THE UNIVERSITY OF

WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

This paper is made available online in accordance with publisher policies. Please scroll down to view the document itself. Please refer to the repository record for this item and our policy information available from the repository home page for further information.

To see the final version of this paper please visit the publisher's website. Access to the published version may require a subscription.

Author(s): M Cattelan, C Varin and D Firth

Article Title: Stochastic dynamic Thurstone-Mosteller models for sports tournaments
Year of publication: 2010
Link to published article:
http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2010/paper 10-19

Publisher statement: None

Stochastic dynamic Thurstone-Mosteller models for sports tournaments

Manuela Cattelan

University of Padua, Italy

Cristiano Varin

Ca' Foscari University, Venice, Italy

David Firth

University of Warwick, UK

September 23, 2010

SUMMARY

In the course of national sports tournaments, usually lasting several months, it is expected that the abilities of teams taking part in the tournament change in time. A dynamic extension of the Thurstone-Mosteller model for paired comparison data is introduced to model the outcomes of sporting contests allowing for time-varying abilities. It is assumed that the development of teams' abilities follows a stationary process and a team-specific home effect is considered. The likelihood function of the proposed model requires the approximation of a high dimensional integral. This difficulty is overcome by means of maximum simulated likelihood via the Geweke-Hajivassiliou-Keane algorithm. Ranking of teams and forecasting future match results are performed through a Metropolis-Hastings algorithm. The methodology is applied to sports data with and without tied contests, namely the 2006-2007 Italian volleyball league and the 2008-2009 Italian Serie A football season.

1

1 Introduction

The analysis of sports data has always aroused great interest among statisticians. Albert *et al.* (2005) collect a number of articles that summarise various statistical aspects of interest in sports data including rating of players or teams, evaluation of sport strategies, enhancement of sport rules, illustration of statistical methods and forecasting of results.

Sports data have been investigated from different perspectives, often with the aim of forecasting the results. A first approach consists in modelling the number of goals of the two teams. Maher (1982) employs independent Poisson distributions for the number of scores of each team with means that depend on the attack and defence strength of teams. Dixon and Coles (1997) propose an *ad hoc* adjustment of the Poisson distribution introducing a dependence parameter that modifies the probabilities of the results 0-0, 0-1, 1-0 and 1-1. Dixon and Coles (1997) introduce also a dynamic element in the model updating the parameter estimates including the results up to the last observation and down-weighting observations distant in time. Karlis and Ntzoufras (2003) suggest to apply a bivariate Poisson distribution with a dependence parameter between the number of goals scored by the two teams and then extend the model to inflate the probabilities of draws.

McHale and Scarf (2007) model the number of shots of the two teams. They propose two different types of Archimedean copula with either Poisson or negative binomial distributions for the marginals to account for the negative dependence between shots-for and shots-against.

Extensions allowing dynamic developments of abilities of the teams are proposed by Rue and Salvesen (2000) and Crowder *et al.* (2002). Rue and Salvesen (2000) assume that the attack and defence strength parameters of each team follow a Brownian motion process.

2

The model is estimated by employing Bayesian inference through Markov chain Monte Carlo methods. Crowder *et al.* (2002) suggest an autoregressive model for the attack and defence abilities of teams. The original model is then replaced by a derived version that is easier to handle by maximum likelihood.

A second approach to the analysis of sports data consists in modelling the difference in scores. Clarke and Norman (1995) perform a linear regression of the difference in scores on the difference in strength of the two teams. Harville (2003) employs a similar specification, but eliminates the incentives for running up the score beyond a predetermined number of points. A dynamic specification of strength in this context is considered in Harville (1980) who proposes an autoregressive process for the strength of teams in different seasons. Also Glickman and Stern (1998) assume that the evolution of week-by-week and seasonal strength follows a first-order autoregressive process. Inference is carried out in a Bayesian framework through Markov chain Monte Carlo algorithms.

Finally, sports data can be analysed by considering only the outcomes of the matches (win-draw-loss). Goddard and Asimakopoulos (2004) use an ordered probit model to determine which covariates, e.g. importance of the match, fouls, yellow and red cards, affect the result of the match. An ordered probit model is adopted also by Koning (2000) who specifies the probability of the outcome as a function of the difference of abilities of the two teams. Kuk (1995) introduces two strength parameters for each team, one denoting the strength when playing at home and the other when playing away.

Barry and Hartigan (1993) propose a dynamic extension for the ability parameters of teams; they employ a choice model assuming a prior distribution for strength of teams that changes slowly in time. Fahrmeir and Tutz (1994) consider three possible specifications for the development of abilities: a first and second order random walk and a local linear trend model. These models are estimated using empirical Bayes methods. Glickman (1999) specifies a logit model assuming a prior with normal increments for abilities of teams and proposes an approximate Bayesian algorithm for ranking purposes. Knorr-Held (2000) employs a logit

model assuming random walk priors for abilities of teams. The variance of the random walk is estimated through four different predictive criteria while the abilities are estimated by means of the extended Kalman filter and smoother.

In this paper, we analyse the results of double round-robin tournaments from the last perspective, that is modelling the outcomes of matches. Since we are interested in studying how the strengths of the teams evolve during the season, we develop a stochastic dynamic paired comparison model. Differently from previous works by Fahrmeir and Tutz (1994), Glickman (1999) and Knorr-Held (2000), we specify a stationary time series model for describing the evolution in time of the ability of each team and estimate model parameters by simulated maximum likelihood.

The paper is organised as follows. Section 2 presents two motivating data sets regarding Italian men's volleyball and football major leagues. Exploratory analyses are carried out to give guidance on model construction. Section 3.1 describes a dynamic version of the Thurstone-Mosteller model suggested by the exploratory analyses. Section 3.2 discusses maximum simulated likelihood estimation of the proposed model while in Section 3.3 a Metropolis-Hastings algorithm is developed for the estimation of the model components and for the prediction of match results. The methodology is applied in Section 4 to the data for the two sports. Some concluding remarks are given in Section 5.

2   Description of the data and first analyses

2.1   Volleyball

As motivating examples we consider data from two major Italian double round-robin tournaments. The first one is the 2006-2007 regular season of the men's Italian A1 volleyball league. There are fourteen teams in the league competing in a double round-robin tournament starting in September 2006 and ending in April 2007. Match results are available from the website `http://www.legavolley.it`. Table 1 orders the teams by the final points

4

Table 1: 2006-2007 regular season men's Italian A1 volleyball league. The table displays: (1) points (`pts`), (2) percentage of home points (`% home`), (3) estimated abilities (`ability`), (4) quasi standard errors (`qse`) and (5) ranks (`rank`) based on the `static` Thurstone-Mosteller model, (6) estimated mean abilities, (7) quasi standard errors and (8) ranks based on the `dynamic` Thurstone-Mosteller model.

	pts	% home	static			dynamic		
			ability	qse	rank	ability	qse	rank
1) Cuneo	57	0.60	0.876	0.296	1	0.716	0.257	1
2) Roma	56	0.61	0.726	0.285	2	0.603	0.251	2
3) Treviso	50	0.68	0.486	0.272	3	0.437	0.250	3
4) Piacenza	49	0.59	0.470	0.271	4	0.365	0.243	4
5) Modena	45	0.64	0.244	0.264	5	0.175	0.243	6
6) Perugia	43	0.54	0.226	0.264	6	0.178	0.239	5
7) Taranto	43	0.67	-0.001	0.262	8	0.012	0.241	8
8) Trentino	41	0.51	-0.090	0.262	9	-0.068	0.229	9
9) Montichiari	40	0.65	0.124	0.262	7	0.084	0.239	7
10) Macerata	34	0.79	-0.106	0.262	10	-0.094	0.248	10
11) Latina	24	0.79	-0.583	0.278	11	-0.481	0.252	11
12) Padova	23	0.78	-0.874	0.298	14	-0.713	0.259	14
13) Vibo Valentia	22	0.68	-0.627	0.280	12	-0.516	0.254	12
14) Verona	19	0.95	-0.871	0.297	13	-0.691	0.259	13

ranking with Cuneo classified first and Verona last at the end of the regular season. The points are assigned as follows: if a match ends 3-0 or 3-1 the winning team gains 3 points and the losing team remains empty handed, while if the match ends 3-2, the winning team gains 2 points and the losing team is rewarded with 1 point.

Let Y_{ij} be the binary random variable which denotes whether the match ended in a victory ($Y_{ij} = 1$) or a loss ($Y_{ij} = 2$) for the home team i against the away team j, with $i, j = 1, \ldots, n$, $i \neq j$. In the specific case of the volleyball tournament there are $n = 14$ teams. Traditional pair comparison models describe the outcome probability as $\mathrm{pr}(Y_{ij} = 1) = F(a_i - a_j)$, where F is a distribution function and a_i is a parameter measuring the ability of team i. This simple choice model is commonly termed the Bradley-Terry model (Bradley and Terry, 1952) or the Thurstone-Mosteller model (Thurstone, 1927; Mosteller, 1951) depending on whether F is the distribution function of a logistic or of a standard normal random variable, respectively.

The second column of Table 1 displays the percentage of points obtained by each team

when playing at home. Between 51% and 95% of the total points gained by a team are collected in home matches. On average, 68% of the points are acquired in home games. The advantage deriving from playing at home is commonly taken into account by including a common home effect parameter h for all teams (Fahrmeir and Tutz, 1994; Knorr-Held, 2000; Harville, 2003), thus leading to the model

$$\text{pr}(Y_{ij} = 1) = F(h + a_i - a_j). \tag{1}$$

Model identifiability requires one constraint in the set of abilities, such as the sum constraint $\sum_{i=1}^n a_i = 0$ or the reference team constraint $a_i = 0$ for some $i \in \{1, \ldots, 14\}$. Column three of Table 1 shows estimates of the abilities $\hat{a}_i$ from the above model using a probit link and the sum constraint for consistency with the analyses later conducted in this paper. The estimated home effect parameter $\hat{h}$ is 0.562 with standard error 0.112, thus confirming the relevant advantage for home teams. Indeed, the model-based estimated probability of a victory for the home team in a match between two teams with the same ability is $\Phi(0.562) = 0.71$, where $\Phi(x)$ denotes the cumulative distribution function of standard normal variable computed at point x.

Column four of Table 1 displays the quasi standard errors (Firth and de Menezes, 2004) of the estimated abilities. These quasi standard errors allow to approximately reconstruct the uncertainty of pairwise differences $\hat{a}_i - \hat{a}_j$ used for comparing teams without the need to report also the covariance between $\hat{a}_i$ and $\hat{a}_j$. The estimated abilities suggest classifying the teams in three groups: a first class group including the best four teams in the final ranking, a second group of teams with an average ability and a final group of four teams that are weaker than the others.

The ranking derived from the estimated abilities closely agrees with the actual final ranking, with a Kendall τ rank correlation index of 0.906. The minor differences between the two rankings are due to the different classification philosophies. For example, Montichiari is ranked 7th by the estimated abilities while it is 9th in the final ranking because in a

relatively large number of matches it won 3-2. Such a result is considered a win in the Thurstone-Mosteller model, but the points gained by the winning team are fewer than if it won with a larger margin (3-1 or 3-0).

2.2 Association football

The second application regards the 2008-2009 Italian Serie A football league. This tournament comprises twenty teams with matches played between August 2008 and May 2009. The teams ranked according to the final points order are listed in Table 2. In the football tournament, the winning team gains 3 points while the losing team gets nothing. If the match is tied, both teams gain 1 point. On average, 65% of the total points are gained in home matches, with percentages ranging from 45% to 79%.

In contrast to volleyball, football matches can also end in a tie, hence random variable Y_{ij} has three categories that we arbitrarily code as follows: 1 if the home team wins, 2 for a tie and 3 for a victory of the guest team. Model (1) is extended to account for ties with a cumulative link specification

$$\mathrm{pr}(Y_{ij} \leq y_{ij}) = F(\delta_{y_{ij}} + h + a_i - a_j), \tag{2}$$

where $-\infty = \delta_0 < \delta_1 < \delta_2 < \delta_3 = \infty$ are cutpoint parameters. Model identifiability requires a further constraint in the cutpoints or to fix the value of the home effect parameter. Here, we prefer the first option and we assume $\delta_1 = -\delta_2$ so as to preserve the model symmetry.

The third column of Table 2 reports the estimates of the abilities which range from -0.505 for Reggina to 0.840 for Internazionale. Again, the ranking derived from the estimated abilities is very similar to the final points ranking, as the Kendall τ rank correlation is 0.95. The estimated home effect parameter is $\hat{h} = 0.396$ with standard error 0.062.

7

Table 2: 2008-2009 Italian Serie A football league. The table displays: (1) points (`pts`), (2) percentage of home points (`% home`), (3) estimated abilities (`ability`), (4) quasi standard errors (`qse`) and (5) ranks (`rank`) based on the `static` Thurstone-Mosteller model, (6) estimated mean abilities, (7) quasi standard errors and (8) ranks based on the `dynamic` Thurstone-Mosteller model.

	pts	% home	static			dynamic		
			ability	qse	rank	ability	qse	rank
1) Internazionale	84	0.56	0.840	0.203	1	0.622	0.171	1
2) Juventus	74	0.53	0.546	0.189	2	0.405	0.160	3
3) Milan	74	0.61	0.545	0.195	3	0.412	0.164	2
4) Fiorentina	68	0.65	0.366	0.196	5	0.279	0.169	5
5) Genoa	68	0.60	0.413	0.188	4	0.326	0.161	4
6) Roma	63	0.68	0.278	0.190	6	0.211	0.164	6
7) Udinese	58	0.66	0.128	0.186	7	0.084	0.160	7
8) Palermo	57	0.75	0.094	0.190	8	0.064	0.168	8
9) Cagliari	53	0.70	0.003	0.187	9	0.028	0.166	9
10) Lazio	50	0.56	-0.133	0.190	11	-0.102	0.166	11
11) Atalanta	47	0.70	-0.150	0.187	12	-0.117	0.164	12
12) Napoli	46	0.76	-0.160	0.187	13	-0.131	0.172	13
13) Sampdoria	46	0.70	-0.110	0.185	10	-0.088	0.159	10
14) Siena	44	0.73	-0.263	0.191	15	-0.194	0.166	14
15) Catania	43	0.79	-0.258	0.190	14	-0.198	0.168	15
16) Chievo	38	0.45	-0.298	0.181	16	-0.230	0.157	16
17) Bologna	37	0.57	-0.377	0.187	17	-0.290	0.161	17
18) Torino	34	0.74	-0.462	0.191	18	-0.355	0.167	18
19) Reggina	31	0.58	-0.505	0.188	20	-0.380	0.159	19
20) Lecce	30	0.63	-0.497	0.186	19	-0.382	0.161	20

2.3 Temporal development of team abilities

In the above static models, parameters a_i measure the average abilities of the teams over a complete season. However, team abilities are expected to change during the season because of injuries to players, tiredness due to participation also in international competitions, team psychology and other factors. Figure 1 shows time-varying estimates of abilities obtained by re-fitting models (1) and (2) using an increasing number of gameweeks for the second halves of the two tournaments. Overall, the team ability trajectories are characterized by a notable persistence through time, except for few trajectories exhibiting smooth drops or rises during specific parts of the season.
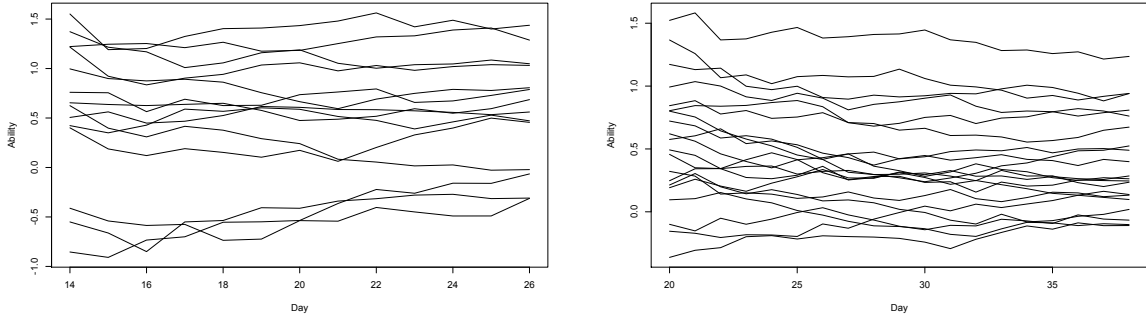
Figure 1: plot of estimated abilities for the second half of the season of the 2006-2007 men's Italian A1 volleyball league (left panel) and of the 2008-2009 Serie A football league (right panel).

3 Dynamic stochastic Thurstone-Mosteller model

3.1 The model

First analyses suggest to model the match results with the dynamic cumulative link model

$$\text{pr}(Y_{ij} \leq y_{ij} | A_{it_{ij}} = a_{it_{ij}}, A_{jt_{ij}} = a_{jt_{ij}}) = F(\delta_{y_{ij}} + h + a_{it_{ij}} - a_{jt_{ij}}),$$

where the random variable $A_{it_{ij}}$ describes the ability of team i in the home match against team j played at gameweek t_{ij} and parameter h counts for the home advantage.

Previous works by Fahrmeir and Tutz (1994), Glickman (1999) and Knorr-Held (2000) model the ability process A_{it} for a single team as random walks of first or second order. The key characteristic of a random walk process is the variance of A_{it} increasing with time without bound. However, the plots in Figure 1 seem not to support the volatility of a random walk process. In other words, we argue that the random-walk assumption of an heterogeneity among teams increasing with time is implausible. Hence, we specify the ability process A_{it} for a specific team i as a stationary Gaussian process with zero mean and exponential covariance function $\text{cov}(A_{it}, A_{it'}) = \sigma^2 \gamma^{|t-t'|}$, where $\sigma^2 > 0$ and $\gamma \in (0,1)$ are two unknown parameters that regulate the variance and the autocorrelation of the ability process, respectively.

With the normal assumptions for the team abilities A_{it}, it is convenient to assume a

9

probit link so that match results $\{Y_{ij} = y_{ij}\}$ can be written as censored normal variables $Z_{ij} \in (\delta_{y_{ij}-1}, \delta_{y_{ij}}]$ where $Z_{ij} = h + A_{it_{ij}} - A_{jt_{ij}} + \epsilon_{ij}$, with latent errors ϵ_{ij} distributed as independent standard normal variables. The resulting model represents a stochastic dynamic version of the Thurstone-Mosteller model.

In contrast to static paired comparison models, the proposed dynamic model induces a cross-dependence among match results; only the results of matches sharing a team are dependent. The degree of dependence between two matches involving the same team is inversely proportional to the temporal distance between the matches. Let $\boldsymbol{\theta} = (h, \delta_2, \sigma^2, \gamma)^{\mathrm{T}}$ be the whole parameter vector. If the sport does not allow for ties, $\delta_2 = 0$. The joint probability for two matches is

$$\mathrm{pr}(Y_{ij} = y_{ij}, Y_{kl} = y_{kl}; \boldsymbol{\theta}) = \int_{\delta_{y_{ij}-1}}^{\delta_{y_{ij}}} \int_{\delta_{y_{kl}-1}}^{\delta_{y_{kl}}} p(z_{ij}, z_{kl}; \boldsymbol{\theta}) \mathrm{d}z_{ij}\, \mathrm{d}z_{kl},$$

where $p(z_{ij}, z_{kl}; \boldsymbol{\theta})$ is the density of a bivariate normal variable with zero mean and covariance

$$\mathrm{cov}(z_{ij}, z_{kl}) = \begin{cases} \sigma^2 \gamma^{|t_{ij}-t_{kl}|} & \text{if } i = k, j \neq l \text{ or } i \neq k, j = l \\ 1 + 2\sigma^2 & \text{if } i = k, j = l, \\ -\sigma^2 \gamma^{|t_{ij}-t_{kl}|} & \text{if } i = l, j \neq k \text{ or } i \neq l, j = k, \\ -2\sigma^2 \gamma^{|t_{ij}-t_{kl}|} & \text{if } i = l, j = k, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Covariance $\mathrm{cov}(z_{ij}, z_{kl})$ is thus equal to zero unless one or both the teams are involved in the two matches. Note that this behaviour is qualitatively different from that of a model in which ability is assumed to change through an autoregressive process of order one. Indeed, as parameter γ approaches its limit value of one, the expression of $\mathrm{cov}(z_{ij}, z_{kl})$ converges to that corresponding to constant-in-time abilities, $A_{it} \equiv A_i$ for all times t, distributed as independent zero-mean normal variables with variance σ^2. With an autoregressive specification, in contrast, the identity $\gamma = 1$ would correspond to a non-stationary random walk model for

10

the time-varying abilities A_{it}.

Let $\boldsymbol{y}$ denote all match results. The likelihood function for $\boldsymbol{\theta}$ is given by the rectangular normal integral

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathcal{D}} p(\boldsymbol{z}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{z},$$

where the integral domain $\mathcal{D}$ is the cartesian product of the intervals $(\delta_{y_{ij}} - 1, \delta_{y_{ij}}]$ and $\boldsymbol{z}$ is the vector of elements z_{ij} for all terms in the set $\{i, j = 1, \ldots, n, i \neq j\}$. The likelihood integrand $p(\boldsymbol{z}; \boldsymbol{\theta})$ is thus a multivariate normal density of dimension $m = n(n-1)$, the number of matches played in the tournament, with zero mean and variance matrix with entries given by expression (3).

Likelihood evaluation is difficult because for each paired comparison there are two abilities correlated with the abilities relating to the same teams in different matches, so the likelihood cannot be split into low-dimensional integrals. The dimension of the likelihood integral is thus equal to the number of matches played in the tournament. The next section shows how the likelihood may be computed approximately by importance sampling.

3.2 Maximum simulated likelihood

The proposed paired comparison model is equivalent to a multivariate probit model whose likelihood is estimated by simulation via the Geweke-Hajivassiliou-Keane (GHK) algorithm (Train, 2003; Masarotto and Varin, 2010). This algorithm approximates the joint distribution of all the outcomes by sequential simulation from univariate truncated normal distributions.

Algorithm implementation requires the assumption of an order for the matches. We choose to arrange the matches in chronological order, with those played at the same time in alphabetic order of the home team. The GHK algorithm is a sequential importance sampling algorithm based on drawing from the conditional density $p(z_{ij}|y_{ij}, \mathcal{Z}_{ij}; \boldsymbol{\theta})$, where $\mathcal{Z}_{ij}$ is the vector of latent variables Z_{kl} preceding Z_{ij} in the chosen order. In other words, the GHK algorithm employs as importance density the normal density $p(z_{ij}|\mathcal{Z}_{ij}; \boldsymbol{\theta})$ truncated over the interval $(\delta_{y_{ij}-1}, \delta_{y_{ij}}]$.

11

Let m_{ij} and s_{ij} be the mean and the standard deviation of the conditional density $p(z_{ij}|\mathcal{Z}_{ij};\boldsymbol{\theta})$, respectively. Then, a draw from the importance density $p(z_{ij}|y_{ij},\mathcal{Z}_{ij};\boldsymbol{\theta})$ is obtained by setting

$$z_{ij}(u_{ij}) = m_{ij} + s_{ij}\Phi^{-1}\{(1-u_{ij})l_{ij} + u_{ij}r_{ij}\},$$

where u_{ij} is a draw from a random variable uniformly distributed in the unit interval, quantities l_{ij} and r_{ij} are defined as

$$l_{ij} = \Phi\left(\frac{-\delta_{y_{ij}-1} - m_{ij}\sqrt{1+2\sigma^2}}{s_{ij}\sqrt{1+2\sigma^2}}\right) \text{ and } r_{ij} = \Phi\left(\frac{-\delta_{y_{ij}} - m_{ij}\sqrt{1+2\sigma^2}}{s_{ij}\sqrt{1+2\sigma^2}}\right).$$

Denote by $z_{ij}^{(b)}$ the bth draw from the above importance density $(b = 1,\ldots,B)$. The GHK algorithm approximates the likelihood of the proposed paired comparison model by the Monte Carlo sum

$$\hat{\mathcal{L}}_{\text{GHK}}(\boldsymbol{\theta};\boldsymbol{y}) = \frac{1}{B}\sum_{b=1}^{B}\left\{\prod_{\text{ord}(i,j)}\frac{p(z_{ij}^{(b)}|\mathcal{Z}_{ij}^{(b)};\boldsymbol{\theta})}{p(z_{ij}^{(b)}|y_{ij},\mathcal{Z}_{ij}^{(b)};\boldsymbol{\theta})}\right\},$$

where the product follows the predetermined match order indicated by $\text{ord}(i,j)$. The GHK algorithm is popular in the econometric literature for approximate inference in multivariate probit models; more details and references can be found in Train (2003).

As with any importance sampling algorithm, the GHK algorithm provides an unbiased estimate of the likelihood function, but it is biased on the scale of the log-likelihood. Since it is more convenient to maximize the log-likelihood, it is opportune to correct the bias of its importance sampling approximation. For this purpose, we implement the correction suggested by Durbin and Koopman (1997).

Given the data dimensionality, some comments on computational aspects are worth noting here. The essential ingredient of the GHK algorithm is the Cholesky factor of the variance matrix of $\boldsymbol{Z}$ used for sampling sequentially from the conditional density $p(z_{ij}|y_{ij},\mathcal{Z}_{ij};\boldsymbol{\theta})$. Computation of the Cholesky factor requires $\mathcal{O}(m^3)$ computations, with m being the number of matches. In a complete season double round-robin tournament $m = \mathcal{O}(n^2)$, thus suggesting

12

that the algorithm may be of little utility except for relatively small tournaments. However, if necessary, relevant computational saving can be obtained by noticing that the variance matrix of $\boldsymbol{Z}$ is increasingly sparse as the number of teams increases because only pairs of matches with a team in common are correlated. In fact, while the variance matrix has dimension $m^2 = \{n(n-1)\}^2$, its non-zero cells number only $2n(n-1)(2n-3)$. For example, the percentage of zero cells in the variance matrix of $\boldsymbol{Z}$ is 73% for the volleyball data, 81% for the football data and it would be 96% in a hypothetical tournament involving 100 teams.

3.3 Estimation of model components, and prediction

Typical interest in statistical analysis of sports tournaments is addressed to ranking teams and forecasting match results. These two targets require estimation of team abilities and the home advantage effect. For these purposes we develop a Metropolis-Hastings algorithm as follows.

Let T be the time of the last match observed, $\mathcal{Y}_T$ be the set of all matches results played up to time T and $\hat{\boldsymbol{\theta}}_T$ be the maximum simulated likelihood estimate of $\boldsymbol{\theta}$ based on all matches played up to time $T \in \{1, \ldots, 2(n-1)\}$. Estimation of team abilities A_{it} for all times $t \leq T$ is based on suitable summaries of the smoothing density $p(A_{it}|\mathcal{Y}_T; \hat{\boldsymbol{\theta}}_T)$. This conditional density can be conveniently estimated through the following Metropolis-Hastings algorithm:

1. Set initially all model components equal to zero, then repeat the following step a sufficiently large number of times

2. Simulate a proposal $\tilde{a}_{it}$ for the ability of each team at each different time $t \leq T$ in turn from the conditional distribution of A_{it} given all the other abilities. The probability of accepting this proposal depends on whether at time t team i plays at home or away. Consider the first case and suppose that team j is playing away from home, then $t = t_{ij} \leq T$ and the proposal $\tilde{a}_{it_{ij}}$ is accepted with probability given by the smaller of

13

1 and the ratio of the match result probabilities

$$\frac{\mathrm{pr}(Y_{ij} = y_{ij} | A_{it_{ij}} = \tilde{a}_{it_{ij}}, A_{jt_{ij}} = a_{jt_{ij}}; \hat{\boldsymbol{\theta}}_T)}{\mathrm{pr}(Y_{ij} = y_{ij} | A_{it_{ij}} = a_{it_{ij}}, A_{jt_{ij}} = a_{jt_{ij}}; \hat{\boldsymbol{\theta}}_T)},$$

otherwise leave $A_{it_{ij}}$ unchanged from its current value $a_{it_{ij}}$. On the other hand, if team i plays away at home of team j, we have $t = t_{ji} \leq T$ and the proposal $\tilde{a}_{it_{ji}}$ is accepted with probability given by the smaller of 1 and the ratio of the match result probabilities

$$\frac{\mathrm{pr}(Y_{ji} = y_{ji} | A_{it_{ji}} = \tilde{a}_{it_{ji}}, A_{jt_{ji}} = a_{jt_{ji}}; \hat{\boldsymbol{\theta}}_T)}{\mathrm{pr}(Y_{ji} = y_{ji} | A_{it_{ji}} = a_{it_{ji}}, A_{jt_{ji}} = a_{jt_{ji}}; \hat{\boldsymbol{\theta}}_T)},$$

otherwise leave $A_{it_{ji}}$ unchanged from its current value $a_{it_{ji}}$.

At convergence, draws from this Metropolis-Hastings algorithm are used for estimation of various summaries of the smoothing densities $p(A_{it}|\mathcal{Y}_T; \hat{\boldsymbol{\theta}}_T)$ for $t \leq T$ and $i = 1, \ldots, n$. Comparisons among the various teams at a certain point of the tournament can be based on the estimated abilities $\hat{A}_{it} = \mathrm{E}(A_{it}|\mathcal{Y}_T; \hat{\boldsymbol{\theta}}_T)$. Expected values $\hat{A}_{it}$ are estimated through the averages of draws a_{it} from the above Metropolis-Hastings algorithm. Finally, at time t teams are ranked on the basis of the average of the estimated time-varying abilities from the beginning of the tournament up to time t, with $t = 1, \ldots, 2(n-1)$.

The Metropolis-Hastings algorithm is also used for prediction of future match results. For example, suppose the target is forecasting the result of the future match between home team i and away team j using information in all the matches played up to time T. Forecast probabilities are then computed as $\mathrm{pr}(Y_{ij} = k|\mathcal{Y}_T; \hat{\boldsymbol{\theta}}_T) = \Phi(\hat{\delta}_k + \hat{h} + \hat{A}_{it_{ij}} - \hat{A}_{jt_{ij}}) - \Phi(\hat{\delta}_{k-1} + \hat{h} + \hat{A}_{it_{ij}} - \hat{A}_{jt_{ij}})$ for $k = 1, 2, 3$, where $\hat{A}_{it_{ij}} = \hat{\gamma}_T^{|t_{ij}-T|}\mathrm{E}(A_{iT}|\mathcal{Y}_T; \hat{\boldsymbol{\theta}}_T)$ is the forecasted ability for home team i at future time t_{ij} with $\hat{\gamma}_T$ being the component pertaining to γ in $\hat{\boldsymbol{\theta}}_T$. The forecast ability for the away team is estimated similarly.

4   Application of the model

We fit the proposed model to the volleyball and football data described in Section 2 with the GHK algorithm using $2,000$ Monte Carlo replications. Adequacy of the Monte Carlo size has been checked by the stability of parameter estimates in repeated estimation using different pseudo-random seeds.

Parameter estimates for the volleyball data are (with standard errors in brackets) $\hat{h} = 0.57$ (0.11), $\hat{\gamma} = 0.96$ (0.03) and $\hat{\sigma} = 0.62$ (0.18). Fitting the model to football data gives qualitatively similar estimates: $\hat{\delta}_2 = 0.39$ (0.04), $\hat{h} = 0.40$ (0.06), $\hat{\gamma} = 0.98$ (0.02) and $\hat{\sigma} = 0.37$ (0.08). The notably large values of the estimate of the autocorrelation parameters confirm the very slowly varying abilities observed in Figure 1. Hence, we re-fit the two data sets with the limiting case of abilities constant in time. In these restricted models parameter estimates for h and δ_2 are essentially unchanged while the estimate of σ is slightly smaller: the estimates for the volleyball data are $\hat{h} = 0.52$ (0.11) and $\hat{\sigma} = 0.47$ (0.11), while those for football are $\hat{\delta}_2 = 0.38$ (0.04), $\hat{h} = 0.39$ (0.06) and $\hat{\sigma} = 0.32$ (0.06). For the volleyball tournament, the maximized log-likelihood for the model with constant-in-time abilities is -106.38 against a value of -105.56 for the dynamic model. Since the restricted models correspond to the limiting case of γ approaching one, the likelihood ratio test statistic does not have the usual asymptotic χ_1^2 distribution under the null hypothesis of constant-in-time abilities, $H_0 : \{A_{it} = A_i, \text{ for all times } t\}$. This difficulty can be overcome by relying on a parametric bootstrap assessment, which produces a p-value equal to 0.10 based on 500 replicates and thus finds no convincing evidence against the null hypothesis. The same conclusion is drawn for football data where the maximized log-likelihoods are -378.92 and -378.36 for the constant and dynamic models, respectively, and the parametric bootstrap test yields a p-value equal to 0.13.

Tables 1 and 2 suggest the possible incorporation of heterogeneity in the home advantage effect, for example by considering team-specific home effects parameters h_i. Although technically possible since a complete double round-robin tournament involves $\mathcal{O}(n^2)$ matches,

15

estimation of the n team-specific home effects h_i is impractical. Instead, we consider also a random-effect type specification where team-specific home effects h_i are realizations of a random variable H supposed independent of the team abilities A_{it} and distributed as a normal variable with mean h and variance σ_{H}^2. In the volleyball data, the parameter estimates for this heterogeneity model are $\hat{h} = 0.58$ (0.14), $\hat{\sigma}_H = 0.27$ (0.27), $\hat{\sigma} = 0.63$ (0.19) and $\hat{\gamma} = 0.96$ (0.04) with a maximized log-likelihood equal to -105.4. These values indicate the absence of a statistically significant home advantage heterogeneity beyond what is captured by the abilities A_{it}. The same conclusion is drawn for the football data, indeed in this case the estimates of a model with heterogeneous home effects are $\hat{\delta}_2 = 0.39$ (0.04), $\hat{h} = 0.40$ (0.08), $\hat{\sigma}_H = 0.19$ (0.13), $\hat{\sigma} = 0.36$ (0.08) and $\hat{\gamma} = 0.98$ (0.02) with a maximized log-likelihood of -377.92. In this instance, a parametric bootstrap assessment of the likelihood ratio test for the hypothesis $H_0 : \sigma_H = 0$ based on 500 replicates produces the estimated p-values equal to 0.22 for the volleyball data and 0.15 for the football data. See also the final discussion in Section 5.

The Metropolis-Hastings algorithm discussed in Section 3.3 is used for estimation of the model components A_{it} with the volleyball and football tournaments. Given the nature of the algorithm and the dependence structure of the stochastic components, successive draws show a noticeable correlation which decays quite slowly. In order to reduce correlation between successive draws, the chain is thinned by saving one draw in ten. A total of $50,000$ draws are saved and the first $10,000$ are discarded to allow for burn-in. Convergence of the last $40,000$ draws is also checked through standard diagnostic tools implemented in the `coda` R package (Plummer *et al.*, 2009). In particular, the convergence diagnostics based on the work of Geweke (1992) and Heidelberger and Welch (1983) do not indicate any problems.

Column six of Table 1 shows the average of the estimated abilities of the volleyball teams for the complete season along with their quasi standard errors. The ranking derived from the mean abilities of the dynamic model closely agrees with the actual final ranking and with the ranking arising from the static model, in fact the Kendall τ rank correlation with these

16

two rankings equals 0.884 and 0.978, respectively.

The left panel of Figure 2 displays the 95% confidence intervals for the estimated mean abilities constructed from the quasi standard errors. The estimates suggest that there may be groups of teams with different strength, even though it should be noticed that the confidence intervals are fairly wide. The best two teams, which coincide with the first two teams in the final points ranking, appear to have an ability higher than the mean ability of other teams in the league. Latina, Vibo Valentia, Verona and Padova appear the weakest teams in the mean ability ranking. In particular, the last two teams both won 6 matches and lost 20 during the season. However, the points system which rewards by one point a match lost by one set creates a final ranking in which Padova is higher than Vibo Valentia even though the latter won 8 matches and lost 18.
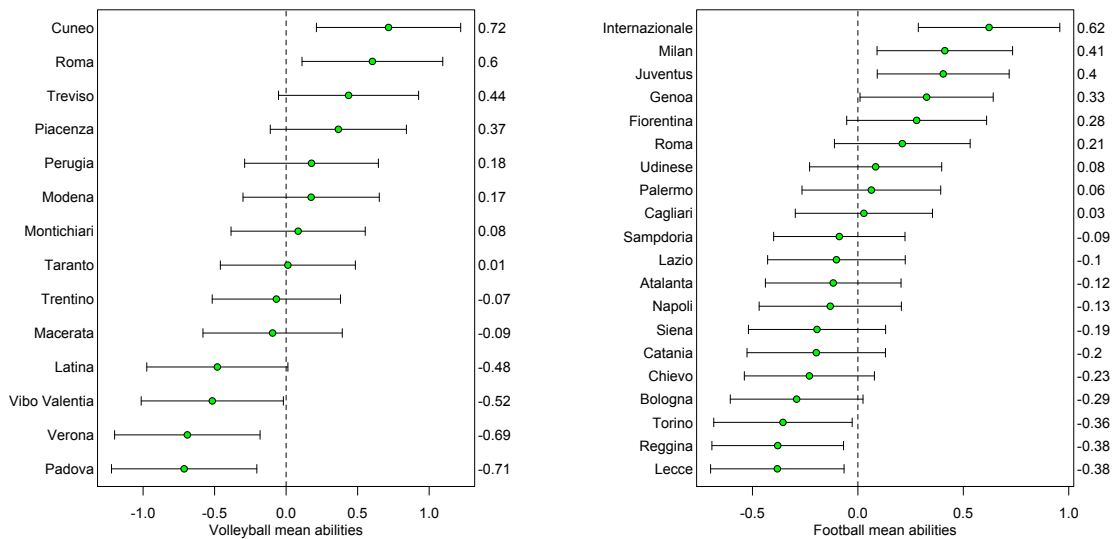

Figure 2: The left panel of the figure displays the 95% confidence intervals for the mean of the estimated abilities based on the quasi standard errors of the teams competing in the 2006-2007 regular season of the men's Italian A1 volleyball league. The right panel displays the 95% confidence intervals for the mean of the estimated abilities based on the quasi standard errors of the teams competing in the 2008-2009 regular season of the men's Italian Serie A league.

The right panel of Figure 2 shows the caterpillar plot of the average of the estimated abilities of the twenty football teams competing in the 2008-2009 Serie A league. The plot

17

reveals that only the best four teams in the league in terms of the points ranking have an estimated average ability higher than zero at an approximate 95% confidence level.

Figures 3 and 4 show the estimated mean abilities of a few teams in the volleyball and football leagues over the course of season. Cuneo and Roma occupy the first and second positions in the final volleyball ranking. Their ability was high for the whole season, but Cuneo was stronger at the beginning of the tournament. Verona experienced an unlucky start to the season with some players injured and consequently a poor performance. After six games the coach was changed and the performance of Verona improved noticeably, nevertheless the team ended the season in last position. The opposite behaviour is shown by Latina which experienced an evident decrease of its ability in the second half of the season: in fact it lost all of the last nine matches.

Regarding the football Serie A league, Figure 4 clearly reveals how much better Internazionale performed than other top teams Milan and Juventus during the whole season. Indeed, Internazionale ended the tournament ten points ahead of Milan and Juventus. The figure also displays the ability for Reggina which is among the weakest teams in the season. The abilities of the other teams are all between those of Internazionale and Reggina. In particular, Napoli shows a significant decrease in its performance a few matches after the beginning of the season. This decline stopped when the team changed coach 13 matches before the end of the tournament.

Table 3 shows model-based forecasts for all the matches of gameweek 24 of the volleyball tournament. Assessment of forecast quality can be based on the Brier score (Brier, 1950) given by

$$\text{BS} = \sum_{l=1}^{k} \left(f_l - o_l \right)^2,$$

where f_l is the forecasted probability of outcome l and o_l is 1 if l occurs and 0 otherwise, k is equal to 3 or 2 depending on whether the sport allows for ties or not, respectively. For each match, a Brier score equal to zero corresponds to a perfect prediction. At the other extreme, a Brier score equal to two corresponds to a completely erroneous prediction. The sum of the
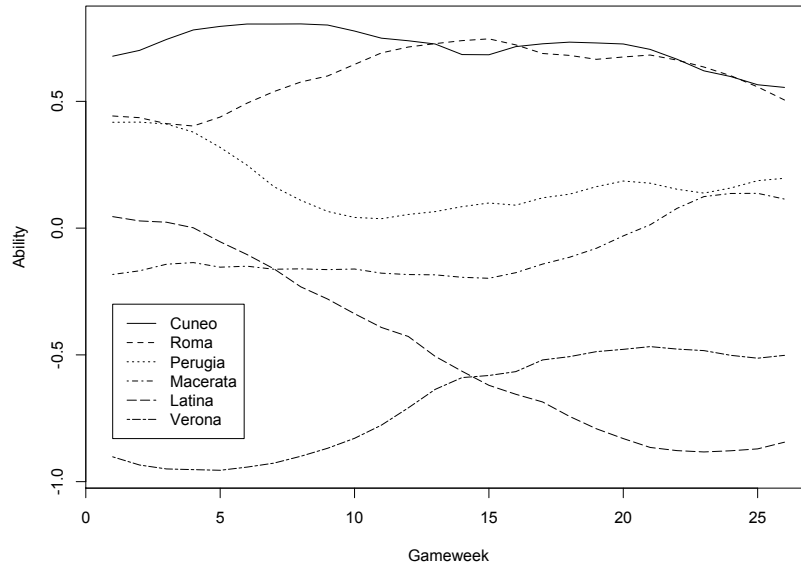
18

Figure 3: 2006-2007 regular season men's Italian A1 volleyball league. Estimated time-varying mid-abilities during the complete season for teams Cuneo, Latina, Macerata, Perugia, Roma, and Verona.
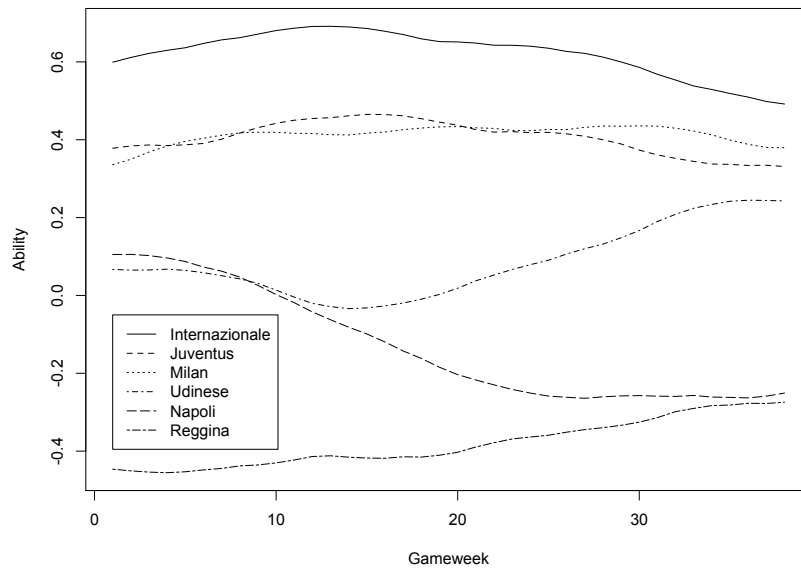

Figure 4: 2008-2009 Italian Serie A football league. Estimated time-varying mid-abilities during the complete season for teams Internazionale, Juventus, Milan, Napoli, Reggina and Udinese.

19

Brier scores for all the matches can be used for an overall assessment of the forecast quality: the lower the sum, the better the forecasts. The total score of our forecasts for the volleyball matches of gameweek 24 is 1.50. If we had simply used a uniform distribution on the results giving each outcome probability 0.5, the Brier score would be considerably worse with a value of 3.5. If we had used the empirical probabilities of home and away win, equal to 0.65 and 0.35, respectively, then the Brier score would have been 2.32.

Table 3: 2006-2007 regular season men's Italian A1 volleyball league. Results, model-based forecasts and Brier scores for all the matches in gameweek 24.

| | | | forecasts | | |
home	away	result	win	loss	Brier
Cuneo	Treviso	win	0.77	0.23	0.11
Padova	Trentino	loss	0.59	0.41	0.69
Piacenza	Latina	win	0.93	0.07	0.01
Vibo Valentia	Montichiari	win	0.57	0.43	0.36
Macerata	Taranto	win	0.67	0.33	0.22
Modena	Verona	win	0.79	0.21	0.09
Roma	Perugia	win	0.89	0.11	0.02

Table 4 reports the model-based predictions for the ten matches of gameweek 29 of the Italian Serie A tournament. The total Brier score for these predictions is 5.19. In this case a uniform distribution on the probabilities of the outcomes $(1/3, 1/3, 1/3)$, would yield a Brier score of 6.67, while the observed proportions of wins, draws and losses $(0.51, 0.25, 0.24)$ observed throughout the season up to gameweek 28, would produce a Brier score of 6.26. The broad conclusion from these forecasts and those not shown here for other gameweeks is that, in both sports, the dynamic model produces substantially better forecasts than the empirical proportions.

5   Conclusions

We have described a dynamic stochastic paired comparison model for the results of matches in sport tournaments. The model specification induces a natural cross-correlation between pairs

20

Table 4: 2008-2009 Italian Serie A football league. Results, model-based forecasts and Brier scores for all the matches in gameweek 29.

home	away	result	forecasts			Brier
			win	draw	loss	
Catania	Lazio	win	0.49	0.29	0.22	0.40
Roma	Juventus	loss	0.43	0.30	0.27	0.81
Bologna	Cagliari	loss	0.42	0.30	0.28	0.79
Chievo	Palermo	win	0.47	0.29	0.24	0.42
Fiorentina	Siena	win	0.61	0.25	0.14	0.23
Genoa	Udinese	win	0.63	0.24	0.13	0.21
Internazionale	Reggina	win	0.82	0.14	0.04	0.05
Lecce	Atalanta	draw	0.37	0.31	0.32	0.72
Napoli	Milan	draw	0.24	0.30	0.46	0.77
Torino	Sampdoria	loss	0.42	0.30	0.28	0.79

of matches sharing a team and takes into account the changes in abilities of teams during the season. Difficulties in likelihood computation deriving from the crossed dependence structure have been overcome by resorting to maximum simulated likelihood through the Geweke-Hajivassiliou-Keane algorithm.

The methodology is illustrated through the analysis of two complete round-robin tournaments where ties are or are not allowed. In both applications it is found that the team abilities are quite constant during the season. A similar conclusion is drawn by Knorr-Held (2000) about the 1996-1997 *Bundesliga* tournament and the 1996-1997 American Nation Basketball Association season.

The issue of whether team-specific home advantages should be included in paired comparison models was considered by many authors with contrasting conclusions. Knorr-Held (1997) does not find much evidence of home advantage heterogeneity among teams in the *Bundesliga*. Neither do the results in Harville and Smith (1994) show much difference in home field advantages among college basketball teams. Analyses for some other contexts do, however, support heterogeneity in home advantages, see Clarke and Norman (1995), Kuk (1995) and Glickman and Stern (1998) for different analyses of the English Premier Football

League.

Following previous analyses of team tournaments, we have assumed equispaced match times corresponding to the gameweeks. However, this assumption could be inappropriate because gameweeks often last more than a day, breaks during the season may occur or matches may be postponed. For example, in many European football leagues matches are played from Friday to Monday with breaks during colder periods or around the matches of the national teams. Furthermore, single matches can be postponed because of extreme weather conditions or concomitant international club matches. These considerations would suggest a continuous-time model specification

$$\text{pr}(Y_{ij} \leq y_{ij} | H_i = h_i, A_i(t_{ij}) = a_i(t_{ij}), A_j(t_{ij}) = a_j(t_{ij})) = F(\delta_{y_{ij}} + h + a_i(t_{ij}) - a_j(t_{ij})),$$

where $A_i(t_{ij})$ denotes the ability of team i at the calendar time t_{ij} of its match against team j. We have fitted this continuous-time model to both the volleyball and football data but, for these two particular tournaments, we found negligible differences in estimated model parameters relative to the approximate discrete-time specification.

The average abilities reproduce closely the final points ranking of the tournaments. In volleyball a ranking even closer to the final one might be obtained by considering four possible categories for the match result, namely decisive loss, narrow loss, narrow win and decisive win corresponding to the results of matches which give to the teams 0, 1, 2 or 3 points.

The focus in this paper has been on modelling match results, but similar ideas can be exploited in the analysis of difference in goals scored in each match. Indeed, the latent process Z_{ij} can be viewed as a model for such a difference.

The model discussed in the paper easily allows the incorporation of time-constant covariates. More difficult but also much more interesting is modelling match results as a function of endogenous time-varying covariates such as the number of goals scored and conceded in previous matches. This type of analysis may be very helpful for forecasting match results, and will be the subject of future research.

22

References

Albert, J., Bennett, J., Cochran, J. J. eds. (2005) *Anthology of Statistics in Sports*, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2005.

Barry, D. and Hartigan, J. A. (1993) Choice models for predicting divisional winners in major league baseball. *Journal of the American Statistical Association* **88**, 766-774.

Bradley, R. A. and Terry, M. E. (1952) Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika* **39**, 324-345.

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

Clarke, S. R. and Norman, J. M. (1995) Home ground advantage of individual clubs in English soccer. *The Statistician* **44**, 509-521.

Crowder, M., Dixon, M., Ledford, A., Robinson, M. (2002) Dynamic modelling and prediction of English football league matches for betting. *The Statistician* **51**, 157-168.

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

Durbin, J. and Koopman, S. J. (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84**, 669-684.

Fahrmeir, L. and Tutz, G. (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association* **89**, 1438-1449.

Firth, D. and de Menezes, R. X. (2004) Quasi-variances. *Biometrika* **91**, 65-80.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in *Bayesian Statistics*, eds. Bernardo J. M., Berger J., Dawid A. P. and Smith, A. F. M. Oxford University Press, Oxford.

Glickman, M. E. (1999) Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics* **48**, 377-394.

Glickman, M. E. and Stern, H. S. (1998) A state-space model for national football league scores. *Journal of the American Statistical Association* **93**, 25-35.

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

Harville, D. A. (1980) Predictions for national football league games via linear-model methodology. *Journal of the American Statistical Association* **75**, 516-524.

Harville, D. A. (2003) The selection or seeding of college basketball or football teams for postseason competition. *Journal of the American Statistical Association* **98**, 17-27.

Harville, D. A. and Smith, M. H. (1994) The home-court advantage: how large is it, and does it vary from team to team? *American Statistician* **48**, 22-28.

Heidelberger, P. and Welch, P. D. (1983) Simulation run length control in the presence of an initial transient. *Operations Research* **31**, 1109-1144.

Karlis, D. and Ntzoufras, I. (2003) Analysis of sports data by using bivariate Poisson models. *Statistician* **52**, 381-393.

Knorr-Held, L. (1997) *Hierarchical Modelling of Longitudinal Data; Applications of Markov Chain Monte Carlo*. Munich: Utz.

Knorr-Held, L. (2000) Dynamic rating of sports teams. *The Statistician* **49**, 261-276.

Koning, R. H. (2000) Balance in competitions in Dutch soccer. *The Statistician* **49**, 419-431.

24

Kuk, A. Y. C. (1995) Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. *The Statistician* **44**, 523-528.

Maher, M. J. (1982) Modelling association football scores. *Statistica Neerlandica* **36**, 109-118.

Masarotto, G. and Varin C. (2010) Gaussian dependence models for non-Gaussian marginal regression. *Submitted.*

McHale, I. and Scarf, P. (2007) Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica* **61**, 432-445.

Mosteller, F. (1951) Remarks on the method of paired comparisons. I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* **16**, 3-9.

Plummer, M., Best, N., Cowles, K., Vines, K. (2009) `coda`: Output analysis and diagnostic for MCMC. R package.

Rue, H. and Salvesen O. (2000) Prediction and retrospective analysis of soccer matches in a league. *The Statistician* **49**, 399-418.

Thurstone, L. L. (1927) A law of comparative judgement. *Psychological Review* **34**, 273-286.

Train, K. E. (2003) *Discrete Choice Methods with Simulation.* Cambridge University Press, Cambridge.