecasters.

4 Monte Carlo

Before applying the above methods to our football data, we replicate the simulation study by Diebold et al. (2021), which reveals the potential of our regularizations. The data-generating process (DGP), assumed to be known by the forecasters, is:

$$\begin{cases} y_t = x_t + \sigma_y e_t, \ e_t \sim \text{ iid } \mathcal{N}(0,1) \\ x_t = \phi_x x_{t-1} + \sigma_x \nu_t, \ \nu_t \sim \text{ iid } \mathcal{N}(0,1), \end{cases} \tag{10}$$

where e and ν are orthogonal at all leads and lags. The variable that we forecast is y, and x_t is the so called long-run component of y_t. Each forecaster receives independent noisy signals about x_t, such that for forecaster k we have

$$z_{kt} = x_t + \sigma_{zk} \eta_{kt}, \ \eta_{kt} \sim \text{ iid } \mathcal{N}(0,1), \tag{11}$$

where η_k and $\eta_{k'}$ are orthogonal at all leads and lags for all forecasters k and k'. We assume forecasters agree that the 1-step-ahead predictive density is Gaussian with variance σ_y^2, and an unknown mean. Forecaster k uses z_{kt} to predict, thus resulting in the following predictive density of forecaster k:

$$p_{kt}(y_{t+1}) = \mathcal{N}(z_{kt}, \sigma_y^2). \tag{12}$$

In the same fashion as Diebold et al. (2021), we consider two parameterizations:

1. DGP 1: $\sigma_{zk} = 1$ for all k
2. DGP 2: $\sigma_{zk} = 1$ for $k = 1, 2, \ldots, \frac{K}{2}$ and $\sigma_{zk} = 5$ for $k = \frac{K}{2} + 1, \ldots, K$,

with $\phi_x = 0.9, \sigma_x = 1, \sigma_y = 0.5$. The difference between the DGPs is the number of good quality signals that are received by forecasters. We expect linear opinion rule to be preferred for DGP 2, at least asymptotically, giving more weight to the first $\frac{K}{2}$ forecasters, who receive better signals. For DGP 1 we can expect equal weights to be assigned to each forecaster, due to the similar prediction accuracy among all forecasters. Therefore, the benchmark model with equal weights (Simple K-Average) will be much harder to beat for DGP 1.

We chose $K = T = 20$, which is around the number of forecasters that we will consider when predicting the football games outcomes. We proceed with our simulation study by first generating data according to the DGP and then estimating mixture weights that will be used for computing the mixture densities. Finally, we generate 1-step-ahead mixture densities and evaluate them using the log score objective function. We repeat the simulation 10,000 times and compute the average log score. In our simulation we will test the following regularization methods: *Simplex*, *Simplex+Ridge* and *Simplex+Entropy*; and one benchmark method, the so-called *Simple Average*. In the latter we simply give the average weight $(1/K)$ to each forecaster.

Finally we choose the penalization strength for simplex+ridge and simplex+entropy in the same fashion as Diebold et al. (2021). We explore 20 penalization strengths for each method, where for simplex+ridge we iterate through 10 equispaced points in each of the following intervals: [1e-15,10] and [15,10000]; and for simplex+entropy we iterate through through 10 equispaced points in each of the following intervals: [1e-15,0.2] and [0.3,20].

The results of the Monte Carlo simulation are summarized in Table 1, where we present the optimized log score for each method under DGPs 1 and 2. Under DGP 1, as we expected, simple averaging performs very well, due to similar forecast accuracy among predictors, while just simplex performs worse than both simplex+ridge and simplex+entropy. Nevertheless, it still outperforms the median forecaster or even the forecaster at 25th percentile. Note that for the benchmarks we evaluate the average performance of each forecaster individually across simulations. Quite remarkably, both simplex+ridge and simplex+entropy outperform the best forecaster, and perform as well as the simple averaging method. A possible explanation is that for the DGP 1, with minimal variation among forecasters, giving every forecaster an average weight is better than limiting yourself to a subset of forecasters. The latter is evident from the fact that simplex+ridge includes every forecaster in the model, instead of only having around 12 as in the case with (unregularized) simplex.

Table 1: **Average Log Scores**

	DGP 1			DGP 2		
Regularization group	L	#	λ^*	L	#	λ^*
Simplex	0.77	11.65	NA	0.78	11.90	NA
Simplex + Ridge	0.49	20.00	7781.11	0.55	17.67	15.0
Simplex + Entropy	0.49	20.00	6.87	0.58	20.00	0.18
Benchmarks	L	#	λ^*	L	#	λ^*
Best	0.52	1	NA	0.61	1	NA
25th Percentile	0.96	1	NA	1.31	1	NA
Median	1.28	1	NA	9.84	1	NA
75th Percentile	1.65	1	NA	inf	1	NA
Worst	2.53	1	NA	inf	1	NA
Simple K-Average	0.49	20	NA	0.73	20	NA

Notes: L is the average log score, # is the number of forecasters selected, λ^* is the ex post optimal penalty parameter, and K is the total number of forecasters. 10,000 Monte Carlo replications are performed.

Under DGP 2, simple average method performs surprisingly well, although worse than under DGP 1. Despite our expectations simplex does not outperform simple averaging, however the two are significantly closer together compared to DGP 1. This result slightly diverges from the result of Diebold et al. (2021), yet it does show us that under greater variation amongst forecasters simplex method should be considered. Observe that the log scores for just simplex are almost the same for both DGPs, which is because for DGP 2 simplex chooses the best 11-12 forecasters, which we know have the same standard deviation (10 of them) as the forecasters for DGP 1. Simplex+ridge and simplex+entropy behave as expected. Furthermore, simplex+entropy outperforms every other method, including simple averaging and the 'best' forecaster. As expected simplex+ridge chose more forecasters from the opinions pool than just simplex, yet unlike simplex+entropy, not every predictor. Simplex+ridge gives the first 10 forecasters with higher accuracy a greater weight, thus weighing down some of the noise provided by the remaining 7-8 forecasters, yet keeping various information they carry.

Based on both sets of results, we can claim that regularized simplex indeed outperforms just simplex. That is because the large estimation error of the unregularized simplex (particularly DGP 2) causes some relevant forecasters to be dropped from the pool, and regularization brings them back. Hence, we have almost all of our forecasters included in the pool for regularized simplex and just over a half for simplex. The best forecaster in the simulated setting could not be outperformed for DGP 1, which is in line with the results of Diebold et al. (2021), yet it could for DGP 2.

As outlined by Diebold et al. (2021) these results are almost impossible to obtain in practice, however *"they document what can be achieved in principle"*. We now turn to the real data set and see empirically whether these methods are applicable to forecasting the density outcomes of FIFA World Cup football matches.

5 FIFA World Cup Forecasts

We now present you with the performance of our regularizations when forecasting the football games outcomes. We will start by evaluating the performance of our methods solely considering the accuracy of our probability forecasts via the log score. Then we will also consider whether the hypothetical improvements in probability forecasts can increase our expected returns from betting. Therefore, we will evaluate the model performance via the expected betting returns; and see if we can achieve systematic profits when information provided by all predictors is considered. There is a total of 23 forecasters in our (opinion) pool, including 20 bookmakers, 2 betting exchanges and a FIFA World Ranking.

5.1 Log Score

The results for the forecast accuracy of our methods and benchmarks are shown in Table 2. We consider several benchmarks, including a range of individual forecasters and a simple average. The latter simply assigns $1/K$ weight to each forecaster. The best forecaster is the one, which gives the lowest (individual) out-of-sample log score; the median forecaster gives the median (individual) out-of-sample log score, and so on.

Table 2: **Log Scores for FIFA World Cup Matches**

Regularization group	L	#	λ^*
Simplex	0.96	8.00	NA
Simplex + Ridge	0.97	23	15.00
Simplex + Entropy	0.96	23	0.02
Benchmarks	L	#	λ^*
Best (*Matchbook*)	0.95	1	NA
25th Percentile (*Betway*)	0.97	1	NA
Median (*bwin*)	0.97	1	NA
75th Percentile (*BoyleSports*)	0.97	1	NA
Worst (*FIFA World Ranking*)	1.00	1	NA
Simple K-Average	0.96	23	NA

Notes: L is the average log score, # is the number of forecasters selected, λ^* is the ex post optimal penalty parameter, and K is the total number of forecasters. The table gives the log scores for 1-game-ahead predictions, i.e. the bookmakers' coefficients (+ any other information needed to forecast) are gathered just before the start of the game. In brackets the names of corresponding forecasters are given. 5-fold cross-validation technique based on 128 games from the last two FIFA World Cups is used when computing the log scores.

Simplex gives a log score of 0.96, which is as low as simple averaging, yet higher than 0.95 - Matchbook (best forecaster). Simplex+entropy performs just as well as just simplex (0.96), yet better than simplex+ridge (0.97). Note that simplex chooses 8 forecasters from the 23 available, while simplex+ridge chooses all the available forecasters. It is important to notice that all of the individual (benchmark) log scores land in between 0.95 and 1.00, which implies that the original density forecasts offer little variation, hence we do not see major improvements via regularizations. The lack of variation in the data is also what limits the performance of regularized simplex. Unlike simplex+entropy, simplex+ridge does not have to choose every predictor from the (opinion) pool. Interestingly, simplex+ridge gives every predictor a non-zero weight, yet it fails to outperform simplex+entropy.

To further evaluate our results we examine the weights that were given to every forecaster for each regularization set up, see Table 4 in appendix A. Simplex, as expected, gives the majority of forecasters a weight of zero, however what is interesting is the size of the weight received by two predictors. Particularly, Matchbook betting exchange received the weight ≈ 0.92 and FIFA World Ranking received the weight ≈ 0.08. Therefore, the remaining six non-zero weights are arbitrarily close to zero. Given that simplex method selects a subset of predictors, and distributes non-zero weights amongst them, we conclude that FIFA World Ranking and particularly Matchbook are (amongst) the most valuable predictors. Note that FIFA World Ranking gives the worst (individual) log score, yet it is considered to be a valuable predictor by the

model. This could be due to the additional information (variation) that is provided by FIFA World Ranking, but not provided by any bookmakers. It is, however, somewhat surprising that Matchbook betting exchange received such a high weight, while Betfair Exchange received a weight of zero. The latter disagrees with the results of Wunderlich and Memmert (2016), who claim Betfair Exchange is more or less the most accurate predictor, as it is the most popular betting exchange, based on the largest trading volumes.[4] However, because Matchbook and Betfair Exchange are both betting exchanges, and they both represent the market-based forecasts, they are highly correlated. Other reasons as to why Matchbook receives high weight could be attraction of more experienced betters, or a more consistent profit margin for the odds, making the (calculated) density forecasts more accurate. Furthermore, we confirmed the results of Wunderlich and Memmert (2016), FIFA World Ranking was considered to be just as valuable (if not more) of a predictor as the bookmakers.

Simplex+entropy has somewhat similar distribution of weights amongst forecasters, except slightly smaller weight was given to Matchbook betting exchange (≈ 0.60) and the rest of the bookmakers received approximately equal weights, which lie in between 0.01 and 0.02. The FIFA World Ranking received almost the same weight as in just simplex (≈ 0.08), which again highlights its accuracy. Simplex+ridge method has kept every forecaster in the pool, perhaps due to little variation amongst forecasters. Except for the slightly higher Matchbook and the FIFA World Ranking weights, the rest of the weights converge towards simple K-averaging. The FIFA World Ranking received approximately the same weight as earlier (≈ 0.08), while Matchbook received the lowest weight (≈ 0.10) out of all regularized mixtures, yet still the highest compared to the other forecasters.

5.2 Expected Profit

As suggested by Wheatcroft (2019), Wunderlich and Memmert (2020) and Lessmann et al. (2010) it is reasonable to measure the forecast accuracy by considering the profitability of our methods. The results are presented in Table 3. Again, several benchmarks are considered, including a range of individual forecasters and a simple average. The best forecaster is the one, which gives the largest (individual) out-of-sample expected profit; the median forecaster gives the median (individual) out-of-sample expected profit, and so on.

[4]The more popular the betting exchange the more information from the betting market it will capture, so the more accurate should be the results.

Table 3: **Expected Profit for FIFA World Cup Matches**

Regularization group	EP	#	λ^*
Simplex	0.10	8	NA
Simplex + Ridge	0.11	23	15.00
Simplex + Entropy	0.16	23	0.02
Benchmarks	EP	#	λ^*
Best (*Matchbook*)	0.24	1	NA
75th Percentile (*Betfred*)	0.10	1	NA
Median (*188Bet*)	0.04	1	NA
25th Percentile (*Betsson)*)	-0.09	1	NA
Worst (*BoyleSports*)	-0.22	1	NA
Simple K-Average	-0.12	23	NA

Notes: EP is the expected profit made from betting (betting returns - bet size), # is the number of forecasters selected, λ^* is the ex post optimal penalty parameter, and K is the total number of forecasters. The table gives the maximal expected profit we would get if we were to bet under the regularized density forecasts, and we were to chose the bookmaker with the highest available odds. In brackets the names of corresponding forecasters are given. Note that log score is still used to compute optimal regularized mixtures of predictive densities, which are then used to compute the expected profit. 5-fold cross-validation technique based on 128 games from the last two FIFA World Cups is used when computing the expected profits.

Unlike in section 5.1, we are now maximizing the expected profit from betting. Note, the density weights are still computed using the log score, however the forecast accuracy is evaluated using the expected betting returns function. Such economic evaluation method yields some interesting results. One of the questions that we considered is whether it is possible to systematically make positive returns, and another is whether regularizations improve our forecasts. For both questions we get positive results. All three sets of regularized mixtures are positive (profit) and have outperformed every benchmark except for the best forecaster, who in this case is Matchbook, see Table 5 in appendix A. Interestingly, simple averaging has one of the worst performances, which is not the case when we evaluate the performance using the log score. Note that we see much more variation among the results, when considering the expected profits. Moreover, if one was to pick a forecaster randomly, his/her expected profits would be around 0 or even negative, as implied by the expected profit of the median forecaster.

In line with our results in Table 3, Matchbook betting exchange (best forecaster) remains unbeaten. Therefore, one should anticipate the highest expected profit, 24%, when betting using only the information (odds) provided by Matchbook betting exchange. While it is somewhat an unexpected result, Matchbook could have achieved such a high expected profit by incorrectly forecasting a probable event. For example, the outcome could more probable than expected by

the market (those participating in the exchange), hence there would higher odds for the outcome that is more likely, yielding higher expected profits. If one considers a good (not best) predictor, for example Betfred (see Table 3), then he/she can profit from betting no more than around 10%. However, if one was to use the regularizations, particularly simplex+entropy he/she would expect to profit up to 16% on their bet. Simplex on its own performs as well as Betfred, giving 10% expected profit. Simplex+ridge, outperforms simplex, and gives expected profit of 11%. Therefore, in terms of profitability, regularizations prove to be somewhat successful when forecasting the outcome of a football match; or at least more successful than when evaluated in terms of forecast accuracy. This goes in line with the results of Wunderlich and Memmert (2020), who argues that often we will observe no improvement in the statistical measure of forecast accuracy, yet a significant improvement in the economic measure of forecast accuracy.[5]

Wunderlich and Memmert (2020) points to the fact that profitability is not exactly a measure of the forecast accuracy. However, Koopman and Lit (2015) and Lessmann et al. (2010) agree that our end goal of improving density forecasts is to maximise betting returns. Thus, when choosing optimal forecaster or forecasting method it is absolutely relevant to consider profitability of the available methods/models. Koopman and Lit (2015) consider whether *"the forecasts from this model are sufficiently accurate to gain a positive return over the bookmaker's odds"*, which is indeed true for our regularized mixtures, yet not for simple averaging. Hence, Table 2 must be viewed with caution because according to the promising results of the simple average method one should use it to bet on the outcomes of the football matches. However, Table 3 clearly shows that such method will lead to negative returns. Instead, Matchbook betting exchange or simplex+entropy regularization should be used, if the sole goal is to maximise the expected profitability of one's bet.

[5]See Wunderlich and Memmert (2020), Lessmann et al. (2010) for more insights as to why this phenomena occurs.

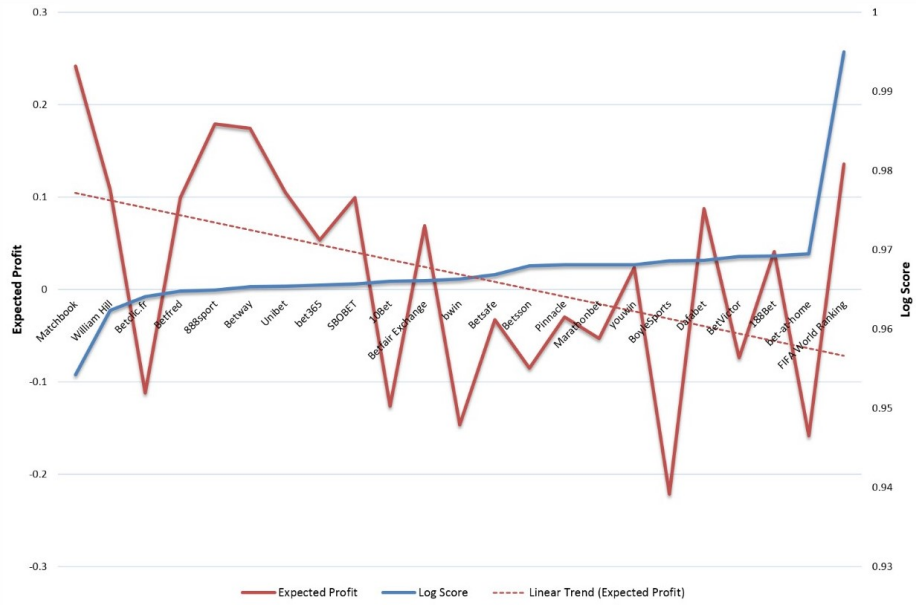

Figure 1: Log Scores and Expected Profits plot for each individual predictor. The plot shows how the expected profits change as the log score (gradually) increases. The dotted line shows the linear trend of the Expected Profit.

To further examine our results, we consider the relation between the log score and expected profit. The changes in the expected profits, as log scores (gradually) increase, are shown in Figure 1. We anticipate the expected profit to decrease as the log score increases. This relationship is only evident, if we look at the linear trend of the expected profits, which is indeed inversely related to the log score plot. Otherwise, the inverse relationship between expected profits and log scores is not necessarily obvious from the plot. For example, looking at the last data point (FIFA World Ranking), we know that FIFA World Ranking gives the worst log score, yet the fourth highest expected profit, see Table 5 in appendix A. This type of relationship between log scores and expected profits follows the results of Wunderlich and Memmert (2020).

6 Conclusion

We examine whether we can improve on density forecasts for the outcomes of FIFA World Cup football matches, by combining forecasts under regularizations as suggested by Diebold et al. (2021). Three alternative regularizations are used, simplex, simplex+ridge and simplex+entropy. 23 forecasters are considered, primarily including bookmakers, but also a couple of betting exchanges and a FIFA World Ranking. To compute optimal weights we use the log score objective function. Furthermore, we evaluate the forecast accuracy in terms of the log score, and expected profit from betting on the games outcomes. For comparison, log scores and expected profits for each individual forecaster are computed, as well as the log scores and expected profits

of a simple average model.

We evaluate our regularizations based on the last two FIFA World Cups (2014, 2018), which gives a total of 128 football matches. To compute optimal weights 5-fold cross-validation is used, and we evaluate and compare only out-of-sample predictions. One of the most notable results, is little improvement in the statistical accuracy (evaluated using log score) of density forecasts when regularizations are applied. As outlined in section 5.1, regularizations insignificantly improve on our benchmarks, meaning that there is no or little advantage considering additional predictors to improve the statistical accuracy of our forecasts. However, it seems that we do somewhat benefit from the regularizations, when looking at the expected profits that we make. Unfortunately, as in the case with the log score evaluation method, none of the regularizations outperform the best forecaster, i.e. Matchbook betting exchange in both cases. However, every other benchmark is defeated by simplex+entropy regularization. Moreover, all regularizations give positive expected profits, which vary between 10% (simplex) and 16% (simplex+ridge). The discrepancy between the two evaluation methods is separately studied by Wunderlich and Memmert (2020), Goddard and Asimakopoulos (2004) and Leitch and Tanner (1991).

In addition to the performance of our methods, we evaluate the optimal weights given to forecasters under each regularization, for results see Table 4 in appendix A. Here, the most notable result is that Matchbook always receives the highest weight, implying it is the most valuable forecaster. Matchbook, is indeed the most accurate forecaster, see Table 5 in appendix A, which justifies it receiving the highest weight. Thus, based on its accuracy (log score) and the weight it receives under regularizations, it is safe to conclude that for the World Cups in 2014 and 2018, Matchbook gives the best density forecasts compared to other forecasters and simple averaging. When evaluating Matchbook based on its profitability, its superiority is also notable. Moreover, FIFA World Ranking always receives the second highest weight, which is another notable result. The result is notable because, based on its log score, FIFA World Ranking has the worst accuracy, yet it has one of the highest weights and expected profits. FIFA World Ranking provides the 'most different' sets of density forecasts, thus it contains some information that is not considered by bookmakers or betting exchanges. To make use of that information a relatively high weight is assigned.

In practice, one can refer to Table 3 to systematically profit from the betting market. Even though regularizations yield lower profits than Matchbook (best forecaster), it might still be useful to consider regularizations since we would expect them to be more consistent in terms of generating the expected profit. Matchbook proved to be very successful in 2014 and 2018 FIFA World Cups, yet that does not mean that for other football tournaments it will be as

successful. Matchbook is a betting exchange, which depends on people actively participating in the betting market. Yet, if for some match the Matchbook trading volumes drop, it will significantly impact the forecast quality as well as the odds, since they are determined by the open betting market. Whereas regularizations would be less sensitive to each forecaster, and if more alternative forecasters are added to the opinion pool, the quality of the density forecasts should increase, both in terms of statistical and economic accuracy. One should be careful, however, because the distribution of the expected profits is difficult to interpret, and thus we cannot accurately measure the statistical significance of the results.

Despite our limiting results, we believe that there is definitely room for the above regularization methods for predicting the outcomes of the football matches. For the future, to improve their performance, one could consider the work of Frick and Wicker (2016), and make use of the economic forecasts, by adding them to the opinions pool. Furthermore, there are many model-based density predictions, which will increase the variability of the opinions pool, thus allowing regularizations to make use of the additional information. Finally, one could look to apply these methods to other sports or tournaments, possibly try and predict the overall winner; or perhaps benefit from a larger data sample, which could better estimate the optimal weights.

References

J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

F. X. Diebold and M. Shin. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, 35(4): 1679–1691, 2019.

F. X. Diebold, M. Shin, and B. Zhang. On the aggregation of probability assessments: Regularized mixtures of predictive densities for eurozone inflation and real interest rates. 2021.

B. Frick and P. Wicker. Football experts versus sports economists: Whose forecasts are better? *European journal of sport science*, 16(5):603–608, 2016.

J. Goddard and I. Asimakopoulos. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66, 2004.

C. W. Granger and R. Ramanathan. Improved methods of combining forecasts. *Journal of forecasting*, 3(2):197–204, 1984.

D. Karlis, I. Ntzoufras, et al. Bivariate poisson and diagonal inflated bivariate poisson regression models in r. *Journal of Statistical Software*, 14(10):1–36, 2005.

S. J. Koopman and R. Lit. A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 167–186, 2015.

G. Leitch and J. E. Tanner. Economic forecast evaluation: profits versus the conventional error measures. *The American Economic Review*, pages 580–590, 1991.

S. Lessmann, M.-C. Sung, and J. E. Johnson. Alternative methods of predicting competitive events: An application in horserace betting markets. *International Journal of Forecasting*, 26 (3):518–536, 2010.

H. O. Stekler, D. Sendor, and R. Verlander. Issues in sports forecasting. *International Journal of Forecasting*, 26(3):606–621, 2010.

E. Štrumbelj. On determining probability forecasts from betting odds. *International journal of forecasting*, 30(4):934–943, 2014.

E. Wheatcroft. Evaluating probabilistic forecasts of football matches: The case against the ranked probability score. *arXiv preprint arXiv:1908.08980*, 2019.

F. Wunderlich and D. Memmert. Analysis of the predictive qualities of betting odds and fifa world ranking: evidence from the 2006, 2010 and 2014 football world cups. *Journal of sports sciences*, 34(24):2176–2184, 2016.

F. Wunderlich and D. Memmert. Are betting returns a useful measure of accuracy in (sports) forecasting? *International Journal of Forecasting*, 36(2):713–722, 2020.

A. Zeileis, C. Leitner, and K. Hornik. Probabilistic forecasts for the 2018 fifa world cup based on the bookmaker consensus model. Technical report, working papers in economics and statistics, 2018.

A Appendix

Table 4: **Optimal Weights Under Regularizations**

Predictor	Simplex	Simplex+Ridge	Simplex+Entropy
10Bet	**0.00**	**0.04**	**0.02**
188Bet	0.00	**0.03**	**0.01**
888sport	**0.00**	**0.05**	**0.02**
bet-at-home	0.00	**0.03**	**0.01**
bet365	0.00	**0.04**	**0.02**
Betclic.fr	0.00	**0.05**	**0.02**
Betfred	**0.00**	**0.05**	**0.02**
Betsafe	0.00	**0.04**	**0.01**
Betsson	0.00	**0.03**	**0.01**
BetVictor	0.00	**0.03**	**0.01**
Betway	0.00	**0.04**	**0.02**
BoyleSports	0.00	**0.03**	**0.01**
bwin	**0.00**	**0.04**	**0.02**
Dafabet	0.00	**0.03**	**0.01**
Marathonbet	0.00	**0.03**	**0.01**
Pinnacle	0.00	**0.03**	**0.01**
SBOBET	**0.00**	**0.04**	**0.02**
Unibet	**0.00**	**0.04**	**0.02**
William Hill	0.00	**0.06**	**0.02**
youwin	0.00	**0.03**	**0.01**
Betfair Exchange	0.00	**0.05**	**0.02**
Matchbook	**0.92**	**0.10**	**0.60**
FIFA World Rankings	**0.08**	**0.08**	**0.09**

Notes: The table summarizes optimal weights computed under each regularization. Bold entries indicate that the corresponding weight is non-zero. Under simplex regularization many predictors receive a (small) non-zero weight that simplifies to **0.00**, when taken to 2 decimal places.

Table 5: **Expected Profit and Log Score by Predictor**

Predictor	Log Score	Expected Profit
Matchbook	0.95	0.24
888sport	0.96	0.18
Betway	0.97	0.17
FIFA World Ranking	1.00	0.14
William Hill	0.96	0.11
Unibet	0.97	0.11
Betfred	0.96	0.10
SBOBET	0.97	0.10
Dafabet	0.97	0.09
Betfair Exchange	0.97	0.07
bet365	0.97	0.05
188Bet	0.97	0.04
youwin	0.97	0.02
Pinnacle	0.97	-0.03
Betsafe	0.97	-0.03
Marathonbet	0.97	-0.05
BetVictor	0.97	-0.07
Betsson	0.97	-0.09
Betclic.fr	0.96	-0.11
10Bet	0.97	-0.13
bwin	0.97	-0.15
bet-at-home	0.97	-0.16
BoyleSports	0.97	-0.22

Notes: The table summarizes log scores and expected profits for each individual predictor. Predictors are ranked based on their expected profits; thus betting via Matchbook gives the highest expected profit, and via BoyleSports the lowest. Note that in calculating the expected profits and log scores for each predictor, only the information (density forecast/odds) provided by one predictor is considered.