European Journal of Operational Research 304 (2023) 1099–1112

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

A unified theory for bivariate scores in possessive ball-sports: The case of handball

Aaditya Singh[a], Phil Scarf[b,*], Rose Baker[c]

[a] CSIS Department, Birla Institute of Technology and Science, Pilani, Goa, India
[b] Cardiff Business School, Cardiff University, Cardiff, United Kingdom
[c] Salford Business School, University of Salford, Salford, United Kingdom

ARTICLE INFO

Article history:
Received 30 September 2021
Accepted 6 May 2022
Available online 13 May 2022

Keywords:
OR in sports
Prediction
Bivariate discrete distributions
Selection bias
Shrinkage

ABSTRACT

We present a unified theory that posits three fundamental models as necessary and sufficient for modelling the bivariate scores in possessive ball-sports. These models provide the basis for perhaps more complicated models that can be used for prediction, experimentation, and explanation. First is the Poisson-match, for when goals are rare, or when goals are frequent but the restart after a goal is contested. Second is the binomial-match, for when goals are frequent and the restart uses the alternating rule. Third is the Markov-match, for when the restart uses the catch-up rule. We describe in detail the new model amongst these, the Markov-match, which is complementary to rather than competing with the binomial-match. The Markov-match is a bivariate generalisation of the Markov-binomial distribution. Its structure (catch-up restart) induces a larger correlation between the scores of competitors than does the binomial-match (alternating restart) but slightly more tied outcomes. The Markov-match is illustrated using handball, a high-scoring sport. In our analysis the time-varying strengths of 45 international handball teams are estimated. This poses some mathematical and computational problems, and in particular we describe how to shrink the strength-estimates of teams that play fewer games in tournaments because they are weaker. For the handball results, the Markov-match gives a better fit to data than the Poisson-match.

1. Introduction

The characteristics of possessive ball-sports are that a match or contest is a sequence of possessions and possession terminates with either a score or loss of possession. Possessive ball-sports include the codes of football (soccer, rugby union, rugby league, Australian rules football, American football, Gaelic football), although most of these are played with both ball in hand and at feet, the "handball" sports (netball, basketball, handball itself, water polo), and "stick and ball" sports (hockey, hurling, lacrosse, ice hockey). Mathematically, it is natural to assign a probability, q say, to the event that a possession terminates with a score and to regard a match as a sequence of possessions in which the competitors each "have their own q". These simplifications are the basis for building stochastic models of the bivariate score that is the match outcome. Thus, if q is small and the outcomes of successive possessions are independent, then the Poisson distribution is a reasonable model for the number of scores by a competitor in the match (Maher, 1982; Heuer, Müller & Rubner, 2010; Martín-González, de Saá Guerra, García-Manso, Arriaza & Valverde-Estévez, 2016), leading to the bivariate Poisson as a model for the match outcome (Dixon & Coles, 1997; Karlis & Ntzoufras, 2003; Koopman & Lit, 2015).

When q is large, and hence the scoring-rate is large, other models are desirable because the Poisson approximation is no longer justifiable. Nonetheless, a match can still be simplified as a Bernoulli sequence, but the nature of the sequence depends on the restart rule. The restart rule determines which competitor starts with possession following a score. Baker, Chadwick, Parma and Scarf (2021) discuss the case of an alternating restart and a model they call the binomial-match. The alternating restart ignores which competitor scores, and possession simply alternates, ABABA… Netball uses this rule. Under the catch-up rule (see e.g. Brams, Ismail, Kilgour & Stromquist, 2018), the conceding competitor has possession at the restart. Crucially, the restart rule impacts the match outcome if the probability that a competitor scores when they restart is different to when their opponent restarts. The catch-up rule gives the conceding competitor a temporary advantage, an op-

[*] Corresponding author.
E-mail addresses: scarfp@cardiff.ac.uk (P. Scarf), r.d.baker@salford.ac.uk (R. Baker).

https://doi.org/10.1016/j.ejor.2022.05.010

portunity to catch up, even though they may be leading. Basketball and handball use this rule. The restart can also be contested (e.g. following a goal in Australian rules football). Then, modelling the restart is no longer relevant, because the competitor with possession following a goal is unknown (unless micro-possession data are available), and the Poisson-match applies. Finally, the suite of models might include a model for a "trailing" restart rule, in which the competitor with the lower score restarts. However, to our knowledge, no possession ball-sport uses this rule, although its potential use in penalty shoot-outs has been studied (Anbarcı, Sun & Unver, 2021; Csató, 2021a,b; Csató & Petróczy, 2022).

In this paper, we discuss the catch-up restart rule and present a new model for the bivariate match outcome. We call this new model a Markov-match. In this way, we complete the suite of models for modelling the bivariate score in possessive ball-sports: the Poisson-match; the binomial-match; and the Markov-match. In this framework, the binomial-match and the Markov-match are not competing models. Instead, they are complementary models whose applicability depends on the restart rule. Nonetheless, these models, in their appropriate setting, compete with the Poisson-match, because we have argued above that the Poisson-match applies when the scoring-rate is low, but not otherwise.

Before we continue with our exposition, for clarity, we note the following. Firstly, when we write "A restarts" we specifically mean that A restarts the match with possession following a score. Secondly, we do not model possession at a micro-level. Thus, when A restarts, we suppose simply that next either A scores or B scores, ignoring the fact that any number of "turnovers" may have occurred in the meantime. This is because typically data are presented as scores. Nonetheless, the models we discuss are extendable when micro-possession data are available. Thirdly, we use the term "goal" for a score, although not all the sports above use this term, and the term "bivariate score" or more simply "the score" for the bivariate quantity that is the outcome of the match. Multiple scoring modes complicate matters (in e.g. American football, Australian rules, and rugby), but we set this aside for now.

The models we discuss are useful for: i) prediction, e.g. in gambling scenarios (Baker & McHale, 2013; Crowder, Dixon, Ledford & Robinson, 2002; Forrest & McHale, 2019; Karlis & Ntzoufras, 2011; Uhrín, Šourek, Hubáček & Železný, 2021); ii) experimentation, e.g. for testing new rules and tournament formats (Kendall & Lenten, 2017; Scarf, Yusof & Bilbao, 2009; Szymanski, 2003; Wright, 2014; Csató, 2021a; 2021b; 2021c; 2022; Lenten & Kendall, 2021); and iii) understanding, e.g. in studies of competitive balance (Koning, 2000; Manasis, Ntzoufras & Reade, 2022; Scarf, Parma & McHale, 2019) or competitor rating (Baker and McHale, 2017; Baker, 2020a; Baker & Scarf, 2021). When using models to do these things, it is important that a suite of models, from which to select the "best" model, is available. It is useful to understand the possibilities for models. The Poisson-match and binomial-match are known possibilities. The Markov-match, which we develop here, completes the set. Of course, these models can be refined, for example, to consider multiple scoring modes (Baker & McHale, 2013; Scarf, Khare & Alotaibi, 2022), and covariates (e.g. Groll, Heiner, Schauberger & Uhrmeister, 2020; Hubáček, Šourek & Železný, 2022) but the same principles apply, and to maintain focus we only study the case of a single scoring mode. Nonetheless, in so doing, we present a unified theory for bivariate scores in possessive ball-sports.

Related works fall under four themes: theory on generalisations of the binomial distribution; modelling in handball; Markov modelling in sport; and potential application of the Markov-match. On the first, a binary sequence in which the probability of "success" at the n+1-th trial depends on the outcome of the n-th was described by Markov himself in 1924 (Dekking & Kong, 2011), and the term Markov-binomial distribution was used in later works. Viveros, Balasubramanian and Balakrishnan (1994) used this binary sequence to model brand-switching. Recently, Baker (2020b) used it to generate a family of under or over-dispersed Poisson distributions. The bivariate extension that we describe is novel, however. On the second, Groll et al. (2020) modelled handball scores using the Poisson-match with covariates. Their purpose was to compare the predictive performance of this model with others that allowed over- and under-dispersion. They do not attempt to model dependence between scores. In the Markov-match, aside from the degenerate case (when $q = 0.5$ for both competitors), we will show that scores are dependant. The work most closely related to ours is Dumangane et al. (2009) because it studies possession sequences. However, they do not model match outcomes. Other published works on handball (e.g. Bilge, 2012; Csató, 2020; Meletakos & Bayios, 2010) have not modelled scores. Works on the third theme are more numerous, although the majority of these relate to models for either sequences of points in service sports (e.g. Klaassen & Magnus, 2001; Sim & Choi, 2020; Štrumbelj & Vračar, P., 2012) or the actions of competitors (e.g. Hirotsu, 2022; Hirotsu & Bickel, 2016; Ötting, 2021; Sandholtz & Bornn, 2020). To our knowledge none relate directly to match outcomes. Finally, the Markov-match could model basketball (the major code of "handball"). Basketball uses the catch-up rule and modelling studies exist (e.g. Kvam & Sokol, 2006; Song & Shi, 2020; Wolfers, 2006), some of which use the Poisson-match (Martín-González et al., 2016; Merritt & Clauset, 2014; Ruiz & Perez-Cruz, 2015). However, the existence of multiple scoring modes (penalty shots, and two- and three-point baskets) makes modelling challenging because, while basketball results (final scores) are widely available, the numbers of each type of score in matches are not.

We motivate the Markov-model using handball. This is because, in handball, scoring is vanilla (one point for a goal), scores are high (30–26 is a typical result), the conceder restarts (catch-up rule), and restarting confers a significant advantage (the restarter scores next typically with probability 0.7). We derive the model and some of its properties theoretically. In the Markov-match framework, competitor strengths are parameterised parsimoniously. We estimate parameters using the method of maximum likelihood. When we want to study mathematically intractable quantities, we use simulation. Results are validated by comparing theoretical and simulated values and by confirming results in earlier studies.

Finally, an interesting challenge is discussed that relates to the scarcity of data for some competitors because they are weak and fail to qualify for some of the tournaments for which we have data. Such selection bias (e.g. Vilkkumaa & Liesiö, 2022) may be a common problem in the analysis of elite sports because by definition elite sport is selective. We model this problem in a novel way.

The structure of the paper is as follows. We present the Markov-match first, and discuss its relationship to the Poisson-match and binomial-match, and the parameterisation of competitors strengths. Then, in Section 3, we present handball, its rules, the source of the data, and an exploratory analysis of the data. Section 4 discusses maximum likelihood estimation and presents the results of fitting the model to handball, including validation and fit comparisons. This section also discusses our novel solution to the selection bias problem. Then, we return to our thesis, the unified theory, in the conclusion, and discuss the limitations and natural extensions of the theory.

2. The Markov-match

Two competitors, A and B, play a "catch-up" contest. Let the total number of goals be N. Then, the match is a sequence of plays numbered 1,…,N. A play is a sequence of possessions between one goal and the next goal. Implicitly we assume that every play ends with a goal, so that unfinished plays (e.g. at the end of the match) are ignored. If A starts a play with possession, we say that "A plays" and likewise for B. The competitor who plays first (at the start of the match) we label A, without loss of generality. At play k (*k*=2,...,*N*), either A or B plays. Crucially, if A scores on play k then B plays next, and vice versa. This is the *catch-up rule.*

If A plays, let A score with probability q_1, so that B scores with probability $1-q_1$, since either A or B must score. If B plays, then let B score with probability q_2, so that A scores with probability $1-q_2$. At play k, if A scores then a 1 is recorded and if B scores then a 0 is recorded. The match is then a binary sequence of length N. This sequence is a Markov-Bernoulli sequence and the number of 1 s has a Markov-binomial distribution (Dekking & Kong, 2011). Thus, the score for A in the match, X_1, is the number of 1 s in this Markov-Bernoulli sequence, and that of B, $X_2 = N - X_1$, is the number of 0 s. The same idea was used by Klaassen and Magnus (2003) in tennis.

We call the score (X_1, X_2) a Markov-match.

In aside, in this setting, if we regard the match-state as the current score, then the state transition probabilities depend on which team scored last, so that the score is a Markov chain. In a more complicated model, the transition probabilities might depend on the current score. This would be interesting to study and to fit to data when the sequence of scores is observed.

The probability distribution of (X_1, X_2), $f(x_1, x_2)$, is derived in Viveros et al. (1994), but iterative computation is simpler. Now, $\Pr(X_1 = x_1, X_2 = x_2 | N)$ is the probability that x_1 goals of $x_1 + x_2 = N$ are scored by A. It can be computed by defining a sequence of probabilities $p_{i,j,k}$ that A has scored i of j goals and A restarts ($k = 1$) or B restarts ($k = 2$). Thus, a recurrence relation is defined by the equations

$$\begin{aligned} p_{i,j,1} &= (1-q_1)p_{i,j-1,1} + q_2 p_{i,j-1,2}, \\ p_{i,j,2} &= q_1 p_{i-1,j-1,1} + (1-q_2)p_{i-1,j-1,2}, \end{aligned} \tag{1}$$

($i \le j = 1, \ldots, N$), with initial conditions $p_{001} = 1$ and $p_{002} = 0$ (because A has first play). Note, the computation time of efficient code is $O((x_1 + x_2)^2)$. When N varies, with some probability distribution h_N,

$$f(x_1, x_2) = \Pr(X_1 = x_1, X_2 = x_2 | N)\, h_N. \tag{2}$$

When there are many competitors, say m, and many matches in a tournament, rather than specify a unique (q_1, q_2) for each competitor pairing, we link (q_1, q_2) to measures of competitor strength. The most parsimonious model specifies the attack strength of A as α_1 and scales that with the factor r to obtain its defence strength, $r\alpha_1$. Then, we can define

$$q_1 = g\{\frac{(\delta/r)\alpha_1}{\alpha_2}\}$$

and

$$q_2 = g\{\frac{(1/r)\alpha_2}{\alpha_1}\}$$

for some suitably chosen function $g : (0, \infty) \to (0, 1)$. Here, δ parameterises home advantage. When $g(x) = x/(1+x)$, we obtain Bradley-Terry type expressions (Dewart & Gillard, 2019):

$$q_1 = \frac{(\delta/r)\alpha_1}{(\delta/r)\alpha_1 + \alpha_2},$$

and

$$q_2 = \frac{(1/r)\alpha_2}{\alpha_1 + (1/r)\alpha_2}.$$

The exponential distribution gives

$$q_1 = 1 - \exp\left\{-\frac{(\delta/r)\alpha_1}{\alpha_2}\right\},$$

and

$$q_2 = 1 - \exp\left\{-\frac{(1/r)\alpha_2}{\alpha_1}\right\}.$$

These models have $m+1$ parameters. However, as strengths are relative, one strength parameter can be specified arbitrarily. Separate defence strengths can be specified for each competitor, whence $q_1 = g\{\delta\alpha_1/\beta_2\}$ and $q_2 = g\{\alpha_2/\beta_1\}$, and the model has $2m+1$ parameters. Baker et al. (2021) used the parsimonious exponential form to model competitor strengths in netball.

For parameter estimation, we can determine the contribution of a match outcome (x_1, x_2) to a likelihood function for parameter estimation as follows. We must specify h_N, the distribution of N, the total score. We assume a Poisson distribution with mean λ, although this is not essential. Instead, a multi-parameter discrete distribution that allows over or under dispersion relative to the Poisson could be used. Further, we set $\alpha_1 = 1$ so that strength is measured relative to competitor 1. Generally, which competitor starts with first play is unknown, so we suppose that each starts with probability ½ (coin toss). Then, the iterative scheme (Eqs. 1) must be modified so that $p_{001} = p_{002} = 1/2$. Matches that are divided into periods (e.g. halves or quarters) are a complication. If end-of-period scores are known, for example, then the likelihood contribution for each period can be specified, assuming the total score in each half has an identical distribution (e.g. Poisson), with the same (e.g. $\lambda_{\text{half}} = \lambda/2$) or different means.

The win-probability, $\Pr(X_1 > X_2)$, can be calculated using a recurrence relation analogous to Eqs. (1). Letting $D_{i,j,k}$ be the probability of a goal difference of i (A goals minus B goals) when j goals have been scored. Then,

$$\begin{aligned} D_{i,j,1} &= (1-q_1)D_{i+1,j-1,1} + q_2 D_{i+1,j-1,2}, \\ D_{i,j,2} &= q_1 D_{i-1,j-1,1} + (1-q_2)D_{i-1,j-1,2}. \end{aligned}$$

($-j \le i \le j$, $j = 1, \ldots, N$), and the probability that A wins is $\sum_N \sum_{i>0} (D_{i,N,1} + D_{i,N,2})\, h_N$. The win-probability is not needed for model fitting but is used in Section 5.1 for prediction and model evaluation. A simpler asymptotic approximation that works well is also used. This is derived in Appendix 1.

In principle, the moments, $E(X_i)$, $\text{var}(X_i)$, $\text{corr}(X_1, X_2)$, can be specified with recurrence relations, but it is simpler to calculate these by simulation.

2.1. The unified theory

Our thesis is that three fundamental models are necessary and sufficient to model bivariate scores in possessive ball-sports: the Poisson-match due to Maher (1982); the binomial-match (Baker et al., 2021); and the Markov-match formulated in this paper. They are necessary and sufficient because there are two basic cases: a) in a possession event the probability of a goal is small e.g. in soccer a typical match has hundreds of possessions and a few goals; b) in a possession event the probability of a goal is large e.g. a netball match at elite level has of the order of 150 possessions and 100 goals—the ball is rarely "turned over". A consequence of b) is that possession at the restart makes a difference, and so the restart rule must be modelled when the restarter is known, and there are two such restart rules: b1) alternating; b2) catch-up. So, there are three cases: a, b1, b2.

For case b2, the Markov-match, described above, is the appropriate model.

For case b1, the binomial-match is the appropriate model. Here, there are $N \sim \text{Poisson}(\lambda)$ plays, A plays first, and then plays alternate regardless of who scores, so that each competitor has $N_1 = N_2 = N/2$ plays if N is even or A has $N_1 = (N+1)/2$ plays and B has $N_1 = (N-1)/2$ if N is odd. Then,

$$(X_1, X_2) = (Y_1 + N_2 - Y_2, Y_2 + N_1 - Y_1),$$

where $Y_1 \sim B(N_1, q_1)$ and $Y_2 \sim B(N_2, q_2)$ independently, because, at the plays of A, successes (goals) acrue to A and failures acrue to B, and vice versa.

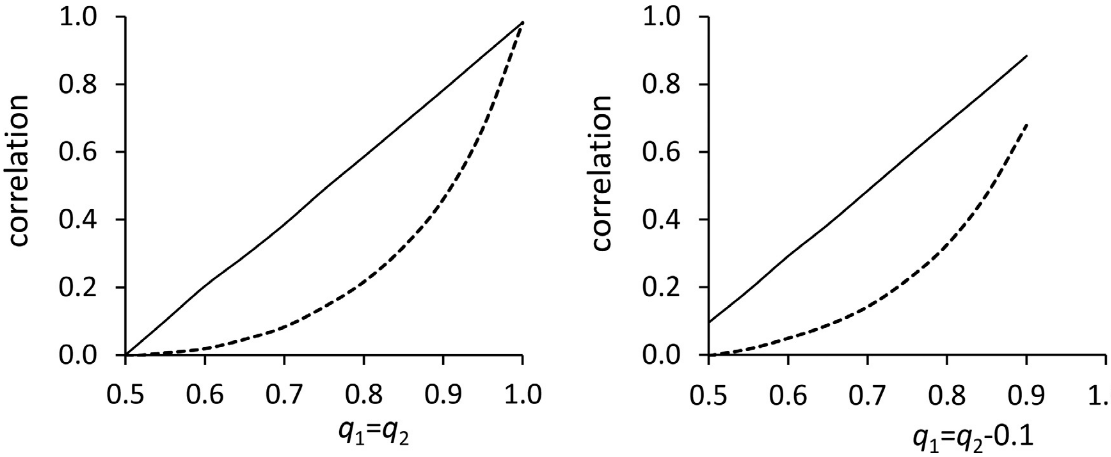

Fig. 1. Corr(X_1, X_2) for Markov-match (solid line) and binomial-match (dashed line) with $N \sim$ Poisson$(\lambda = 56)$. Left: equal strengths. Right: unequal strength.

For case a), the Poisson approximation to the binomial justifies the Poisson-match, in which $X_1 \sim \text{Poisson}(\mu_1)$ and $X_2 \sim \text{Poisson}(\mu_2)$. Further, when the restarter is unknown, because the restart is contested, assuming there are $N \sim \text{Poisson}(\lambda)$ plays implies that: $N_1 \sim \text{Poisson}(q\lambda)$ and $N_2 \sim \text{Poisson}((1-q)\lambda)$; and N_1 and N_2 are independent. Thus, the Poisson-match is obtained. Note, N_1 is a thinned Poisson, and by symmetry so is N_2; N_1 and N_2 are independent because if they were not then their sum would not be Poisson.

Common to the three models is the notion that the model parameters are not unique for each paired contest but instead depend on the attack and defence strengths of the competitors in the contest. In this way, a tournament with m competitors can be parameterised with m strength paramaters rather than $m(m-1)$ pairs (μ_1, μ_2) or (q_1, q_2).

Each model can be generalised in many ways, e.g. time-varying strengths, multiple scoring modes. All three models are approximations to the reality, e.g. ends of periods (half-time) interrupt plays. Nonetheless, the fundamental principles remain.

2.2. Comparison of the Markov-match with the binomial-match

It is interesting to compare the outcome uncertainty and the correlation of scores in the Markov-match (catch-up restart) with those in the binomial-match (alternating restart). This is because, for fixed competitor strengths, the restart rule may influence outcome uncertainty (outcomes may be closer on average) and this may be explained by dependence (e.g. correlation) between scores (Scarf et al., 2022). Note, the purpose of this comparison is not to decide which model is better, because, as we discuss above, the models are complementary. Rather, it is to compare restart rules within a theoretical framework (the unified theory). Relevant quantities were simulated, noting 10^6 repetitions of a match were required to ensure the standard deviation of the win percentage was less than 0.1. For the two match-types, all conditions were identical except the restart rule. Thus, for both $N \sim$ Poisson. We use $\lambda = 56$ because this is the mean total score in a match in the Handball World Championships, which we consider later. This was a slight departure from the definition of the binomial-match in Baker et al. (2021), in which each competitor received a random Poisson but equal number of starts. This leads to a subtly different correlation structure (Appendix 2). To be consistent with the calculations in Appendix 2, we use the Pearson correlation (Fig. 1), although because the assignment of competitors to A is arbitrary, the intraclass correlation may be more appropriate. We can see that the correlation is higher when the teams are unequal strength-wise (left/right comparison of the figures), and that scores have a higher correlation under the catch-up rule (Markov-match) than the alternating restart rule (binomial-match).

Table 1 indicates that the alternating restart (binomial match) has slightly fewer tied matches and favours the weaker competitor in comparison to the catch-up restart (Markov-match). This is rather counter-intuitive, particularly when the correlation for the Markov-match is generally larger than the binomial-match. However, the guaranteed, occasional extra play (restart) that one competitor receives under the alternating restart rule (when N is odd) may be the influencing factor here, because this one-play advantage, which occurs in half of matches, accrues to the weaker or stronger competitor with equal probability. Thus, the weaker competitor may not lose as often as it might because in 50% of matches it has an extra play.

2.3. The multi-modal Markov-match

In aside, here, we outline some ideas for modelling a catch-up contest with multiple scoring modes, e.g. basketball with three scoring modes for 1, 2 and 3 points respectively, although this is a simplification. In principle, it is possible to estimate score probabilities for competitors for each mode of scoring when only the total points-value is known if some assumptions are made about the relative rates of different scoring modes. However, we will suppose the numbers of scores for each score type (mode) for each competitor are known.

Proceeding specifically, using the example of basketball, if A restarts then there are 6 possible outcomes: A scores a 1 or 2 or 3 point-basket and B has the restart or B scores a 1 or 2 or 3 point-basketball and A retains the restart. So, we can use the Markov-match framework but with extra probabilities relating to mode of scoring. Two parameterisations are natural.

Let A score with probability q_1 if restarting, otherwise B scores with probability $1-q_1$. Then, conditional on A scoring, 1, 2 or 3 points are scored with probabilities b_{11},b_{12},b_{13} respectively, where these probabilities sum to unity. Similarly, conditional on B scoring, the corresponding probabilities are b_{21},b_{22},b_{23}, which sum to unity. The strength parameters for team A are then e.g. q_1,b_{12},b_{13}.

In the second parameterisation, the probabilities of scoring are absolute, so that $b_{11}+b_{12}+b_{13}=q_1$. These three probabilities are the strength parameters, and similarly for team B.

$$\frac{b_{11}+b_{12}+b_{13}+(1-b_1)b_{21}}{b_2}/+\frac{(1-b_1)b_{22}/b_2+(1-b_1)b_{23}}{b_2}=1,$$

Then, recurrence relations to calculate the likelihood (probability distribution) of the outcome $(x_{11}, x_{21}, x_{31}, x_{21}, x_{22}, x_{23})$ conditional

Table 1
Percentages of outcomes (win, tie, loss) and win-loss ratio in Markov-match (catch-up) and binomial-match (alternating) as a function of q_1 (rows) and q_2 (columns). $N \sim \text{Poisson}(\lambda = 56)$.

		0.5				0.6				0.7				0.8			
		win	Tie	loss	ratio	win	tie	loss	ratio	win	tie	loss	ratio	win	tie	loss	ratio
0.5	*catch-up*	47.4	5.3	47.4	1.00												
	alternating	47.3	5.3	47.4	1.00												
0.6	*catch-up*	75.0	4.5	20.4	3.67	46.5	6.4	47.1	0.99								
	alternating	75.3	4.0	20.7	3.64	47.4	5.3	47.3	1.00								
0.7	*catch-up*					75.8	5.4	18.8	4.02	45.8	8.1	46.2	0.99				
	alternating					76.3	4.1	19.6	3.89	46.8	5.9	47.2	0.99				
0.8	*catch-up*									77.0	6.4	16.6	4.64	44.8	10.6	44.6	1.00
	alternating									78.4	4.3	17.4	4.51	46.6	6.6	46.7	1.00

on $N \sim \text{Poisson}(\lambda)$ can be found. Competitor strengths can then be parameterised as described in Section 3.

3. Handball

Two teams, of seven players, compete on a court with a goal on the floor at each end of the court. The ball is moved by hand. A player in possession can dribble, pass, shoot, or hold the ball for up to three seconds or take three steps. Shooting is not allowed within six metres of the goal. A goal is scored when the ball is thrown into the goal. The team with the most goals wins. Normal playing time is two halves of 30 min each. Overtime (two halves of five minutes each) is played when a match is tied at the end of normal time and a winner has to be determined. Overtime can be repeated once, and if necessary is followed by a penalty shootout. Play starts each half with a contested “throw-off”. Following a goal, the conceding team restarts with possession.

We scraped data of 753 matches at eight editions (2007–2021) of the World Men’s Handball Championship from https://www.ihf.info/competitions, including for each match: date, location, round, teams, and scores in each period including overtime and penalties. The format of the tournament has changed over this period with varying numbers of teams and matches (Csató, 2021c). Broadly, qualifiers from the initial group stage proceed to knockout (KO) stages while non-qualifiers play a “consolation” tournament, the Presidents Cup. In total, 45 national teams participated in the eight tournaments (Table 3, Appendix 3). The scores recorded for two matches in 2021 that were voided because Cape Verde withdrew from the tournament were omitted (Cape Verde vs Germany, Cape Verde vs Uruguay). Then, of the 751 matches, 43 were tied at the end of normal time. Of these 43, 12 played first overtime, and a further 4 played second overtime, and a further one match ended with a penalty shootout. The remaining 31 tied matches occurred in the group stages. Note, we might have omitted matches in the President’s Cup, because arguably there may be less incentive for teams to play at full strength despite the results contributing to world rankings.

The scoring rate is high (Figs. 2,3) and has not changed much over time (figures omitted). An approximate representation of team strengths is shown in Fig. 4. Analysis of scoring rate by tournament round, comparing scoring rate in group stage, President Cup and KO stages (results omitted for brevity), indicates that scoring-rate is not quality related. In a match at the highest level (see Table 4, Appendix 4), we see $q_i > 0.7$ and $> 50\%$ of plays with only one possession (no turn-over).

We fitted the Markov-match to the scores in these matches. We discuss that in the next section. Our purpose is to demonstrate the model and how to fit it to match results. This then provides some indication of the usefulness of the model. In Fig. 4, we can see that many teams were absent from tournaments. This is because they failed to qualify on those occasions. Therefore, the dataset is biased—there is more information about stronger teams than weaker teams and ~~very~~ little information about some teams e.g. Cape Verde (one match only). We address this important issue in the analysis that follows.

4. Estimation of parameters

In a simple approach, one might assume that competitors possess strengths that are fixed for all time. Then, the (vanilla) log-likelihood function is the sum of the log-likelihood contributions for each match (Eq. (2)). However, strengths are continuously evolving. Therefore, at a time instant *t*, competitors’ strengths are estimated using the results of matches up to time *t*. However, three issues arise: how to discount (down-weight) past matches; how to handle a competitor that has played no matches by time *t*; and how to adjust the estimation when such absence of match results is related to strength. The first issue is discussed in detail in Baker et al. (2021) and we use the same approach here. Thus, results from a match played τ time units ago are weighted (discounted) by the term $\exp(-\xi\tau)$. Since matches in a single tournament are played close together in time, our approach is close to that of block-discounting.

For the second, shrinkage is now standard. With shrinkage, at a particular time point, the log-likelihood is a weighted sum of terms $\Delta\ell_i = -2(\log\alpha_i - \mu)^2/\sigma^2$, one for each competitor, and the (discounted) term corresponding to the results of all matches to date. In $\Delta\ell_i$, μ and σ^2 are notionally the mean and variance of the (log-) strength of an “unknown” competitor (no games played to date). Thus, if competitor *i* is unknown, $\Delta\ell_i$ is the only place that its strength appears in the log-likelihood, and the maximum likelihood estimate of α_i is $\hat{\alpha}_i = \hat{\mu}$, and $\hat{\mu}$ is estimated as the mean log-strength of “known” competitors. Conversely, the more matches a competitor has played the less their estimate is shrunk.

In the dataset that we use for handball, the absence of match results is related to strength. This is because a team is absent from a tournament because they failed to qualify and weaker teams are less likely to qualify. The results of qualification matches were not available. Therefore, an alternative to standard shrinkage is desirable. We would expect the log-strength of an unknown competitor to be somewhat less than the mean log-strength of “known” competitors. Therefore, we propose a method that is adapted for qualification. This is an original contribution of this paper, and is motivated by Heckman (1977) who pointed out that a sample of the wages of those working is biased. We discuss this next.

4.1. Modelling qualification for a tournament

We use an empirical Bayes approach, and essentially suppose that the strength of a randomly chosen competitor arises from a normal distribution. However, strength “seen” in qualification, denoted *U*, and strength “seen” in tournament results, denoted *V*, are assumed different but correlated, with a correlation coefficient of ρ. As qualification is a binary process, it is sufficient to specify

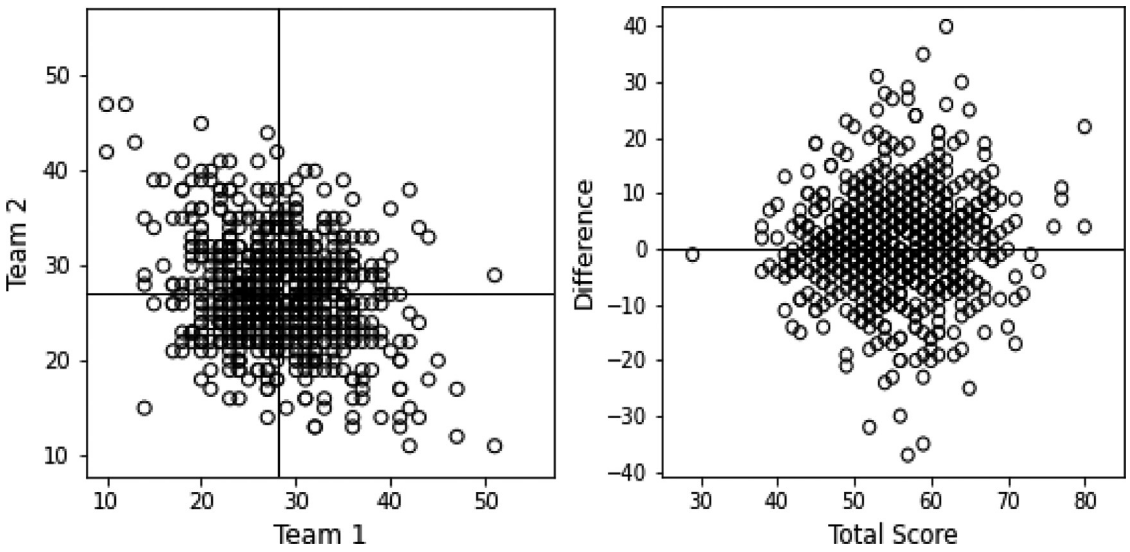

Fig. 2. Match scores in the World Men's Handball Championship 2007–2021. Left: score of team 1 (first named on score sheet) vs score of team 2 (second named). Mean scores shown as horizontal and vertical lines. Right: total score versus score difference.

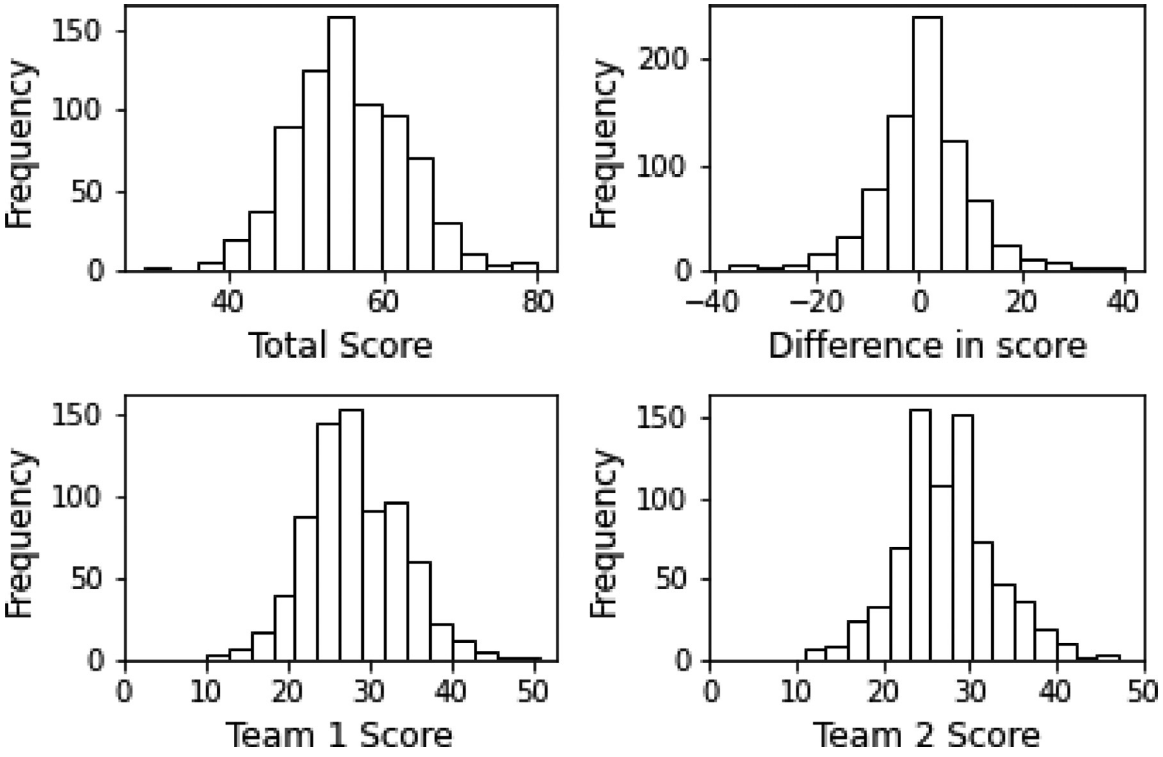

Fig. 3. Histograms of scores in World Men's Handball Championship 2007–2021. Note, a goodness-of-fit test for total score rejects Poisson (p<0.001).

$U \sim N(0,1)$ and to suppose a competitor qualifies if $U > a$. We suppose $V \sim N(\mu, \sigma^2)$, as in the standard shrinkage approach. Setting $W = (V-\mu)/\sigma$, the distribution of U and W is then bivariate normal $BN(0,1,0,1,\rho)$ and the joint pdf is given by

$$f_{U,W}(u,w) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{u^2+w^2-2\rho uw}{2(1-\rho^2)}\right\}.$$

The probability of qualification Q is given by

$$Q = \Pr(U > a) = \int_a^\infty f_{U,W}(u,w)du = \Phi\left\{\frac{(\rho w - a)}{\sqrt{1-\rho^2}}\right\}. \qquad (3)$$

where Φ is the standard normal distribution function and $w = (\log\alpha_i - \mu)/\sigma$. The terms $\log Q$ (from Eq. (3)) are added to the log-likelihood for each occasion that a competitor qualifies and terms $\log(1-Q)$ are added for each occasion that it does not. Since $\log(1-Q)$ is maximised when $Q=0$, that is, when $\mu = \infty$, non-qualification tends to inflate μ, the mean strength of the other, qualifying teams. Thus, equivalently, the more often a competitor fails to qualify, the more its strength decreases relative to the overall mean strength.

Note, in aside, when V is estimated with standard error ν, it can be shown that $Q = \Phi\{(\rho w - a)/\sqrt{1-\rho^2+\rho^2\nu^2/\sigma^2}\}$. That is, the error on V causes Q to shrink towards 1/2. Also, further, considering a single tournament and qualification for that tournament,

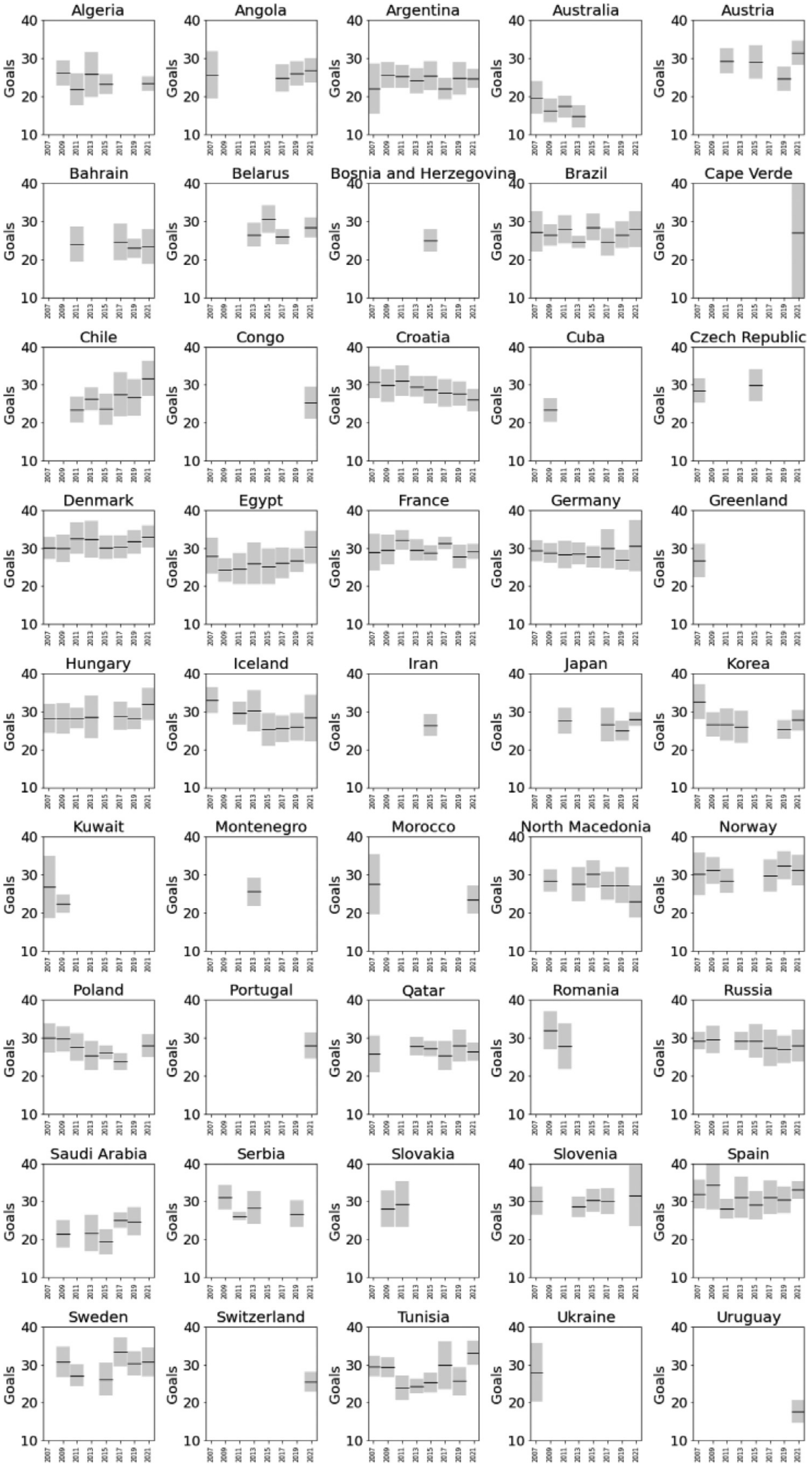

Fig. 4. Mean goals per match (+/- 2 standard error of mean) by team and tournament for 45 teams that competed in the World Men's Handball Championships 2007–2021.

the full ‘prior’ pdf is a generalization of Azzalini’s skew-normal distribution (Azzalini & Capitanio, 2014) with pdf

$$f(v|\rho) = \Phi\left(\frac{\rho w - a}{\sqrt{1-\rho^2}}\right)\frac{\exp(-w^2/2)}{\sqrt{2\pi\sigma^2}\Phi(-a)},$$

and defining S as 1 if the competitor qualifies, else −1, the mean of this prior distribution is

$$E(V|S) = \mu + \frac{S\rho\sigma\exp\{-a^2(1-\rho^2)/2\}}{\sqrt{2\pi}\Phi(-aS)}.$$

Returning to the method we use, the qualification threshold parameter, a, is estimated using the proportion of competitors that qualify for a tournament. Thus, in a tournament with q competitors, a is the solution of $\Pr(U > a) = q/45$ and $U \sim N(0,1)$ so $a = \Phi^{-1}(1 - q/45)$. When 21 of 45 competitors fail to qualify (tournaments in 2007 to 2019), $a = -0.084$. The correlation ρ is estimated using the modified likelihood function, discussed later.

The estimated strength of competitors that have not yet played a tournament can be estimated in a second step by using these estimates. The additional term in the log-likelihood for strength of a competitor that has failed to qualify for n consecutive tournaments is

$$l = \frac{1-\exp(2n\xi)}{1-\exp(2\xi)}\log\{\Phi(\frac{\rho w - a}{\sqrt{1-\rho^2}})\} - w^2/2.$$

Here, the leading factor arises from the block discounting, and the geometric series of weights has been summed. The probability Φ is the probability of not qualifying, and the last term is the normal prior. Constants have been discarded. To determine the standard error of the estimated strength, this likelihood was maximised for w using Newton-Raphson iteration, which requires analytic first and second derivatives of l. Then, the variance of w, and hence $\log\alpha$, can be found from the Hessian, the matrix of second derivatives.

Failure to qualify pulls the estimated strength down more as n increases. However, the discounting term is large, so the estimated strength and its standard error soon reach a constant. This seems realistic: the more times a competitor fails to qualify, the weaker we may consider it, but discounting allows for variation in competitor strength with time, so even a competitor that failed to qualify many times may now not be considered very weak after all.

4.2. Outline of computation strategy

Tournaments, at two-yearly intervals, were labelled 1,…,*T*. Strengths were estimated at times immediately following the end of each of these T tournaments. Model parameters were the competitor strengths α plus the other ancillary parameters. Library function minimisers were used to maximise the adjusted log-likelihood. One team’s strength is held constant, as all observable results depend on the ratio of strengths. Denmark was chosen, having played in all tournaments.

Qualification-adapted shrinkage (as described in Section 4.1) was used. The full-modified likelihood included terms for: results of matches; qualification or otherwise for each tournament; and shrinkage for teams who had played few matches. Teams with no matches to date (at the time of estimation) were omitted in this first step of likelihood maximisation. Data comprised scores from each half of a match for the subset of the 45 teams who competed at each tournament. Each half is modelled separately because the *catch-up* rule does not carry through the half time interval.

Two parameters could not be estimated by maximum-likelihood, the discounting rate ξ and the σ^2, variance of the prior distribution of logged strengths. The estimation of these two parameters is described next.

4.3. Estimation of ξ and σ^2

The discounting parameter ξ cannot be estimated by maximum-likelihood because it changes the effective amount of data in the analysis. However, the weighted likelihood (pseudo-likelihood) can be corrected, as follows. The prior term is omitted, leaving the major part of the log-likelihood that pertains to match scores. Denote this by l'. Then, by calculating the (common) expectation of the contribution of match i of n with weight w_i to the log-likelihood, it can be shown that $C = nl'/\sum_{i=1}^{n} w_i$ has the expectation that ℓ would have without discounting. Therefore, ξ can be chosen to minimise C.

Now, σ^2 (the variance of the ‘prior’ distribution) cannot be estimated with the other model parameters because the prior term in the likelihood l is such that $l \to \infty$ as $\sigma^2 \to 0$ and $\log\alpha_i \to \mu$ for all i. Therefore, it is sensible to maximise the marginal log-likelihood of σ^2,

$$l_m(\sigma^2) = \log\int\exp(l(\theta))\mathrm{d}\theta, \quad (4)$$

where θ is the vector of all parameters except σ^2. The integration in Eq. (4) is intractable, therefore, we use Laplace’s approximation

$$l(\theta) \approx l(\hat\theta) - (\theta - \hat\theta)^T \mathrm{D}\,(\theta - \hat\theta)/2,$$

where $\hat\theta$ is the maximum-likelihood estimate of θ, and the matrix D is the negative of the Hessian, that is, $\mathrm{D}_{ij} = -\partial^2 l/\partial\theta_i\partial\theta_j|_{\theta=\hat\theta}$. This approximation can be integrated analytically to give

$$l_m(\sigma^2) = l(\hat\theta) - (1/2)ln|\mathrm{D}| - (n/2)\log\sigma^2,$$

where |D| is the determinant of D, and then maximised to determine $\hat\sigma^2$.

5. Results of fitting the Markov-match

Fig. 5 shows the strength estimates for the 45 teams at $t=8$ (2021). In these plots, estimated strengths are rescaled so that $\sum\log(\hat\alpha_i) = 0$. Otherwise, the constraint $\alpha_{\mathrm{DEN}} = 1$ used in the parameter estimation step would imply a time-constant strength for Denmark. The large standard error for Cape Verde indicates that their strength is estimated using one match only (vs Hungary). Table 2 shows estimates of the other (ancillary) model parameters with standard errors for the two models, and the percentage of correct win/lose forecasts, plus other performance statistics such as Brier score (Brier, 1950) and the mean absolute error on predicted score difference. The final at the 2021 World Championships was Denmark vs Sweden (see Appendix 4). Their strength estimates, $(\alpha_1, \alpha_2) = (1.465, 1.396)$, imply $(q_1, q_2) = (0.699, 0.664)$, $E(X_1, X_2) = (28.7, 27.3)$, and win (for Denmark) and tie probabilities of 0.57 and 0.05; the actual outcome was (26,24).

The Poisson-match we fit uses the most parsimonious parameterisation of strengths: $\mu_1 = \delta\alpha_1/r\alpha_2$ and $\mu_2 = \alpha_2/r\alpha_1 X_1$, and assumes X_1 and X_2 are independent. For this model, the total number of goals is also Poisson with mean $\mu_1 + \mu_2$. The win probability (for team 1) is approximately $P = \Phi\{(\mu_1 - \mu_2)/\sqrt{\mu_1 + \mu_2}\}$, and with a continuity correction, so that a difference less than ½ is regarded as a tie, it is

$$P = \frac{1}{2}\left[\Phi\left\{\frac{\mu_1 - \mu_2 - 1/2}{\sqrt{\mu_1 + \mu_2}}\right\} + \Phi\left\{\frac{\mu_1 - \mu_2 + 1/2}{\sqrt{\mu_1 + \mu_2}}\right\}\right].$$

Time was measured on a continuous scale with unit of time as one year, and exponential discounting was used. Therefore, strengths were effectively estimated using the results of approximately $1 + e^{-2\xi} + e^{-4\xi} + \ldots$tournaments, that is, ~1.5 tournaments when $\xi = 0.55$.

The log-likelihood is higher for the Markov-match, which has one extra parameter, the mean total number of goals scored. The

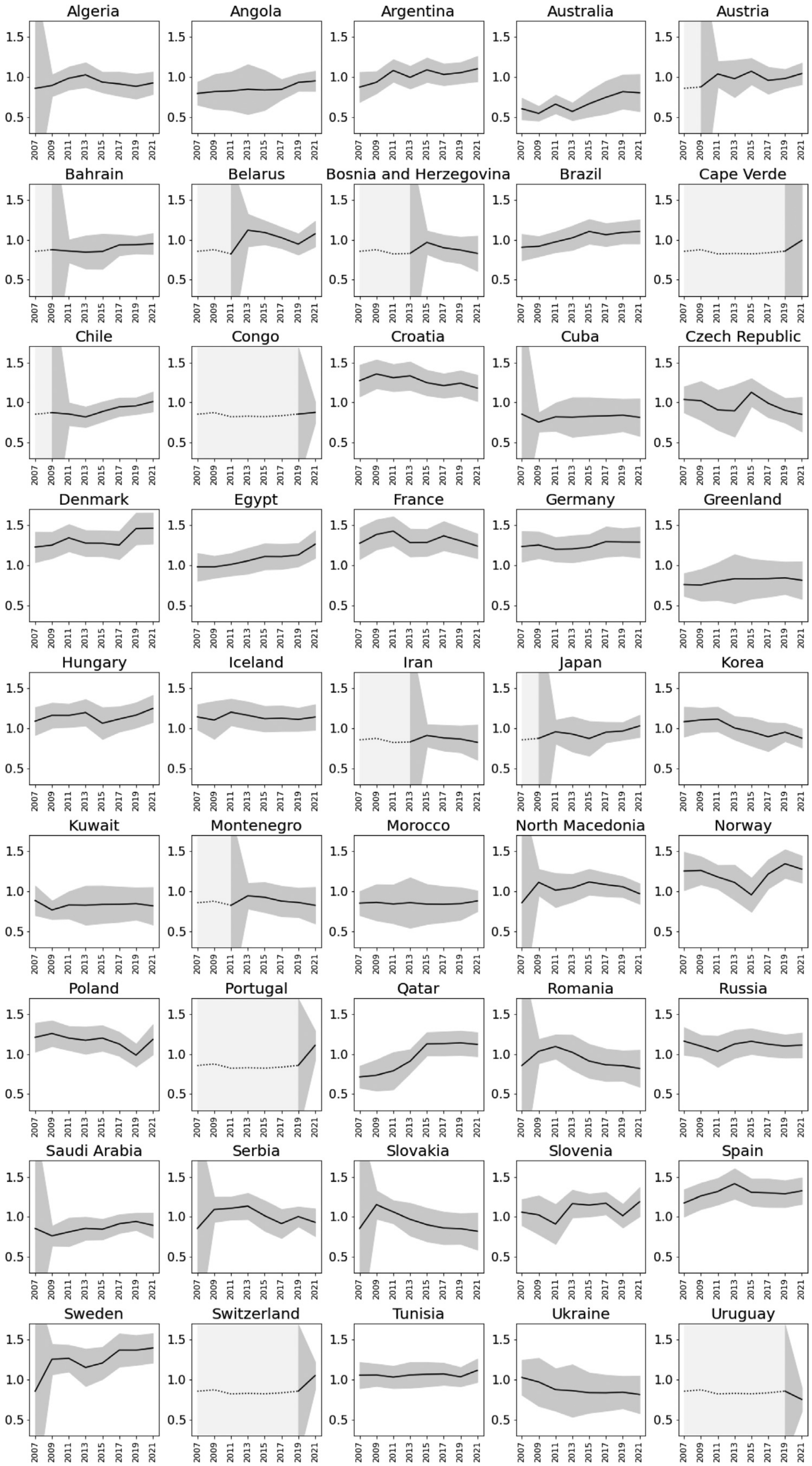

Fig. 5. Evolution of estimated strengths (ribbons at +/− two standard errors) for the Markov-match. Estimates shown dotted prior to first qualification. When $(\alpha_1, \alpha_2) = (1, 1)$, $q_1 = q_2 = 0.682$, $E(X_1, X_2) = (28, 28)$, and win and tie proportions are 0.46 and 0.08. When $(\alpha_1, \alpha_2) = (1.1, 0.9)$, $(q_1, q_2) = (0.753, 0.608)$, $E(X_1, X_2) = (31, 25)$, and win and tie proportions are 0.86 and 0.04.

Table 2
Quantities of interest in the estimation of the Poisson-match and Markov-match in 2021. Some selected, rescaled strength estimates are shown.

	Poisson-match		Markov-match	
	coeff.	s.e.	coeff.	s.e.
log-likelihood	−1093.13	–	−874.73	–
mean total goals ($\hat{\lambda}$)	–	–	55.98	0.81
R	0.072	0.001	0.874	0.065
Ξ	0.45	–	0.55	–
Σ	0.234	–	0.288	–
P	0.961	0.021	0.906	0.046
M	−0.231	0.046	−0.418	0.069
α_{DEN}			1.465	0.097
α_{SWE}			1.396	0.092
% correct results	74.09	–	74.25	–
Brier score	0.169	–	0.168	–
Score difference error	5.05	–	4.99	–
Hosmer-Lemeshow χ^2	15.70	-	10.99	–
γ	1.233	0.091	1.021	0.091

percentage of wins correctly predicted in “one-step ahead” forecasts is however very similar to the Poisson-match. Results were also forecast using a simple ‘strawman’ system, based on the cumulative percentage of total goals in matches scored by particular teams in earlier tournaments. This forecast 72.4% wins correctly, slightly worse than either of the parametric models.

Since a model without home advantage had a lower AIC, δ is not estimated.

The strength seen in qualification is highly correlated with strength seen in tournament results ($\rho = 0.906$). We expect a high correlation. One might argue that the parameter ρ is not needed in the model, so that qualification could be modelled directly using “tournament strength”. However, this part of the model regularises the likelihood, so that it is “smooth” with respect to inclusion or exclusion of teams, thus preventing numerical difficulties with likelihood maximisation.

With $\mu = -0.418$, the strength of the “average team” at 2021 estimation point is 0.66, indicating that strengths are positively skewed, so that there are (at least at 2021) a relatively small number of teams that are very strong. This “average team” includes non-qualifiers and this to an extent explains the lower value.

5.1. Goodness of model fit

Model fit for complicated models is difficult to assess. We base goodness of fit tests on the forecast probabilities of winning. Correction for parameter uncertainty shrinks the win probability towards 1/2. Win-probabilities were computed using Monte-Carlo simulation (10,000 realizations of the multivariate normal distribution of parameter values) allowing for parameter uncertainty.

We use a standard measure of fit from logistic regression, the Hosmer-Lemeshow statistic (Hosmer, Lemeshow & Sturdivant, 2013), on the predicted (out-of-sample) probability of winning the game. The reference team (for the purpose of defining a win) was the team with the lower ID code. The predicted winning probabilities were sorted into ascending order and divided into ten groups. In each group, the observed and expected numbers of reference team wins were used to form a chi-squared, χ^2. The results are given in the Table 2. The fit of the Markov-match is acceptable ($p = 0.257$), while the Poisson-match ($p = 0.048$) shows some evidence of lack of fit.

Another approach to assess goodness-of-fit adds extra model parameters and assesses whether they are needed. In this way, the forecast probabilities of winning p_{F} were transformed to $q_{\mathrm{F}} = h(p_{\mathrm{F}})$. One could, for example, transform the win probability for the reference competitor as: $p_{\mathrm{F}} \rightarrow p_{\mathrm{F}}^{\gamma}$. However, a suitable transformation must give the same result, whether applied to p_{F} or to $1 - p_{\mathrm{F}}$, and hence to q_{F} and $1 - q_{\mathrm{F}}$. A monotone transformation with this property is implied by $\log(q_{\mathrm{F}}/(1 - q_{\mathrm{F}})) = \gamma \log(p_{\mathrm{F}}/(1 - p_{\mathrm{F}}))$. Thus, the log-odds are scaled, that is

$$q_{\mathrm{F}} = \frac{p_{\mathrm{F}}^{\gamma}}{p_{\mathrm{F}}^{\gamma} + (1 - p_{\mathrm{F}})\gamma}.$$

This transformation pulls probabilities towards $q = 1/2$ if $\gamma < 1$, pushes them towards extremes if $\gamma > 1$, and it still works when competitors cannot be meaningfully distinguished. Estimates of γ for the two models are shown in the Table 2. Under the Markov-match, the hypothesis that $\gamma = 1$ cannot be rejected ($p = 0.751$), the Poisson-match is shown to need $\gamma > 1$, that is, the win probabilities it forecasts are too central ($p = 0.0171$), reinforcing that the Markov-match is a good fit but the Poisson-match is not.

5.2. Other remarks

The numbers of goals scored by the two teams are observed to be negatively correlated. Any ratio-of-strengths model will give such a result, because the numbers of goals are stochastic functions $f(\alpha_1/\alpha_2)$ and $f(\alpha_2/\alpha_1)$ respectively. For example, in the Poisson-match, mean numbers of goals are $\alpha_1/r\alpha_2$ and $\alpha_2/r\alpha_1$ respectively. A distribution of strengths α will then induce a negative correlation between the scores of the competitors. Note that the intraclass correlation must be used to measure this, as competitor labels of 1 or 2 are arbitrary in the absence of a home competitor. Further, it would be interesting to investigate this correlation directly in a more balanced tournament, such as the European Handball Federation men’s Champions League (see Csató, 2020).

The two models yield different values of team strength, which are highly correlated (correlation coefficient 0.992). There is no reason why the two sets of strengths should be identical.

Ties at the end of the second half of play are common (nearly 6% of games end thus). This is similar to the figures reported in the simulation study in Table 1.

Note that the prior and qualification terms satisfy an invariance property that the full likelihood must have: if a different competitor is chosen to have $\alpha = 1$ the likelihood describing results of games does not change. The qualification term has the same property, because if $\alpha \rightarrow c\alpha$, then $\mu \rightarrow \mu + \ln(c)$, and so $\log(\alpha) - \mu$ does not change.

6. Conclusion

We present a unified suite of models, a theory, for the bivariate scores in possessive ball-sports. One is old, the Poisson-match, one is recent, the binomial-match, and one is new, the Markov-match, which is described in detail in this paper. These models are appropriate in the cases when: goals are rare or the restart is contested; when goals are frequent; and for the latter when the restart alternates or when it favours the conceder (catch-up rule). These cases include all the well-known possessive ball-sports. Therefore, we claim these three models provide the fundamental basis for all useful models of possessive ball-sports.

The paper describes in detail the Markov-match, wherein: when competitor 1 plays (takes the restart) they score with probability q_1 and concede with probability $1-q_1$; when competitor 2 plays they score with probability q_2 and concede with probability $1-q_2$; and the conceding competitor restarts. We describe how to compute the probability distribution of the final score (X_1, X_2), how to parameterise the model, and how to estimate the parameters. The model is illustrated using handball, a high-scoring sport. We show that the Markov-match is a better fit to the handball results than the Poisson-match. The Markov-match is compared to the binomial match, particularly in terms of the dependence that the models induce between X_1 and X_2. Correlation between the scores is generally higher in the Markov-match.

An interesting problem of selection bias in tournament scores is tackled. In this problem, competitors have fewer observations (fewer matches) in a typical dataset because they are weaker. In our solution, strength estimates of competitors with fewer results are shrunk below the overall mean strength rather towards it, as is customary in standard shrinkage. Also, we present a new solution to the problem of estimating the rate at which the results of past matches are discounted. The test of goodness of fit for the models is also new, as are the computations of the likelihood and win probability (exact and approximate).

The models are important because they form the basis of models that are used for prediction (e.g. in prediction “competitions”), experimentation (e.g. testing new formats); and understanding (e.g. rating/ranking competitors). They provide a suite of models from which to choose the “best” model, although the Markov-match and the binomial-match are complementary, each competing in its appropriate setting with the Poisson-match.

In future work, we will look at basketball, water polo, and European handball. Basketball and water polo use the catch-up rule. Basketball has multiple scoring modes. In water polo, the scoring-rate is moderate, and it will be interesting to see which model, the Poisson-match or the Markov-match, is preferred. European teams dominate men’s handball (the highest placed non-European team in Table 3, Appendix 3, is Tunisia in 20th), so that the tournaments we have analysed are rather unbalanced, although our analysis adjusts for this. It would be interesting therefore to study the performance of the Markov-match in the more balanced biennial European Men’s Handball Championship. Modelling strength evolution in the manner of Bolsinova, Maris, Hofman, van der Maas and Brinkhuis (2022) would also be interesting, but this would require a large number matches to obtain meaningful results.

Appendix 1. Asymptotic approximation for the win probability under the Markov-match

Since the numbers of goals are large, renewal theory can give a good approximation to the probability that team 1 wins. The ‘times’ (numbers of goals from either team) at which team 1 (say) restarts with possession form a discrete renewal process (Noortwijk & van der Weide, 2006), with probabilities P_j of length j given by $P_1 = 1 - q_1$, $P_{k>1} = q_1(1-q_2)^{k-2}q_2$, so that team 1 either fail to score and regain possession immediately, or scores and then team 2 fails to score $k-2$ times and then team 2 scores, returning possession to team 1. This random variable has a distribution that is a mixture of 1, that is, the distribution that assigns probability 1 to the value 1, and a shifted geometric distribution. Each renewal counts 1 goal for team 2. The pgf (probability generating function) is

$$G(z) = \frac{(1-q_1)z + q_1 q_2 z^2}{1-(1-q_2)z} \tag{A1}$$

because the pgf of the geometric distribution starting at 2 is

$$\frac{q_2 z^2}{1-(1-q_2)z}.$$

Let T_2 be the time (number of goals by either side) for team 2 to score when team 1 is initially in possession. From (A1), the mean time to score for team 2 when team 1 is in possession is $E(T_2) \equiv k_2 = 1 + q_1/q_2$ and the $\mathrm{var}(T_2) \equiv \nu_2 = (2 - q_1 - q_2)q_1/q_2^2$. Since each renewal is a goal for team 2, the number of goals scored by team 2 is the renewal function at time N (the total number of goals).

There are two complications. One is that N is a random variable. The Poisson distribution is the obvious choice, but in fact we only need to specify the mean E(N) and variance var(N). The other complication is that team 1 do not always start in possession.

Noortwijk and van der Weide (2006) show that the mean number of renewals (team 1 goals) is such that $E(X_1) \to (N+1/2)/k_1 + \nu_1/2k_1^2 - 1/2$ as $N \to \infty$ when team 2 starts in possession. They also show that the variance of the number of renewals $\phi^2 \equiv N\mathrm{var}(T)/E(T)^3$, and the number of renewals X_1 is asymptotically normal. Here, the total number of goals N is a random variable.

To find the probability that team 1 wins, we study the goal difference D, where when team 1 starts with possession

$$D = X_1 - X_2 = 2X_1 - N \approx \left(1 - \frac{2}{k_2}\right)E(N) - \frac{1}{k_2} - \frac{\nu_2}{k_2^2} + 1.$$

When team 2 starts with possession, by symmetry

$$-D = X_2 - X_1 = 2X_2 - N \approx \left(1 - \frac{2}{k_1}\right)E(N) - \frac{1}{k_1} - \frac{\nu_1}{k_1^2} + 1.$$

Averaging over these expressions for D, when the game starts with a coin toss

$$E(D) = X_1 - X_2 \approx \left(E(N) + \frac{1}{2}\right)\left(\frac{1}{k_1} - \frac{1}{k_2}\right) + \frac{(k_1 - k_2)\nu_1}{2k_1^3}.$$

Note that the leading term in D is $N(q_1 - q_2)/(q_1 + q_2)$. That is, it is as if we have a series of Bernoulli trials where team 1 wins with probability $q_1/(q_1 + q_2)$. The formula for E(D) uses the fact that $\nu_1/k_1^3 = \nu_2/k_2^3$. Hence the formula is symmetric between team 1 and 2 as it must be. The asymptotic variance is

$$\mathrm{var}(D) = \left(\frac{1}{k_1} - \frac{1}{k_2}\right)^2 \mathrm{var}(N) + 4\left(\frac{\nu_1}{k_1^3}\right)E(N).$$

The team 1 winning probability approximation is $P = \Phi\{E(D)/\sqrt{\mathrm{var}(D)}\}$.

In the dataset, the mean average error (MAE) on P was 0.00109, using the Poisson approximation with Poisson mean E(N)=var(N) $\simeq$ 56.

By regarding a score in the range (-½, ½) as a draw, counting as half a win, the probability of winning becomes

$$P = \frac{1}{2}\left[\Phi\left\{\frac{E(D-1/2)}{\sqrt{\mathrm{var}(D)}}\right\} + \Phi\left\{\frac{E(D+1/2)}{\sqrt{\mathrm{var}(D)}}\right\}\right],$$

which has a MAE of 0.000911.

Appendix 2: Correlation in the binomial-match

For brevity, we consider only the case $q_1 = q_2 = q$. First, assume $N \sim \text{Poisson}(\lambda)$. Then, conditional on $N = n$, let $Y_1 \sim N(nq/2, nq(1-q)/2)$, and $Y_1 \sim N(nq/2, nq(1-p)/2)$, independently, and let $X_1 = Y_1 + n/2 - Y_2$ and $X_2 = Y_2 + n/2 - Y_1$. This approximates the *true* binomial-match, and avoids the technicalities that arise because N can be odd.

Now, $E(X_1) = E(X_2) = n/2$.

Also, $\text{var}(X_1|N=n) = \text{var}(Y_1) + \text{var}(Y_2) = nq(1-q)$, because Y_1 and Y_2 are independent, so that

$$E({X_1}^2|N=n) = \text{var}(X_1|N=n) + \{E(X_1|N=n)\}^2 = nq(1-q) + n^2/4.$$

Therefore,

$$E({X_1}^2) = \lambda q(1-q) + E(N^2)/4 = \lambda q(1-q) + (\lambda + \lambda^2)/4,$$

and thus $\text{var}(X_1) = \lambda q(1-q) + (\lambda + \lambda^2)/4 - \lambda^2/4 = \lambda(1 + 4q - 4q^2)/4$. Notice that when $q = 1/2$, $\text{var}(X_1) = \lambda/2 = E(X_1)$, so that X_1 has the same coefficient of variation as a Poisson$(\lambda/2)$ random variable.

Now,

$$\begin{aligned} E(X_1X_2|N=n) &= E\{(Y_1 + n/2 - Y_2)(Y_2 + n/2 - Y_1)\} \\ &= 2E(Y_1Y_2) - E({Y_1}^2) - E({Y_2}^2) + n^2/4 \\ &= -nq(1-q) + n^2/4, \end{aligned}$$

so that $E(X_1X_2) = -\lambda q(1-q) + (\lambda + \lambda^2)/4$, and $\text{cov}(X_1, X_2) = -\lambda q(1-q) + \lambda/4$, and

$$\text{corr}(X_1, X_2) = \frac{1 - 4q + 4q^2}{1 + 4q - 4q^2}.$$

When $q = 1/2$, $\text{corr}(X_1, X_2) = 0$, and when $q = 1$, $\text{corr}(X_1, X_2) = 1$. This is as an addendum to Baker et al. (2021), in which the binomial-match is defined in a subtly different way. Therein, each competitor has $N \sim \text{Poisson}(\lambda/2)$ restarts, so that the total number of restarts has mean λ and variance 2λ. This is different to the above.

Appendix 3

Table 3
Summary of performance of national teams in the biennial World Men's Handball Championship 2007–2021.

Team	matches	appearances	winner	runner-up	semi-final	mean score
Denmark	73	8	2	2	6	31.30
Spain	72	8	1	0	4	31.06
Norway	50	6	0	2	2	30.58
Romania	16	2	0	0	0	30.19
Slovenia	41	5	0	0	2	30.05
Sweden	50	6	0	1	2	29.76
France	74	8	4	0	7	29.66
Czech Republic	15	2	0	0	0	29.13
Croatia	71	8	0	1	3	29.10
Hungary	55	7	0	0	0	28.80
Russia	52	7	0	0	0	28.69
Slovakia	16	2	0	0	0	28.63
Iceland	51	7	0	0	0	28.61
Germany	65	8	1	0	2	28.58
Austria	27	4	0	0	0	28.56
Serbia	31	4	0	0	0	28.13
Belarus	25	4	0	0	0	28.04
Portugal	6	1	0	0	0	28.00
Ukraine	4	1	0	0	0	28.00
Tunisia	57	8	0	0	0	27.72
Poland	57	7	0	1	3	27.56
North Macedonia	40	6	0	0	0	27.35
South Korea	42	6	0	0	0	27.26
Cape Verde	1	1	0	0	0	27.00
Qatar	43	6	0	1	1	26.84
Brazil	54	8	0	0	0	26.76
Japan	27	4	0	0	0	26.74
Greenland	6	1	0	0	0	26.67
Chile	42	6	0	0	0	26.52
Iran	7	1	0	0	0	26.43
Egypt	56	8	0	0	0	26.34
Angola	26	4	0	0	0	25.81
Montenegro	7	1	0	0	0	25.57
Switzerland	6	1	0	0	0	25.50
Congo	6	1	0	0	0	25.33
Morocco	13	2	0	0	0	25.31
Bosnia and Herzegovina	7	1	0	0	0	25.00
Argentina	55	8	0	0	0	24.40
Algeria	36	5	0	0	0	24.14
Bahrain	27	4	0	0	0	23.78
Kuwait	13	2	0	0	0	23.77
Cuba	9	1	0	0	0	23.44
Saudi Arabia	37	5	0	0	0	22.41
Australia	29	4	0	0	0	16.83
Uruguay	5	1	0	0	0	17.60

Appendix 4

Table 4
Men's Handball World Championship Final 2021, Denmark (D) vs Sweden (S), score 26–24. Denmark had 25 plays (restarts); Sweden 27. Sweden had first play. Two plays ended at the buzzer. $q_{\text{DEN}} = 0.80$, $q_{\text{SWE}} = 0.74$. Plays with one possession: 29/52. (We collected these data directly from match video.).

first half						second half					
play	score		Play	score		play	score		play	score	
S	0	1	SDS	8	7	D	14	13	D	21	20
D	1	1	DS	8	8	S	14	14	SD	22	20
SDS	1	2	DSD	9	8	D	15	14	SD	23	20
D	2	2	S	9	9	SDS	15	15	S	23	21
S	2	3	D	10	9	D	16	15	DSD	24	21
D	3	3	S	10	10	S	16	16	S	24	22
SDS	3	4	DS	10	11	DS	16	17	D	25	22
D	4	4	DS	10	12	D	17	17	SDSDS	25	23
SD	5	4	D	11	12	SDS	17	18	DSD	26	23
SDS	5	5	S	11	13	D	18	18	SDS	26	24
D	6	5	DSD	12	13	S	18	19	D		
S	6	6	SD	13	13	D	19	19			
DSDSD	7	6	S			SD	20	19			
SDSD	8	6				S	20	20			

References

Anbarcı, N., Sun, C.-. J., & Unver, M. U. (2021). Designing practical and fair sequential team contests: The case of penalty shootouts. *Games and Economic Behavior, 130*, 25–43.

Azzalini, A., & Capitanio, A. (2014). *The skew-normal and related families*. Cambridge, UK: Cambridge University Press.

Baker, R. (2020a). New order-statistics-based ranking models and faster computation of outcome probabilities. *IMA Journal of Management Mathematics, 31*, 33–48.

Baker, R (2020b). Discrete distributions from a Markov chain. *arXiv* 2006.13766.

Baker, R., & McHale, I. (2013). Forecasting exact scores in National Football League games. *International Journal of Forecasting, 29*, 122–130.

Baker, R., & Scarf, P. A. (2021). Modifying Bradley–Terry and other ranking models to allow ties. *IMA Journal of Management Mathematics, 32*, 451–463.

Baker, R. D., Chadwick, S., Parma, R., & Scarf, P. A. (2021). The binomial-match, outcome uncertainty, and the case of netball. *Journal of the Operational Research Society*. https://doi.org/10.1080/01605682.2021.1931496.

Bilge, M. (2012). Game analysis of Olympic, World and European Championships in men's handball. *Journal of Human Kinetics, 35*, 109–118.

Bolsinova, M., Maris, G., Hofman, A. D., van der Maas, H. L., & Brinkhuis, M. J. (2022). Urnings: A new method for tracking dynamically changing parameters in paired comparison systems. *Journal of the Royal Statistical Society Series C, 71*, 91–118.

Brams, S., Ismail, M., Kilgour, D., & Stromquist, W. (2018). Catch-up: A rule that makes service sports more competitive. *The American Mathematical Monthly, 125*, 771–796.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review, 78*, 1–3.

Crowder, M., Dixon, M. J., Ledford, L., & Robinson, M. (2002). Dynamic modelling and prediction of English Football League matches for betting. *The Statistician, 51*, 157–168.

Csató, L. (2020). Optimal tournament design: Lessons from the men's handball Champions League. *Journal of Sports Economics, 21*, 848–868.

Csató, L. (2021a). A comparison of penalty shootout designs in soccer. *4OR, 19*, 183–198.

Csató, L. (2021b). *Tournament design: How operations research can improve sports rules*. Cham, Switzerland: Palgrave Macmillan.

Csató, L. (2021c). A simulation comparison of tournament designs for the World Men's Handball Championships. *International Transactions in Operational Research, 28*, 2377–2401.

Csató, L. (2022). How to design a multi-stage tournament when some results are carried over? *OR spectrum* in press. https://doi.org/10.1007/s00291-022-00671-2.

Csató, L., & Petróczy, D. G. (2022). Fairness in penalty shootouts. Is it worth using dynamic sequences? *arXiv* 2004.09225.

Dekking, M., & Kong, D. (2011). Multimodality of the Markov binomial distribution. *Journal of Applied Probability, 48*, 938–953.

Dewart, N., & Gillard, J. (2019). Using Bradley–Terry models to analyse test match cricket. *IMA Journal of Management Mathematics, 30*, 187–207.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics, 46*, 265–280.

Dumangane, M., Rosati, N., & Volossovitch, A. (2009). Departure from independence and stationarity in a handball match. *Journal of Applied Statistics, 36*, 723–741.

Forrest, D., & McHale, I. G. (2019). Using statistics to detect match fixing in sport. *IMA Journal of Management Mathematics, 30*, 431–449.

Groll, A., Heiner, J., Schauberger, G., & Uhrmeister, J. (2020). Prediction of the 2019 IHF World Men's Handball Championship –a sparse Gaussian approximation model. *Journal of Sports Analytics, 6*, 187–197.

Heckman, J. (1977). Sample selection bias as a specification error. *Econometrica : journal of the Econometric Society, 47*, 153–162.

Heuer, A., Müller, C., & Rubner, O. (2010). Soccer: Is scoring goals a predictable Poissonian process? *Europhysics Letters, 89*, 38007.

Hirotsu, N. (2022). Soccer as a Markov process: Modelling and estimation of the zonal variation of team strengths. *IMA Journal of Management Mathematics*. https://doi.org/10.1093/imaman/dpab042.

Hirotsu, N., & Bickel, J. E. (2016). Optimal batting orders in run-limit-rule baseball: A Markov chain approach. *IMA Journal of Management Mathematics, 27*, 297–313.

Hosmer, D. W., J. r, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). New York: Wiley.

Hubáček, O., Šourek, G., & Železný, F. (2022). Forty years of score-based soccer match outcome rediction: An experimental review. *IMA Journal of Management Mathematics, 33*, 1–18.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society Series D, 52*, 381–393.

Karlis, D., & Ntzoufras, I. (2011). Robust fitting of football prediction models. *IMA Journal of Management Mathematics, 22*, 171–182.

Kendall, G., & Lenten, L. J. A. (2017). When sports rules go awry. *European Journal of Operational Research, 257*, 377–394.

Klaassen, F. J. G. M., & Magnus, J. R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association, 96*, 500–509.

Klaassen, F. J. G. M., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research, 148*, 257–267.

Koning, R. H. (2000). Balance in competition in Dutch soccer. *Journal of the Royal Statistical Society Series D, 49*, 419–431.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society Series A, 178*, 167–186.

Kvam, P., & Sokol, J. (2006). A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics, 53*, 788–803.

Lenten, L. J. A., & Kendall, G. (2021). Scholarly sports: Influence of social science academe on sports rules and policy. *Journal of the Operational Research Society* in press. https://doi.org/10.1080/01605682.2021.2000896.

Manasis, V., Ntzoufras, I., & Reade, J. J. (2022). Competitive balance measures and the uncertainty of outcome hypothesis in European football. *IMA Journal of Management Mathematics, 33*, 19–52.

Martín-González, J. M., de Saá Guerra, Y., García-Manso, J. M., Arriaza, E., & Valverde-Estévez, T. (2016). The Poisson model limits in NBA basketball: Complexity in team sports. *Physica A: Statistical Mechanics and its Applications, 464*, 182–190.

Meletakos, P., & Bayios, I. (2010). General trends in European men's handball: A longitudinal study. *International Journal of Performance Analysis in Sport, 10*, 221–228.

Merritt, S., & Clauset, A. (2014). Scoring dynamics across professional team sports: Tempo, balance and predictability. *EPJ Data Science, 3*. https://doi.org/10.1140/epjds29.

Noortwijk, J. M., & van der Weide (2006). Computational techniques for discrete–time renewal processes. In *Safety and Reliability for Managing Risk, vols 1-3. Book Series: Proceedings and Monographs in Engineering Water and Earth Sciences Pages* (pp. 571–578).

Ötting, M. (2021). Predicting play calls in the National Football League using hidden Markov models. *IMA Journal of Management Mathematics, 32*, 535–545.

Ruiz, F., & Perez-Cruz, F. (2015). A generative model for predicting outcomes in college basketball. *Journal of Quantitative Analysis in Sports, 11*, 39–52.

Sandholtz, N., & Bornn, L. (2020). Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball. *The Annals of Applied Statistics, 14*, 1122–1145.

Scarf, P. A., Khare, A., & Alotaibi, N. (2022). On skill and chance in sport. *IMA Journal of Management Mathematics, 33*, 53–73.

Scarf, P. A., Parma, R., & McHale, I. (2019). On outcome uncertainty and scoring rates in sport: The case of international rugby union. *European Journal of Operational Research, 273*, 721–730.

Scarf, P. A., Yusof, M. M., & Bilbao, M. (2009). A numerical study of designs for sporting contests. *European Journal of Operational Research, 198*, 190–198.

Sim, M. K., & Choi, D. G. (2020). The winning probability of a game and the importance of points in tennis matches. *Research Quarterly for Exercise and Sport, 91*, 361–372.

Song, K., & Shi, J. (2020). A gamma process based in-play prediction model for National Basketball Association games. *European Journal of Operational Research, 283*, 706–713.

Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting, 28*, 532–542.

Szymanski, S. (2003). The economic design of sporting contests. *Journal of Economic Literature, 41*, 1137–1187.

Uhrín, M., Šourek, G., Hubáček, O., & Železný, F. (2021). Optimal sports betting strategies in practice: An experimental review. *IMA Journal of Management Mathematics, 32*, 465–489.

Vilkkumaa, E., & Liesiö, J. (2022). What causes post-decision disappointment? Estimating the contributions of systematic and selection biases. *European Journal of Operational Research, 296*, 587–600.

Viveros, R., Balasubramanian, K., & Balakrishnan, N. (1994). Binomial and negative binomial analogues under correlated Bernoulli trials. *The American Statistician, 48*, 243–247.

Wolfers, J. (2006). Point shaving: Corruption in NCAA basketball. *American Economic Review, 96*, 279–283.

Wright, M. B. (2014). OR analysis of sporting rules–a survey. *European Journal of Operational Research, 232*, 1–8.