A Family of Solutions Related to Shin's Model For Probability Forecasts

MELIS KIZILDEMIR[a], ERTUGRUL AKIN[b] and ALTUG ALKAN[c*]

[a]Middle East Technical University, Informatics Institute
[b]Istanbul Technical University, Department of Civil Engineering
[c]Piri Reis University, Institute of Graduate Studies

May 16, 2024

Abstract

This work develops a family of solutions related to Shin's model in producing accurate probability forecasts. More precisely, closed-form solutions are derived based on an analytical approach to known solution and evaluated experimentally for sports betting data sets.

Keywords: Probability forecasting, Shin's model, Sports forecasting

1 Introduction

In the realm of sports betting, accurate probability forecasts are essential for both bettors and bookmakers alike. The efficacy of these forecasts significantly influences decision-making processes and financial outcomes. Bookmakers, who are better at predicting event results than uninformed bettors, adjust odds not just to balance their books, but also to exploit bettor biases, highlighting the critical role of accurate forecasting in setting profitable odds (Levitt, 2004).

Probability forecasts are known to originate from three primary sources. First, there are the forecasts from the betting market itself. Second, forecasts can stem from complex models or traditional statistical approaches grounded in the fundamental aspects of sports or variables representing these traits. Third, experts like bookmakers, insiders or sports commentators also provide forecasts on the probable results of sporting events (Stekler, Sendor and Verlander, 2010).

Betting odds are widely used in determining probability forecasts because of two main reasons: their availability and reliability. Betting odds are readily visible and accessible on any bookmaker's platform and are presumed to be highly

[*]All correspondences should be addressed to: altug.alkan@alumni.pru.edu.tr

precise, as any inaccuracies could potentially harm bookmakers financially (Wunderlich and Memmert, 2020). Hvattum and Arntzen (2010) also supports this, as they highlighted the superior predictive quality of betting odds compared to various quantitative models in forecasting soccer results. They demonstrated that a consensus model based on betting odds provided more accurate forecasts for the European championship in soccer than methods using the ELO rating and FIFA World Ranking.

Different models also exist for determining probability forecasts from different main sources such as past knowledge of related features. Dixon and Coles (1997) developes a model that uses Poisson distribution based on number of goals with parametrization in terms of their historical performance metrics. Spann and Skiera (2009) compare the predictive accuracy of different models and they focus on rule based combination to manage systematic gain, including those derived from betting odds, emphasizing the diversity of models available for forecasting probabilities in sports betting. Constantinou, Fenton and Neil (2013) explore the use of Bayesian networks to model football match outcomes, employing the application of advanced statistical methods to extract forecast probabilities from betting markets.

Štrumbelj (2014) demonstrates that the probabilities derived from betting odds through Shin's model yield more precise forecasts compared to those obtained through basic normalization or regression models. He utilizes Shin's model with fixed point iteration to address scenarios in the market where there are more than two outcomes, aiming to calculate Shin probability equation of Jullien and Salanié (1994). Here, we construct an approach based on similar motivation to determine probabilities from raw odds for scenarios with three outcomes.

This paper is structured as follows: In section 2, we first review the known solution of Shin's model for $n = 2$. Then we introduce our novel closed from solutions related to Shin's model for $n >= 3$. In section 3, we compare and evaluate proposed solutions with Shin and basic normalization methods on a public dataset. Finally in section 4, we summarize and conclude our findings.

2 Analytical Solutions

2.1 Known solution of Shin's Model for $n = 2$

Shin (1993) introduced a model designed to measure the incidence of insider trading in financial markets, which have insider trading potential since individuals can access non-public information about future events to gain an unfair advantage. In the model, the expected profit of bookmakers is measured, accounting for the presence of two types of bettors: insiders, who possess insider information, and outsiders, who do not strategize in their betting. The bookie expects to payout to insider $z\sum_{i=1}^{n} p_i/\pi_i$ and $(1-z)\sum_{i=1}^{n} p_i^2/\pi_i$ to the outsiders where π_i is the array of the bookie's prices, z is the prevalence of insider trading and p_i presents array of winning probabilities. Assuming bookie's revenue is constant at 1, the expected profit (EP) is calculated by

$$EP(\pi) = 1 - \sum_{i=1}^{n} \frac{zp_i + (1-z)p_i^2}{\pi_i} \tag{2.1}$$

Considering that, in economic equilibrium, the expected profit of a bookmaker will be 0, π_i can be solved in terms of z and p.

$$\pi_i = \sqrt{zp_i + (1-z)p_i^2} \sum_{j=1}^{n} \sqrt{zp_j + (1-z)p_j^2} \tag{2.2}$$

Shin employed an iterative approach in his model, dealing with linearized versions of his equations (Jullien and Salanié, 1994). However, Jullien and Salanié (1994) showed that a straightforward nonlinear estimation method can be implemented by inverting Eq. (2.2) to yield Shin probabilities (Štrumbelj, 2014).

$$p_i = \frac{\sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\beta}} - z}{2(1-z)} \tag{2.3}$$

where $\beta = \sum_{i=1}^{n} \pi_i$.

Štrumbelj (2016) presented that z can be solved using fixed-point iteration starting at $z_0 = 0$ by using condition $\sum_{i=1}^{n} p_i = 1$

$$z_{m+1} = \sum_{i=1}^{n} (\sqrt{z_m^2 + 4(1-z_m)\frac{\pi_i^2}{\beta}}) - 2 \tag{2.4}$$

For the specific scenario where there are only two possible outcomes ($n = 2$), Eq. (2.4) possesses an analytical solution (Štrumbelj, 2016).

$$z = \frac{(\pi_+ - 1)(\pi_-^2 - \pi_+)}{\pi_+(\pi_-^2 - 1)} \tag{2.5}$$

where $\pi_+ = \pi_1 + \pi_2$ and $\pi_- = \pi_1 - \pi_2$.

2.2 New solutions related to Shin's Model for $n >= 3$

In this section, we construct an approach to use analytical solution to go further. In the case of 3 possible outcomes ($n = 3$) we propose similar symmetry with $\pi_{-i} = \pi_i - \sum_{k \neq i}^{3} \pi_k$ where $\pi_i > \beta - 1$. What we mean, in its simplest form, is to select one of the 3 outcomes, treating the other two as a single whole then apply the known solution and scan the combinations. This reduces 3-outcome option to

2-outcome and requires that z parameter in Eq. (2.5) become dependent on i-th outcome.

$$z_i = \frac{(\beta - 1)(\pi_{-i}^2 - \beta)}{\beta(\pi_{-i}^2 - 1)} \tag{2.6}$$

In this case Eq. (2.3) transforms into

$$p_i^* = \frac{\sqrt{z_i^2 + 4(1 - z_i)\frac{\pi_i^2}{\beta}} - z_i}{2(1 - z_i)} \tag{2.7}$$

p_i^* is then normalized to obtain p_i

$$p_i = \frac{p_i^*}{\sum_{i=1}^{3} p_i^*} \tag{2.8}$$

We apply our probability calculation for 3-outcomes on the market data, which includes odds for match result outcomes of 10 matches from UEFA Conferance League, given in Table 1.

Event Name	o_1	o_2	o_3	p_1	p_2	p_3	$\sum_{i=1}^{3} p_i$
Spartak Moskova - FC Orenburg	1.22	4.57	6.54	0.80059	0.13615	0.06326	1.0
Fiorentina - Maccabi Hayfa	1.22	4.63	6.38	0.80048	0.13253	0.06699	1.0
Club Brugge - Molde	1.17	4.97	7.57	0.83962	0.11833	0.04205	1.0
Slavia Prag - AC Milan	2.52	3.05	2.14	0.33278	0.25650	0.41072	1.0
Lille - Sturm Graz	1.45	3.52	4.57	0.65659	0.20782	0.13559	1.0
West Ham - Freiburg	1.44	3.67	4.49	0.66244	0.19627	0.14129	1.0
Atalanta - Sporting Lizbon	1.64	3.30	3.61	0.56884	0.22996	0.20120	1.0
M Tel Aviv - Olympiacos	2.74	3.37	1.89	0.29799	0.22257	0.47944	1.0
Viktoria Plzen - Servette	1.78	2.87	3.54	0.51507	0.27895	0.20598	1.0
PAOK - Dinamo Zagreb	1.58	3.34	3.83	0.59361	0.22440	0.18199	1.0

Table 1: Odds for given events and calculated values of p_i based on Eq (2.8).

In order to see details for behaviour of our approach, we propose p_i with $p_i = f(\beta)p_i^*$.

$$p_i = f(\beta)\frac{\sqrt{z_i^2 + 4(1 - z_i)\frac{\pi_i^2}{\beta}} - z_i}{2(1 - z_i)} \tag{2.9}$$

where $\sum_{i=1}^{n} p_i = 1$.

Assuming $\pi_1 = \pi, \pi_2 = a\pi, \pi_3 = b\pi$;

$$\beta = \pi(1 + a + b) \tag{2.10}$$

$$\pi_{-1} = \pi(1 - a - b) \tag{2.11}$$

$$\pi_{-2} = \pi(a - 1 - b) \tag{2.12}$$

$$\pi_{-3} = \pi(b - 1 - a) \tag{2.13}$$

Eq. (2.6) becomes as below.

$$z_1 = \frac{(\beta - 1)\left(\left(\frac{\beta(1-a-b)}{(1+a+b)}\right)^2 - \beta\right)}{\beta\left(\left(\frac{\beta(1-a-b)}{(1+a+b)}\right)^2 - 1\right)} \tag{2.14}$$

$$z_2 = \frac{(\beta - 1)\left(\left(\frac{\beta(a-1-b)}{(1+a+b)}\right)^2 - \beta\right)}{\beta\left(\left(\frac{\beta(a-1-b)}{(1+a+b)}\right)^2 - 1\right)} \tag{2.15}$$

$$z_3 = \frac{(\beta - 1)\left(\left(\frac{\beta(b-1-a)}{(1+a+b)}\right)^2 - \beta\right)}{\beta\left(\left(\frac{\beta(b-1-a)}{(1+a+b)}\right)^2 - 1\right)} \tag{2.16}$$

$$f(z_i, \beta) = \frac{1}{\sum_{i=1}^{3} \frac{\sqrt{{z_i}^2 + 4(1-z_i)\frac{\pi_i^2}{\beta}} - z_i}{2(1-z_i)}} \tag{2.17}$$

Now, computational experimentation on $f(z_i, \beta)$ for (a, b) pairs with $\pi_i > \beta - 1$ is summarized in Table 2.

β	$f(z_i, \beta)$
1.05	1.0256
1.10	1.0526
1.15	1.0811
1.20	1.1111
1.25	1.1429
1.30	1.1765

Table 2: β and $f(z_i, \beta)$ values for (a, b) pairs with $\pi_i > \beta - 1$. That is, if we have $\pi_i > \beta - 1$ for all i, we get $f(z_i, \beta)$ results independent of the choice of a and b values.

Table 2 suggests that we can use the power of symmetry which is $a = b = 1$, in the special case of all the odds are equal to each other $(\forall o_i, o_i = o)$, we obtain $\beta = 3\pi$ since $(\forall \pi_i, \pi_i = \pi)$and $p_i = 1/3$. Substituting β into Eq. (2.6) we obtain

$$z^* = \frac{(\beta - 1)(\beta - 9)}{(\beta - 3)(\beta + 3)} \tag{2.18}$$

and substituting 1/3 into p_i, $f(\beta)$ is obtained and z_i can be written as a norm in order to avoid range restrictions.

$$f(\beta) = \frac{1}{3\left(\frac{\sqrt{\left(\frac{(\beta-1)(\beta-9)}{(\beta-3)(\beta+3)}\right)^2 + \frac{4\beta}{9}\left(1-\frac{(\beta-1)(\beta-9)}{(\beta-3)(\beta+3)}\right)} - \frac{(\beta-1)(\beta-9)}{(\beta-3)(\beta+3)}}{2\left(1-\frac{(\beta-1)(\beta-9)}{(\beta-3)(\beta+3)}\right)}\right)} \tag{2.19}$$

$$z_i = \left|\frac{(\beta - 1)(\pi_{-i}^2 - \beta)}{\beta(\pi_{-i}^2 - 1)}\right| \tag{2.20}$$

When we substitute $f(\beta)$ in Eq. (2.9) we obtained analytical solution for p_i

$$p_i = \frac{1}{3\left(\frac{\sqrt{z^{*2} + \frac{4\beta}{9}(1-z^*)} - z^*}{2(1-z^*)}\right)} \frac{\sqrt{z_i^2 + 4(1 - z_i)\frac{\pi_i^2}{\beta}} - z_i}{2(1 - z_i)} \tag{2.21}$$

We apply the solution found in Eq (2.21) on the market data given in Table 1 and obtain same results in that precision.

In case of expanding the range of odds which falls within the range of 1.01 to 10, sum of their probabilities that is $\sum_{i=1}^{3} p_i = 1$ in 10^{-10} tolerance. It shows that our approach ensures the probability distribution in a certain region.

z^* can be generalized, when the number of outcomes are more than 3, as in $n = 3$ using the same special case with $\beta = n\pi$ and $p_i = 1/n$.

$$z^* = \left|\frac{(\beta - 1)((n - 2)^2\beta - n^2)}{((n - 2)^2\beta^2 - n^2)}\right| \tag{2.22}$$

$$z_i = \left|\frac{(\beta - 1)(\pi_{-i}^2 - \beta)}{\beta(\pi_{-i}^2 - 1)}\right| \tag{2.23}$$

$$p_i = \frac{1}{n\left(\frac{\sqrt{z^{*2} + \frac{4\beta}{n^2}(1-z^*)} - z^*}{2(1-z^*)}\right)} \frac{\sqrt{z_i^2 + 4(1 - z_i)\frac{\pi_i^2}{\beta}} - z_i}{2(1 - z_i)} \tag{2.24}$$

We applied our probability calculation (Eq 2.24) for $n = 4$ on the market data, which includes odds for total goal outcomes of 10 matches given in Table 3. For $n \geq 4$, detail investigation of our analytical approach is beyond the scope of this study but early results suggest that closed form solutions work in a certain region.

Event Name	o_1	o_2	o_3	o_4	p_1	p_2	p_3	p_4	$\sum_{i=1}^{4} p_i$
Fenerbahçe - Gaziantep	6.1	2.09	2.58	5.75	0.07785	0.47295	0.35881	0.09039	1
Arsenal - Nottingham Forest	5.22	1.98	2.74	6.99	0.11221	0.50635	0.33022	0.05122	1
Adana Demirspor - Rizespor	5.73	2.04	2.64	6.21	0.09083	0.48768	0.34762	0.07387	1
Watford - Plymouth	4.24	1.86	2.99	9.15	0.16255	0.54808	0.28851	0.00087	1
Newcastle United - Aston Villa	4.21	1.87	3.05	9.58	0.17004	0.5437	0.28361	0.00266	1
Lille - Nantes	4.24	1.87	3.04	9.45	0.16758	0.54374	0.28474	0.00394	1
Wolverhampton - Brighton	4.93	1.95	2.8	7.55	0.12609	0.51613	0.32025	0.03753	1
Almeria - Real Madrid	5.82	2.05	2.62	6.08	0.08679	0.48465	0.35102	0.07754	1
Tottenham - Sheffield United	6.16	2.08	2.54	5.53	0.06841	0.47542	0.36413	0.09204	1
Bournemouth - Arsenal	4.32	1.87	2.95	8.91	0.15686	0.54442	0.29424	0.00448	1

Table 3: Odds for given events and calculated values of p_i based on Eq (2.24).

3 Experimental Results

We evaluate the forecasting accuracy of probabilities using Ranked Probability Score (RPS) (Epstein, 1969). Ranked Probability Score can be defined as follows

$$RPS = \frac{1}{r-1} \sum_{i=1}^{r} (\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j)^2 \tag{3.1}$$

where p_j is the probability estimate of outcome j, r is the number of outcomes and e_j is the actual probability of corresponding outcome j.

In our case we have three outcomes: Home, Draw, Away and p_1 refers to the probability estimate of home team winning, p_2 refers to the probability estimate of the event ending in a draw and p_3 refers to the probability estimate of away team winning. Similarly, e_1 refers to the actual probability of home team winning, e_2 refers to the actual probability of the event ending in a draw and e_3 refers to the actual probability of away team winning.

Since results of events (H, D, A) can be considered ordinal, we believe RPS is a good evaluation metric. We compute scores for our approach, Shin and Basic Normalization methods for comparison.

We evaluate the proposed method on events played in the English Premier League between 2000-2001 season and 2020-2021 season, a total of 21 seasons (Lixi, 2021). The dataset consists of 6590 matches played in total by 40 different teams. We use odds from four different providers: Bet365, Interwetten, Landbrokes and William Hill. Table 4 and Table 5 suggest that the results are similar across three methods and it shows that our closed form solution works well.

Additionally, we contrast the outcome uncertainty for basic normalization, Shin's model and our proposed method by computing Shannon Entropy. Shannon entropy serves as a metric for assessing the level of uncertainty or randomness within a dataset or information set and it is used for an investigation with similar motivation (Štrumbelj, 2014). Shannon entropy reaches its maximum value when all outcomes are equally likely, indicating the highest level of uncertainty or ran-

Method	Mean	Median
Basic Normalization	0.1985	0.1592
Shin	0.1982	0.1592
Proposed	**0.1981**	**0.1586**

Table 4: Mean and median ranked probability scores of methods.

Odd Provider	Proposed		Shin		Basic Normalization	
	Mean	Median	Mean	Median	Mean	Median
Bet365	**0.1953**	0.1570	**0.1953**	**0.1554**	0.1954	0.1565
Interwetten	**0.1995**	**0.1590**	0.1997	0.1605	0.2002	0.1610
Landbrokes	**0.1984**	0.1595	0.1985	0.1599	0.1987	**0.1589**
William Hill	**0.1992**	**0.1590**	0.1993	0.1609	0.1995	0.1604

Table 5: Provider based mean and median ranked probability scores of methods.

domness and conversely, it reaches its minimum when there is no uncertainty, such as when all outcomes have a probability of 1 or 0, indicating complete certainty.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot \log_2(P(x_i)) \tag{3.2}$$

Method	Mean	Standard Deviation
Basic Normalization	1.4932	0.5224
Shin	1.4265	0.1809
Proposed	**1.4085**	0.1991

Table 6: Comparison of outcome uncertainities.

Table 6 and Table 7 suggest that the results are consistent and similar across three methods.

League	Proposed		Shin		Basic Normalization	
	Mean	Std	Mean	Std	Mean	Std
Bet365 A	**1.3915**	0.2179	1.4062	0.2022	1.4790	0.5608
Interwetten	**1.4238**	0.1805	1.4439	0.1602	1.5003	0.4868
Landbrokes	**1.4044**	0.2043	1.4228	0.1865	1.4897	0.5261
William Hill	**1.4144**	0.1936	1.4329	0.1748	1.5039	0.5158

Table 7: Provider based average uncertainity and standard deviation values.

4 Conclusion

In this study, we propose novel closed form solutions related to Shin's model for $n = 3$. We evaluate the proposed solution on a public dataset and compare our findings with Shin and basic normalization methods. Results confirm the findings of Štrumbelj (2014) regarding the potential of Shin's model and our findings also suggest that closed-form solutions which this study proposes can be an alternative perspective in producing accurate probability forecasts for sports betting.

Author's contributions

Mathematical foundations and proposed approach of this study is developed by Altug Alkan who is the corresponding author of this research. Ertugrul Akın and Melis Kızıldemir contributed to the computational parts of this study. All authors collaborated in the preparation of manuscript. All authors read and approved the final manuscript.

Data Availability

The original data presented in the study are openly available at https://www.kaggle.com/datasets/louischen7/football-results-and-betting-odds-data-of-epl/.

Acknowledgements

Altug Alkan would like to thank Erik Štrumbelj for some helpful suggestions and valuable feedbacks.

Conflict of Interest

The authors declare no conflicts of interest.

References

Constantinou, Anthony Costa, Norman Elliott Fenton and Martin Neil. 2013. "Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks." *Knowledge-Based Systems* 50:60–86.

Dixon, Mark J and Stuart G Coles. 1997. "Modelling association football scores and inefficiencies in the football betting market." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46(2):265–280.

Epstein, Edward S. 1969. "A scoring system for probability forecasts of ranked categories." *Journal of Applied Meteorology (1962-1982)* 8(6):985–987.

Hvattum, Lars Magnus and Halvard Arntzen. 2010. "Using ELO ratings for match result prediction in association football." *International Journal of forecasting* 26(3):460–470.

Jullien, Bruno and Bernard Salanié. 1994. "Measuring the incidence of insider trading: A comment on Shin." *The Economic Journal* 104(427):1418–1419.

Levitt, Steven D. 2004. "Why are Gambling Markets Organised so Differently from Financial Markets?" *The Economic Journal* 114(495):223–246.
URL: *https://doi.org/10.1111/j.1468-0297.2004.00207.x*

Lixi, Chen. 2021. "Football results and betting odds data of EPL.".
URL: *https://www.kaggle.com/datasets/louischen7/football-results-and-betting-odds-data-of-epl/*

Shin, Hyun Song. 1993. "Measuring the Incidence of Insider Trading in a Market for State-Contingent Claims." *The Economic Journal* 103(420):1141–1153.
URL: *http://www.jstor.org/stable/2234240*

Spann, Martin and Bernd Skiera. 2009. "Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters." *Journal of Forecasting* 28(1):55–72.

Stekler, Herman O, David Sendor and Richard Verlander. 2010. "Issues in sports forecasting." *International Journal of Forecasting* 26(3):606–621.

Štrumbelj, Erik. 2016. "A comment on the bias of probabilities derived from betting odds and their use in measuring outcome uncertainty." *Journal of sports economics* 17(1):12–26.

Wunderlich, Fabian and Daniel Memmert. 2020. "Are betting returns a useful measure of accuracy in (sports) forecasting?" *International Journal of Forecasting* 36(2):713–722.
URL: *https://www.sciencedirect.com/science/article/pii/S016920701930233X*

Štrumbelj, Erik. 2014. "On determining probability forecasts from betting odds." *International Journal of Forecasting* 30(4):934–943.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207014000533*