The Statistician (2000)
49, *Part* 3, *pp.* 399–418

Focus on Sport

Prediction and retrospective analysis of soccer matches in a league

Håvard Rue and Øyvind Salvesen

Norwegian University of Science and Technology, Trondheim, Norway

[Received October 1997. Final revision March 2000]

Summary. A common discussion subject for the male part of the population in particular is the prediction of the next week-end's soccer matches, especially for the local team. Knowledge of offensive and defensive skills is valuable in the decision process before making a bet at a bookmaker. We take an applied statistician's approach to the problem, suggesting a Bayesian dynamic generalized linear model to estimate the time-dependent skills of all teams in a league, and to predict the next week-end's soccer matches. The problem is more intricate than it may appear at first glance, as we need to estimate the skills of all teams simultaneously as they are dependent. It is now possible to deal with such inference problems by using the Markov chain Monte Carlo iterative simulation technique. We show various applications of the proposed model based on the English Premier League and division 1 in 1997–1998: prediction with application to betting, retrospective analysis of the final ranking, the detection of surprising matches and how each team's properties vary during the season.

Keywords: Dynamic models; Generalized linear models; Graphical models; Markov chain Monte Carlo methods; Prediction of soccer matches

1. Introduction

Soccer is a popular sport all over the world, and in Europe and South America it is the dominant spectator sport. People find interest in soccer for various reasons and at different levels, with a clear dominance for the male part of the population. Soccer is an excellent game for different forms of betting. The outcome of a soccer match depends on many factors; among these are the home ground–away ground effect, the effect of injured players and various psychological effects. Good knowledge about these factors only determines the final result up to a significant, but not too dominant, random component.

Different models for predicting the outcomes of soccer matches for betting have existed for a long time. A quick tour around the World Wide Web, perhaps starting at

```
http://dmiwww.cs.tut.fi/riku,
```

shows an impressive activity with links to small companies selling ready-to-go computer programs for prediction and betting, discussion groups and bookmakers who operate on the Web. The most popular ideas in the prediction programs are to consider win, draw and loss sequences, goals

Address for correspondence: Håvard Rue, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7034 Trondheim, Norway.
E-mail: Havard.Rue@math.ntnu.no

0039–0526/00/49399

scored over five matches and the points difference in the current ranking. Most of the programs have a Bayesian flavour and allow the user to include his or her expert knowledge in various ways. The (pure) statistical side of soccer is not so well developed and widespread. Ridder *et al.* (1994) analysed the effect of a red card in a match, Kuonen (1996) modelled knockout tournaments, Lee (1997) provided a simplified generalized linear model with application to final rank analysis and Dixon and Coles (1997) and Dixon and Robinson (1998) have provided a more comprehensive model.

In this paper, we model the results of soccer matches played in a league, where the teams play against each other twice (home and away) relatively regularly and during a limited time period. We shall use the history of the matches played to estimate what we think are the two most important (time-dependent) explanatory variables in a Bayesian dynamic generalized linear model: the strengths of attack and defence. It is more intricate than we might think to estimate the properties like the strengths of attack and defence for each team. Assume that teams A and B play against each other with the result 5–0. One interpretation of this result is that A has a powerful attack; another that B has a weak defensive strength. The properties for A and B conditional on the result are therefore dependent. As each team plays against all the other teams in a league, we soon reach full dependence between the (time-dependent) properties for *all* teams. We can analyse such problems by using Markov chain Monte Carlo (MCMC) techniques (Gilks *et al.*, 1996) to generate dependent samples from the posterior density. The model proposed and the power of MCMC sampling can be used to make predictions for the next round in the league and to answer other interesting questions as well, like ‘what is the expected ranking at the end of the season?’ and ‘were Arsenal lucky to win the Premier League in 1997–1998?’.

A simplified and stripped-down version of our model is similar to the generalized linear model developed independently by Lee (1997). Dixon and Coles (1997) and Dixon and Robinson (1998) presented a more comprehensive model than Lee (1997), trying to mimic time-varying properties by downweighting the likelihood. They could not provide a coherent model for how the properties develop in time. In this paper we provide a Bayesian model which models the time variation of all properties simultaneously, present a new parameterization and ideas in goal modelling and show how we can make inference and do a retrospective analysis of a season using the power of MCMC sampling. We need a joint model for the properties in time to do a retrospective analysis of a season, to be able to estimate each team’s properties at time t using data from matches *both* before and after time t. Our approach provides a coherent model which is easy to extend to account for further refinement and development in the art of predicting the outcomes of soccer matches. Other similar sports problems could be approached in the same manner.

The rest of the paper is organized as follows. We start in Section 2 with the basic explanatory variables and derive the model step by step. Section 3 describes how we estimate global parameters by using historical data and make inference from the model by using MCMC techniques. In Section 4 we apply our model to analyse the English Premier League and division 1 in 1997–1998, with a focus on prediction with application to betting, retrospective analysis of the final ranking, locating surprising matches and studying how each team’s properties vary during the season.

2. The model

In this section we derive our model for analysing soccer matches in a league. We start with the basic explanatory variables, continue by linking these variables with a goal model and model their dependence in time. In Section 2.4 we collect all the pieces in the full model.

2.1. Basic explanatory variables

Many explanatory variables will influence the result of a forthcoming soccer match. Which factors we should include in the model depend on what kind of data is available. To keep the problem simple, we shall only make use of the result in a match, like the home team wins 3–1 over the away team. We therefore ignore all other interesting data like the number of near goals, corner kicks, free kicks and so on. Our next step is to specify which (hidden) explanatory variables attached to each team will influence the result of a match.

The two important properties of each team are their defending and attacking skills. The strengths of defence and attack are represented as the random variables d and a respectively. A high value of d and a means a good defence and attack respectively. Let $\mathbf{e}_{\mathrm{A}} = (a, d)_{\mathrm{A}}$ denote the properties for team A, and further let $\mu_{a,\mathrm{A}}$ and $\sigma^2_{a,\mathrm{A}}$ be the prior mean and variance for a_{A}, and similarly for the strength of the defence and the other teams.

2.2. The goal model

The next step is to specify how the result $(x_{\mathrm{A,B}}, y_{\mathrm{A,B}})$ depends on the properties of home team A and away team B. A reasonable assumption is that the number of goals that A scores $(x_{\mathrm{A,B}})$ depends on A's strength of attack and B's strength of defence. Similarly, the number of goals that B scores $(y_{\mathrm{A,B}})$ depends on B's strength of attack and A's strength of defence. Additionally, we include a psychological effect; team A will tend to underestimate the strength of team B if A is a stronger team than B. Let $\Delta_{\mathrm{AB}} = (a_{\mathrm{A}} + d_{\mathrm{A}} - a_{\mathrm{B}} - d_{\mathrm{B}})/2$ measure the difference in strength between A and B. We assume further that

$$x_{\mathrm{A,B}} | (\mathbf{e}_{\mathrm{A}}, \mathbf{e}_{\mathrm{B}}) \overset{\mathrm{d}}{=} x_{\mathrm{A,B}} | a_{\mathrm{A}} - d_{\mathrm{B}} - \gamma \Delta_{\mathrm{AB}},$$

$$y_{\mathrm{A,B}} | (\mathbf{e}_{\mathrm{A}}, \mathbf{e}_{\mathrm{B}}) \overset{\mathrm{d}}{=} y_{\mathrm{A,B}} | a_{\mathrm{B}} - d_{\mathrm{A}} + \gamma \Delta_{\mathrm{AB}},$$

where γ is a small constant giving the magnitude of the psychological effect. We assume that the strengths of team A and B are not much different since we are analysing teams in the same league, so it is reasonable to expect $\gamma > 0$. (The opposite effect $(\gamma < 0)$ might occur if team A is so superior compared with team B that the latter develops an inferiority complex facing team A, which we do not expect will happen in the same league.)

To motivate our probability law for $x_{\mathrm{A,B}}$ and $y_{\mathrm{A,B}}$, we display in Fig. 1 the histogram of the results in 924 matches in the Premier League in 1993–1995. The histogram and nature of the game itself suggest to a first approximation a Poisson law for $x_{\mathrm{A,B}}$ and $y_{\mathrm{A,B}}$. Thus, as a first approximation we may assume that the number of goals conditioned on the teams' properties is Poisson distributed with mean $\lambda^{(x)}_{\mathrm{A,B}}$ and $\lambda^{(y)}_{\mathrm{A,B}}$, where

$$\begin{aligned} \log(\lambda^{(x)}_{\mathrm{A,B}}) &= c^{(x)} + a_{\mathrm{A}} - d_{\mathrm{B}} - \gamma \Delta_{\mathrm{AB}}, \\ \log(\lambda^{(y)}_{\mathrm{A,B}}) &= c^{(y)} + a_{\mathrm{B}} - d_{\mathrm{A}} + \gamma \Delta_{\mathrm{AB}}. \end{aligned} \tag{1}$$

Here, $c^{(x)}$ and $c^{(y)}$ are global constants describing (roughly) the logarithm of the empirical mean of the home and away goals.

Although independence of $x_{\mathrm{A,B}}$ and $y_{\mathrm{A,B}}$ has been verified empirically to be quite reasonable (Lee, 1997), it does not imply that $x_{\mathrm{A,B}}$ and $y_{\mathrm{A,B}}$ are independent conditional on $(\mathbf{e}_{\mathrm{A}}, \mathbf{e}_{\mathrm{B}})$. Dixon and Coles (1997) proposed therefore to use the joint conditional law for $(x_{\mathrm{A,B}}, y_{\mathrm{A,B}})$

$$\pi_{\mathrm{gl}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}} | \lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}) = \kappa(x_{\mathrm{A,B}}, y_{\mathrm{A,B}} | \lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}) \,\mathrm{Po}(x_{\mathrm{A,B}} | \lambda^{(x)}_{\mathrm{A,B}}) \,\mathrm{Po}(y_{\mathrm{A,B}} | \lambda^{(y)}_{\mathrm{A,B}}) \tag{2}$$

where $\mathrm{Po}(x_{\mathrm{A,B}} | \lambda^{(x)}_{\mathrm{A,B}})$ is the Poisson law for $x_{\mathrm{A,B}}$ with mean $\lambda^{(x)}_{\mathrm{A,B}}$, and κ is a correction factor given as

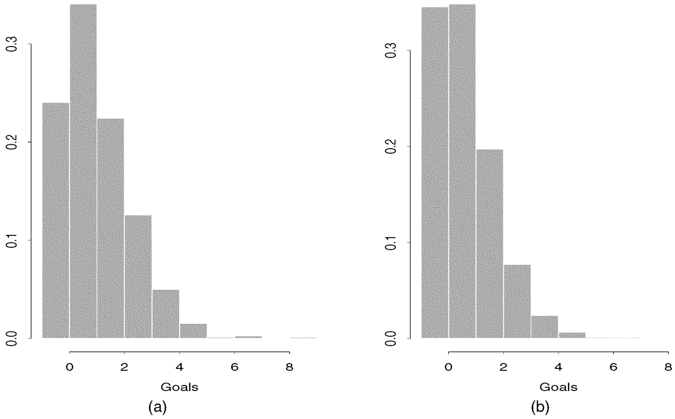


Fig. 1. Histograms of the number of (a) home and (b) away goals in 924 matches in the Premier League, 1993–1995

$$\kappa(x_{\mathrm{A,B}}, y_{\mathrm{A,B}} | \lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}) = \begin{cases} 1 + 0.1\lambda^{(x)}_{\mathrm{A,B}}\lambda^{(y)}_{\mathrm{A,B}} & \text{if } x_{\mathrm{A,B}} = 0,\ y_{\mathrm{A,B}} = 0, \\ 1 - 0.1\lambda^{(x)}_{\mathrm{A,B}} & \text{if } x_{\mathrm{A,B}} = 0,\ y_{\mathrm{A,B}} = 1, \\ 1 - 0.1\lambda^{(y)}_{\mathrm{A,B}} & \text{if } x_{\mathrm{A,B}} = 1,\ y_{\mathrm{A,B}} = 0, \\ 1.1 & \text{if } x_{\mathrm{A,B}} = 1,\ y_{\mathrm{A,B}} = 1, \\ 1 & \text{otherwise.} \end{cases}$$

The correction factor κ increases the probability of 0–0 and 1–1 results at the cost of 1–0 and 0–1 results. All the other joint probabilities remain unchanged. Note further that the (conditional) marginal laws of $x_{\mathrm{A,B}}$ and $y_{\mathrm{A,B}}$ from equation (2) are $\mathrm{Po}(x_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}})$ and $\mathrm{Po}(y_{\mathrm{A,B}}|\lambda^{(y)}_{\mathrm{A,B}})$ respectively. We found it necessary to modify equation (2) in two ways: the first modification is about the Poisson assumption; the second is a robustness adjustment.

Although the Poisson model seems reasonable, it may not be if one of the teams scores many goals. This is highly demotivating for the other team, and in most cases implies a contradiction with our underlying model assumption that the goal intensity does not depend on the goals scored during the match. We correct for this by truncating the Poisson law $\mathrm{Po}(x_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}})$ (and similarly for $y_{\mathrm{A,B}}$) in equation (2) after 5 goals. Denote by π^*_{g1} the resulting truncated law. The results 7–0 and 6–5 will be interpreted as 5–0 and 5–5 respectively. The goals that each team scores after their first 5 are non-informative on the two teams' properties.

It is our experience that the result of a match is less informative on the properties of the teams than equation (1) and π^*_{g1} suggest. We found it necessary to use a more model robust goal model by forming a mixture of laws,

$$\pi_{\mathrm{g}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}) = (1-\epsilon)\pi^*_{\mathrm{g1}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}) + \epsilon\pi^*_{\mathrm{g1}}\{x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\exp(c^{(x)}), \exp(c^{(y)})\}. \tag{3}$$

Here, ϵ is a parameter with the interpretation that only $100(1-\epsilon)\%$ of the 'information' in the match result is informative on $\mathbf{e}_{\mathrm{A}}$ and $\mathbf{e}_{\mathrm{B}}$, and the remaining $100\epsilon\%$ is not informative. The non-informative part of π_{g} uses the average goal intensities, $\exp(c^{(x)})$ and $\exp(c^{(y)})$, and π_{g} therefore shrinks π^*_{g1} towards the law for an average match. The value of ϵ is found in Section 3.2 to be around 0.2.

2.3. Time model

It is both natural and necessary to allow the attack and defence variables to vary with time. For the discussion here, assume that t' and $t'' \geqslant t'$ are two following time points (the number of days from a common reference point) where team A plays a match. Let us consider the strength of attack where similar arguments are valid also for the strength of defence. We need to specify how $a_{\mathrm{A}}^{t''}$ (with superscript t for time) depends on $a_{\mathrm{A}}^{t'}$ and possibly on previous values. Our main purpose is to predict matches in the near future, so only a reasonable local behaviour for a_{A} in time is needed. As a first choice, we borrow ideas from dynamic models (West and Harrison, 1997) and use Brownian motion to tie together a_{A} at the two time points t' and t'', i.e. conditional on $\{a_{\mathrm{A}}^{t},\ t \leqslant t'\}$:

$$a_{\mathrm{A}}^{t''} \overset{\mathrm{d}}{=} a_{\mathrm{A}}^{t'} + \left\{ B_{a,\mathrm{A}}\left(\frac{t''}{\tau}\right) - B_{a,\mathrm{A}}\left(\frac{t'}{\tau}\right) \right\} \frac{\sigma_{a,\mathrm{A}}}{\surd\{1-\gamma(1-\gamma/2)\}}. \tag{4}$$

(Recall that $\sigma_{a,\mathrm{A}}^2$ is the prior variance for a_{A}.) Here, $\{B_{\cdot,\cdot}(t),\ t \geqslant 0\}$ is standard Brownian motion starting at level 0 and where the subscript marks that the strength of attack belongs to team A. The last factor is a scaling factor which we motivate in the next paragraph. The characteristic time parameter τ is common to all teams and gives the inverse loss of memory rate for a_{A}^{t}, $\mathrm{var}(a_{\mathrm{A}}^{t''} - a_{\mathrm{A}}^{t'}) \propto \sigma_{a,\mathrm{A}}^2/\tau$. We model the strengths of attack and defence for all teams as in equation (4) and assume that the Brownian motion processes are independent.

Using Brownian motion is a weak and flexible way of modelling the development of the properties over time. Although the process itself is non-stationary with a variance increasing towards ∞ with increasing time, the conditioning on the match results will ensure smooth posterior realizations of the properties and reasonable extrapolation into the near future.

The common parameters γ and τ control the psychological effect and the loss of memory rate. These parameters have a nice interpretation if we consider the conditional (on the past $\{t < t''\}$) expected value and variance in the Gaussian conditional density for $\log(\lambda_{\mathrm{A,B}}^{(x),t''})$ (and $\log(\lambda_{\mathrm{A,B}}^{(y),t''})$). If we assume for simplicity that $\sigma_{a,\mathrm{A}}^2 = \sigma_{d,\mathrm{A}}^2 = \sigma_{a,\mathrm{B}}^2 = \sigma_{d,\mathrm{B}}^2$, we obtain

$$E\{\log(\lambda_{\mathrm{A,B}}^{(x),t''})|\mathrm{past}\} = c^{(x)} + a_{\mathrm{A}}^{t'} - d_{\mathrm{B}}^{t'} - \gamma\Delta_{\mathrm{AB}}^{t'}$$

and

$$\mathrm{var}\{\log(\lambda_{\mathrm{A,B}}^{(x),t''})|\mathrm{past}\} = 2\sigma_{a,\mathrm{A}}^2 \frac{t''-t'}{\tau}. \tag{5}$$

Thus, γ adjusts the conditional expected value, and τ controls the conditional variance of $\log(\lambda_{\mathrm{A,B}}^{(x)})$. The scaling with γ in equation (4) makes γ and τ orthogonal in this sense. We interpret τ and γ as the main and secondary parameter of interest respectively.

2.4. Full model

We can now build the full model on the basis of the previous assumptions. The properties of each team obey equation (4) for the time development and equation (1) for the result of each match. We only need to keep records of which teams play against each other, when and where. Fig. 2 shows the situation schematically when four teams play 3 × 2 matches at time t_0, t_1 and t_2, and the fourth round at time t_3 is not played. (We choose to play the matches on the same days, so the notation is simplified. In practice this is not the case.) The graph is called the *directed acyclic graph* (Whittaker, 1990) and the directed edges in the graph show the flow of information or causal relationships between parent and child nodes in the graph. If we construct the corresponding moral graph (Whittaker, 1990) of Fig. 2, we shall find a path between each node in the graph soon after the league starts. We must therefore make inference for all the properties for each team at all time points, simultaneously.

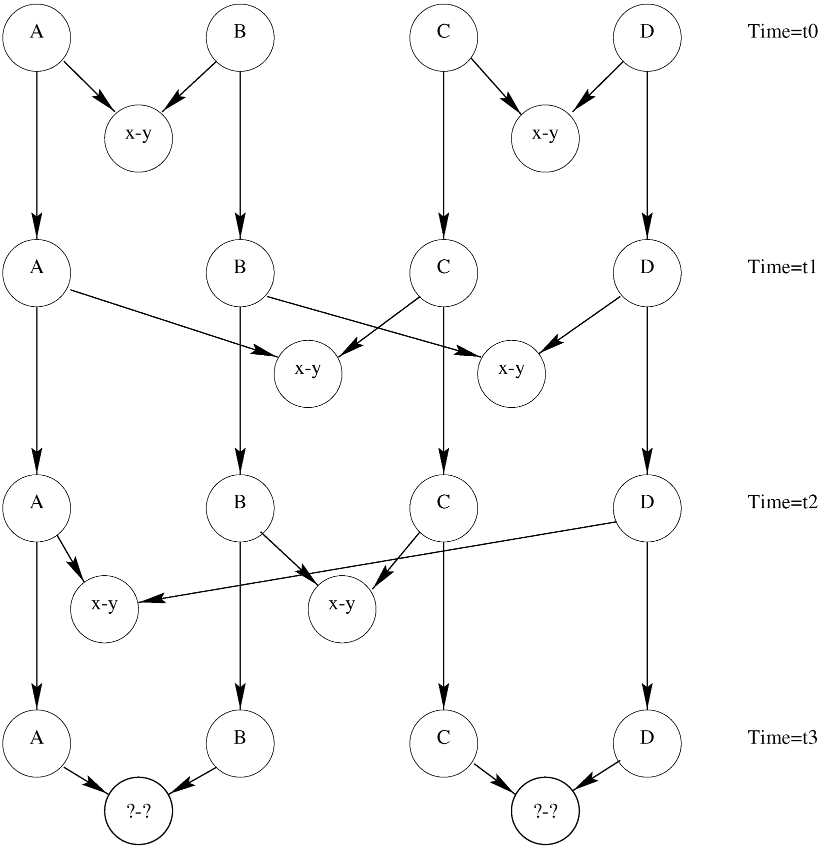


Fig. 2. Directed acyclic graph describing the causal structure in our model with four teams and eight matches

We write $\boldsymbol{\theta}$ for all the variables in the model and for the moment keep parameters ϵ, γ and τ fixed. The variables are the properties $\mathbf{e}_{\mathrm{A}}^{t_0}$, $\mathbf{e}_{\mathrm{B}}^{t_0}$, $\mathbf{e}_{\mathrm{C}}^{t_0}$ and $\mathbf{e}_{\mathrm{D}}^{t_0}$ at time t_0, the result of the match between A and B, and C and D at time t_0, the properties $\mathbf{e}_{\mathrm{A}}^{t_1}$, $\mathbf{e}_{\mathrm{B}}^{t_1}$, $\mathbf{e}_{\mathrm{C}}^{t_1}$ and $\mathbf{e}_{\mathrm{D}}^{t_1}$ at time t_1, the result of the match between A and C, and B and D at time t_1, and so on. The joint density for all variables in the model $\boldsymbol{\theta}$ is easy to find if we make use of Fig. 2. The joint density of $\boldsymbol{\theta}$ is the product of the conditional density of each node given its parents. By using $\pi(\cdot|\cdot)$ as a generic notation for the density of its arguments and indicate the time by superscripts, we obtain, starting from the top of the graph where each line of equation (6) corresponds to each row of the graph in Fig. 2,

$$\begin{aligned}\pi(\boldsymbol{\theta}) = {} & \pi(\mathbf{e}_{\mathrm{A}}^{t_0})\,\pi(\mathbf{e}_{\mathrm{B}}^{t_0})\,\pi(\mathbf{e}_{\mathrm{C}}^{t_0})\,\pi(\mathbf{e}_{\mathrm{D}}^{t_0}) \\ & \times \pi(x_{\mathrm{A,B}}^{t_0}, y_{\mathrm{A,B}}^{t_0}|\mathbf{e}_{\mathrm{A}}^{t_0}, \mathbf{e}_{\mathrm{B}}^{t_0})\,\pi(x_{\mathrm{C,D}}^{t_0}, y_{\mathrm{C,D}}^{t_0}|\mathbf{e}_{\mathrm{C}}^{t_0}, \mathbf{e}_{\mathrm{D}}^{t_0}) \\ & \times \pi(\mathbf{e}_{\mathrm{A}}^{t_1}|\mathbf{e}_{\mathrm{A}}^{t_0})\,\pi(\mathbf{e}_{\mathrm{B}}^{t_1}|\mathbf{e}_{\mathrm{B}}^{t_0})\,\pi(\mathbf{e}_{\mathrm{C}}^{t_1}|\mathbf{e}_{\mathrm{C}}^{t_0})\,\pi(\mathbf{e}_{\mathrm{D}}^{t_1}|\mathbf{e}_{\mathrm{D}}^{t_0}) \\ & \times \pi(x_{\mathrm{A,C}}^{t_1}, y_{\mathrm{A,C}}^{t_1}|\mathbf{e}_{\mathrm{A}}^{t_1}, \mathbf{e}_{\mathrm{C}}^{t_1})\,\pi(x_{\mathrm{B,D}}^{t_1}, y_{\mathrm{B,D}}^{t_1}|\mathbf{e}_{\mathrm{B}}^{t_1}, \mathbf{e}_{\mathrm{D}}^{t_1}) \\ & \times \ldots . \end{aligned} \tag{6}$$

Here $\pi(\mathbf{e}_{\mathrm{A}}^{t_0})$ is the prior density for $\mathbf{e}_{\mathrm{A}}$, which will be commented on in Section 3.2, and $\pi(x_{\mathrm{A,B}}^{t_0}, y_{\mathrm{A,B}}^{t_0}|\mathbf{e}_{\mathrm{A}}^{t_0}, \mathbf{e}_{\mathrm{B}}^{t_0})$ is the mixture law given in equation (3). Note that the mixture law depends on $\mathbf{e}_{\mathrm{A}}^{t_0}$ and $\mathbf{e}_{\mathrm{B}}^{t_0}$ only through $\lambda_{\mathrm{A,B}}^{(x),t_0}$ and $\lambda_{\mathrm{A,B}}^{(y),t_0}$, as given in equations (1) and (3). Further, $\pi(\mathbf{e}_{\mathrm{A}}^{t_1}|\mathbf{e}_{\mathrm{A}}^{t_0}) = \pi(a_{\mathrm{A}}^{t_1}|a_{\mathrm{A}}^{t_0})\,\pi(d_{\mathrm{A}}^{t_1}|d_{\mathrm{A}}^{t_0})$ where

$$a_{\mathrm{A}}^{t_1}|a_{\mathrm{A}}^{t_0} \sim N\left(a_{\mathrm{A}}^{t_0}, \frac{t_1 - t_0}{\tau}\sigma_{a,\mathrm{A}}^2\right)$$

and similarly for $d_{\mathrm{A}}^{t_1}|d_{\mathrm{A}}^{t_0}$. We denote by $N(\mu, \sigma^2)$ the Gaussian distribution with mean μ and variance σ^2. The rest of the terms in $\pi(\boldsymbol{\theta})$ have similar interpretations; only the teams and times differ.

We have not included the properties in between times t_0, t_1, t_2 and t_3, as their (conditional) distributions are known to be Brownian bridges conditional on the properties at time $t_0, \ldots, t_3$.

We can view the full model as a Bayesian dynamic model in the general framework of West and Harrison (1997), where equation (3) represents the observation equation and equation (4) the state space equation (taking all teams and properties into account).

To make inference for the properties of each team conditionally on the observed match results, we need the conditional (posterior) density derived from equation (6). It is difficult to analyse the posterior with direct methods because of the complexity and an intractable normalization constant. We can, however, make use of MCMC methods to analyse our model, and this will be further discussed in the next section, leaving the details for Appendix A.

3. Inference

In this section we shall discuss how we can make inferences from the model by making use of MCMC methods from the posterior density, for fixed values of the τ-, γ- and ϵ-parameters. We shall then discuss how we choose the constants $c^{(x)}$ and $c^{(y)}$, and how we estimate τ, γ and ϵ to maximize the predictive ability of the model.

3.1. The Markov chain Monte Carlo algorithm

We can make inferences from the posterior density proportional to equation (6) by using (dependent) samples from the posterior produced by MCMC methods. There is now an extensive literature on MCMC methods and Gilks *et al.* (1996) provide a comprehensive overview of the theory and the wide range of applications. In brief, to generate realizations from some density $f(\mathrm{d}\mathbf{z})$ we construct a Markov chain using an irreducible aperiodic transition kernel which has $f(\mathrm{d}\mathbf{z})$ as its equilibrium distribution. The algorithm runs as follows; suppose that the current state of the Markov chain is $\mathbf{z}$, and we propose a move of type j that moves $\mathbf{z}$ to $\mathrm{d}\mathbf{z}'$ with probability $q_j(\mathbf{z}, \mathrm{d}\mathbf{z}')$. The move to $\mathbf{z}'$ is accepted with probability $\min\{1, R_{\mathbf{z},\mathbf{z}'}\}$, where

$$R_{\mathbf{z},\mathbf{z}'} = \frac{f(\mathrm{d}\mathbf{z}')\,q_j(\mathbf{z}', \mathrm{d}\mathbf{z})}{f(\mathrm{d}\mathbf{z})\,q_j(\mathbf{z}, \mathrm{d}\mathbf{z}')}; \tag{7}$$

otherwise we stay in the original state $\mathbf{z}$. When q_j is symmetric, equation (7) reduces to $f(\mathrm{d}\mathbf{z}')/f(\mathrm{d}\mathbf{z})$ and the sampler is known as the Metropolis algorithm. Perhaps the best-known construction is the Gibbs sampler: take q_j to be the conditional density of the component(s) to be updated given the remaining components. Because $R_{\mathbf{z},\mathbf{z}'}$ is 1 in this case we always accept the new state.

In theory we need two different move types to implement an MCMC algorithm for our model: update the result for those matches which are not played and update the strengths of attack and defence for each team at each time that a match is played. However, to ease the implementation of the MCMC algorithm we reformulate the mixture model and attach an independent Bernoulli variable to each match. Each Bernoulli variable is updated during the MCMC algorithm and indicates which one of the distributions on the right-hand side of equation (3) we currently use. We refer to Appendix A for the details of the MCMC algorithm and a discussion of various MCMC issues. Refer also to chapter 15 in West and Harrison (1997) for a discussion of simulation-based inference in dynamic models.

The average acceptance rate for the MCMC algorithm proposed is around 55%. Even if our single-site updating algorithm is not that specialized, we obtain quite reasonable computational costs. The algorithm does 1000 updates of all variables in a case with 380 matches, in about 12 s on a Sun Ultra 4 computer with a 296 MHz SUNW UltraSPARC-II central processor unit.

3.2. Inference for $c^{(x)}$, $c^{(y)}$, τ, γ and ϵ

In this section we shall discuss how we choose various constants, validate our Gaussian prior distribution for the properties and estimate the important parameters τ, γ and ϵ by using historical data from the Premier League and division 1.

We used 1684 matches from the Premier League for 1993–1997 and 2208 matches from division 1 for 1993–1997 to estimate two sets of global constants $c^{(x)}$ and $c^{(y)}$. For simplicity we used the logarithm of the empirical mean of the home and away goals. The estimates are $c^{(x)} = 0.395$ and $c^{(y)} = 0.098$ for the Premier League and $c^{(x)} = 0.425$ and $c^{(y)} = 0.062$ for division 1. These values are close to those which we obtained with a more accurate and comprehensive Bayesian analysis.

It is tempting to use Gaussian priors for the properties of each team. To validate this assumption, we used 924 matches from the Premier League, assuming that each of the 924 match results were realizations from matches with a common distribution for $\log(\lambda^{(x)})$ and $\log(\lambda^{(y)})$. The posterior densities of $\log(\lambda^{(x)})$ and $\log(\lambda^{(y)})$ were close to Gaussian distributions, so our assumption seems reasonable. We therefore took the prior for a and d for all teams to be (independent) Gaussian priors with an average variance $1/37$ found from the estimates. This implies a common loss of memory rate for all teams (Section 2.3). Although we expect the strengths of attack and defence to be (positively) dependent, we choose prior independence for simplicity and to be non-informative. Further, the prior variance is confounded with the τ-parameter for all matches apart from the first in each league (see equation (5)).

The conditional mean and variance for the goal log-intensity are controlled by the parameters τ and γ, and the goal model depends on the mixture parameter ϵ. The predictive properties of the model depend on these parameters. It is tempting from the Gaussian structure in the time model to make use of the conjugate gamma density for the inverse variances (precisions); if the prior for τ is gamma distributed so is the posterior. Our experience with this approach tells us that there is not much information in the match results about the loss of memory rate τ, although the parameter is important for the predictive abilities for the model. We chose therefore τ, γ and ϵ to optimize the predictive ability of the model on historical data. We ran our model on the second half of the four seasons from 1993 to 1997 both for the Premier League and division 1, and predicted successively the second half of each season each round. (We need to use the first half of each season to learn about the various teams.) To quantify the quality of the predictions, we computed the geometrical average of the probabilities for the observed results (home win, draw and away win) for each match played so far. This has a flavour of being a normalized pseudolikelihood measure as it is a product of conditional probabilities. We denote this measure PL. We repeated this procedure for various values of τ, γ and ϵ on a three-dimensional grid. (This was a computationally expensive procedure!) To our surprise, there seems to be a common set of values for the parameters that gave reasonable results overall for both leagues and for all seasons. These values were $\tau = 100$ days, $\gamma = 0.1$ and $\epsilon = 0.2$. (The value of $\tau = 100$ corresponds to a prior variance of $\sigma_0^2 = 1/37$, so the 'optimal' τ is linked to the prior variance through $\sigma_0^2/\tau = 1/3700$.) Although we found parameter values giving higher values of PL for each of the seasons and leagues, we chose to use the common set of parameter values as these simplify both the interpretation of the model and its use.

4. Applications and results

This section contains applications of the model proposed for the Premier League and division 1 for 1997–1998, in prediction, betting and retrospective analysis of the season.

For simplicity we selected values that were uniformly spaced in the interval from -0.2 to 0.2 as the prior means ($\mu_{a,\mathrm{A}}$, $\mu_{\mathrm{A,B}}$ and so on) for the strengths of attack and defence in the Premier League, based on our prior ranking of the teams. A similar procedure was chosen for division 1. Another approach is to use the mean properties (possibly adjusted) from the end of the previous season, at least for those teams that stay in the same league. As the prior mean is only present in the first match for each team, all reasonable prior mean values give approximately the same predictions for the second half of the season. The prior mean values are most important for predicting the first rounds of each season. We further used the parameter values suggested in Section 3.2. In the forthcoming experiments, the number of iterations was checked to give reliable results. We used 100000 iterations in the prediction application and 1 million iterations in the retrospective analysis.

4.1. Prediction and betting

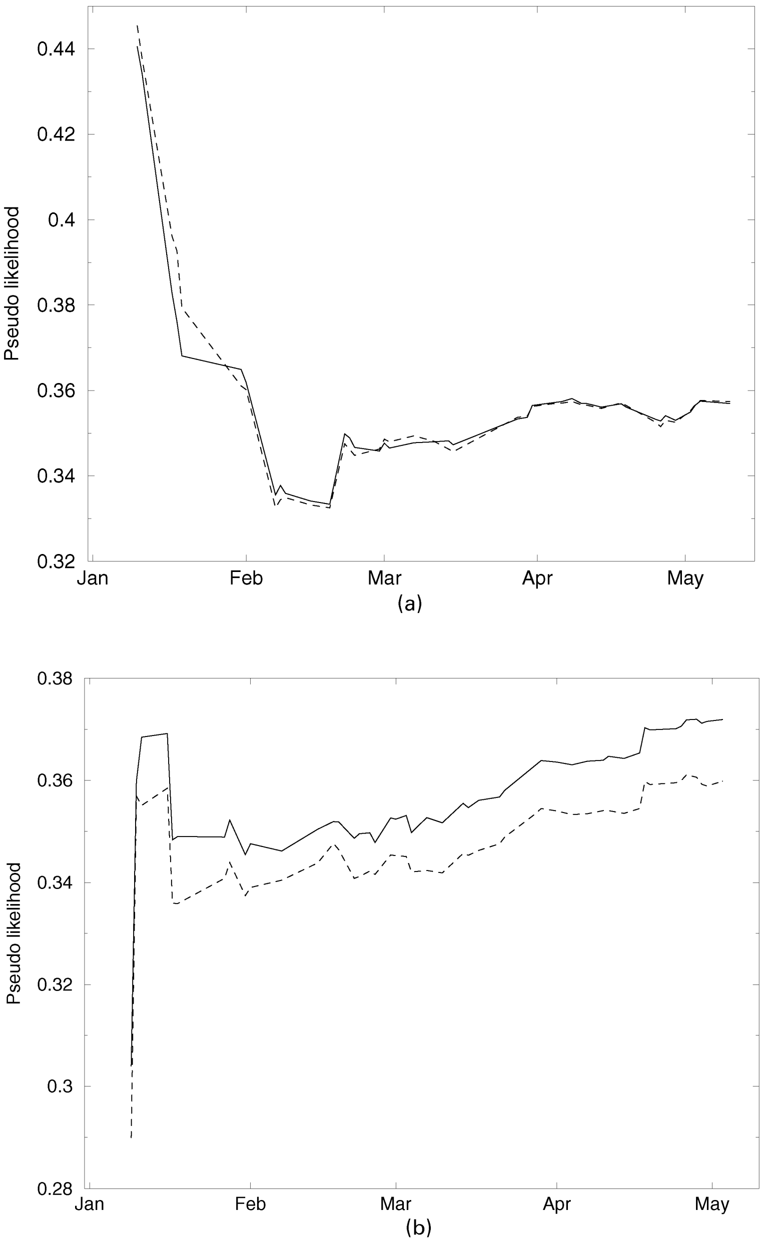
To compare our predictions and to simulate the betting experiments, we used the odds provided by one of the largest international bookmakers operating on the Web, Intertops. Although the firm is officially located in Antigua in the West Indies, it is easily accessible from anywhere in the world via the Internet at `http://www.Intertops.com`. From their odds for the Premier League and division 1, we computed the corresponding (predictive) probabilities for a home win, draw or an away win.

4.1.1. *Predictions*

Fig. 3 shows PL as a function of time for the second half of the 1997–1998 season in the Premier League and division 1, using the first half of each season to learn about the various teams. Both leagues are nearly equally predictable from the bookmaker's point of view, with final PL values of 0.353 and 0.357 for the Premier League and division 1 respectively. Our model does surprisingly well compared with the bookmaker, with final PL values of 0.357 and 0.372 for the Premier League and division 1 respectively. The predictions are especially good for division 1. Recall that the model only makes use of the match results and not all the other information which is available to those who set the bookmaker's odds. It seems that the bookmakers provide better odds for the Premier League than for the lower divisions, which might be natural as the majority of the punters bet on the Premier League.

4.2. Single bets

We can simulate a betting experiment against Intertops by using the above predictions. Assume that $\mathcal{B}$ is the set of matches that we can bet on. Which matches should we bet on and how much? This depends on our utility for betting but, as we decide ourself which matches to bet on, we are



(a)

(b)

Fig. 3. Pseudolikelihood measure for the predictions made by the model (———) and the odds from Intertops (- - - - -) for 1997–1998 for (a) the Premier League and (b) division 1

playing a favourable game as the posterior expected profit is positive. The statement is conditional on a 'correct' model, and a betting experiment will therefore validate our model. For favourable games, Epstein (1995) suggested that we bet on outcomes with a positive expected profit but place the bets so that we obtain a low variance of the profit. In addition to a positive expected profit, this strategy will make the probability of ruin, which is an absorbing state, small. We chose therefore to place the bets to maximize the expected profit while we keep the variance of the profit lower than some limit. An equivalent formulation is to maximize the expected profit minus the variance of the profit, which determines how we should place our bets up to a multiplicative constant. This constant can be found if we choose a specific value or an upper limit for the variance of the profit. Let μ_i^j and $\sigma_i^{2,j}$ be the expected profit and variance for betting a unit amount on outcome i in match j, where $i \in \{\text{home win, draw, away win}\}$. These values are found from the probability p_i^j and odds o_i^j for outcome i in match j. Let β_i^j be the corresponding bet, where for simplicity we restrict ourselves to place no more than one bet for each match. The optimal bets are found to be

$$\arg\max_{\beta_i^j \geqslant 0}[U(\{\beta_i^j\})], \qquad \text{where } U(\{\beta_i^j\}) = E(\text{profit}) - \text{var}(\text{profit}) = \sum_{j \in \mathcal{B}} \beta_i^j(\mu_i^j - \beta_i^j \sigma_i^{2,j}).$$

The solution is $\beta_i^j = \max(0, \mu_i^j/2\sigma_i^{2,j})$, where additionally we choose the outcome i with maximal $\beta_i^j \mu_i^j$ for match j to meet the ‘not more than one bet for each match’ requirement. Fig. 4 shows the profit (scaled to have $\Sigma_j \beta_i^j = 1$ for all bets made so far) using the predictions and odds in Fig. 3, together with an approximate 95% interval given as

$$\text{posterior mean} \pm 2 \text{ posterior standard deviation.}$$

The results are within the upper and lower bound, although the lower bound at the end of the season is negative, still indicating a risk of losing money. For the Premier League the final profit was 39.6% after we won on 15 of a total of 48 bets on 17 home wins, five draws and 26 away wins. For division 1 the final profit was 54.0% after we won on 27 of a total of 64 bets on 30 home wins, six draws and 28 away wins. The final bounds were (−47.2%, 87.8%) and (−29.5%, 58.0%) for the Premier League and division 1 respectively. The β_i^j varied from 0.001 to 0.211 with an average of 0.047 for division 1.

It is not enough to predict the match result just slightly better than the bookmakers to earn money when we bet on single matches. The bookmakers take some percentage of the bets as tax by reducing their odds to less than 1 over their probabilities for a home win, draw and away win, and the odds from Intertops satisfy $1/o_{\text{home}}^j + 1/o_{\text{draw}}^j + 1/o_{\text{away}}^j \approx 1.2$.

The high profit at the end of January for the Premier League is due to a single match, Manchester United *versus* Leicester City on January 31st. Intertops gave high odds for an away win, 13.8, whereas our model predicted a chance of 0.184 for an away win. As Leicester City won 0–1, this bet gave a significant pay-out.

4.2.1. Combination bets

Intertops also provides the opportunity for ‘combination’ bets: bets on the correct results for more than one match simultaneously. We have investigated the profit if we chose this option where we (for simplicity) place our bets on the correct results of three matches simultaneously. The probability for predicting three matches correctly is $p_i^j p_{i'}^{j'} p_{i''}^{j''}$ (approximately only, as the teams' properties are dependent), and this event has odds $o_i^j o_{i'}^{j'} o_{i''}^{j''}$. How should we now place our bets $\beta_{i,i',i''}^{j,j',j''}$? The same argument as for single bets applies: place the bets to maximize the expected profit minus the variance. Although the idea is similar, we now have dependences between the various bets as some matches can be in more than one combination of three matches. Let $\boldsymbol{\beta}$ be the vector of bets and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the vector of expected values and the covariance matrix for all the available combinations with a unit bet. We should place our bets proportionally to

$$\arg\max_{\boldsymbol{\beta} \geqslant 0}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}). \qquad (8)$$

This is a standard quadratic linear programming problem which is easily solved through well-

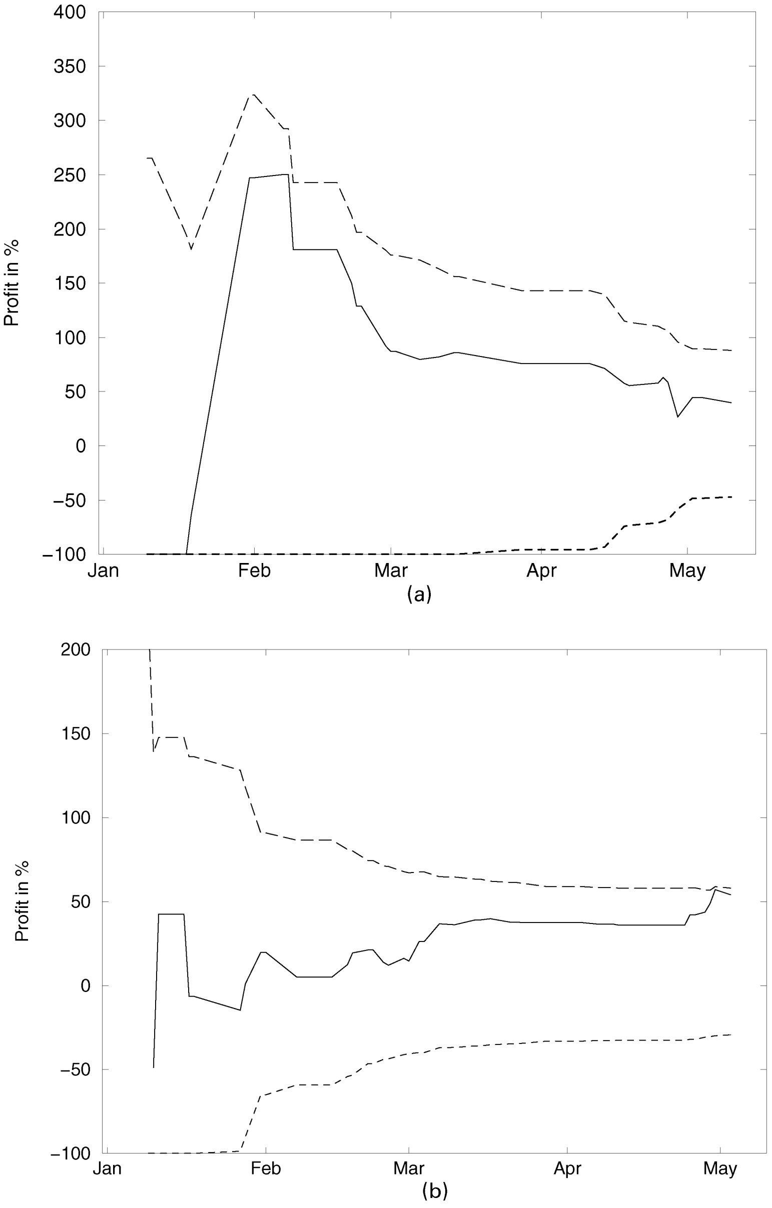


Fig. 4. Observed profit in the simulated betting experiments (on single matches against the odds provided by Intertops) for the 1997–1998 season for (a) the Premier League and (b) division 1: ———, model; - - - - -, lower bound; – – –, upper bound

known algorithms, although the covariance matrix $\mathbf{\Sigma}$ is somewhat tedious to calculate. We choose our candidates only from those outcomes which we bet on in the single-bet case to obtain a reasonable dimension of problem (8). The simulated combination betting experiment gave less satisfying results. The final profits were -100% (140.2%) after 35 bets and 80.3% (109.7%) after 63 bets for the Premier League and division 1, with the posterior standard deviation given in parentheses. If we merge the results for the two divisions, the profit was 9.7% with a large variance compared with the variance that was obtained using single bets. It seems to be both easier and more reliable to bet on single matches compared with combination bets.

4.3. Retrospective analysis of Premier League, 1997–1998

According to the model assumptions the match results in the Premier League for 1997–1998 update information about defending and attacking strengths for all the teams throughout the season. Given this information, it is interesting to know whether Arsenal were lucky to win the Premier League in 1997–1998. Similar questions arise for other teams: were Everton lucky to stay in the league? Did Aston Villa deserve their seventh place? It is easy to provide the answer from the model for such questions using the power of MCMC methods by playing a new match for each of the 380 matches using samples from the *joint* posterior densities for all properties and at all times. By collecting the points and goals scored we can compute a conditional sample of the final ranking. We repeat this procedure after each iteration in the MCMC algorithm and at the end compute the estimates that we are interested in from all the samples. (In this analysis we increased the prior variance for all the properties by a factor of 10, and similarly with τ to keep the conditional variance in equation (5) unchanged. These near non-informative priors make more sense in a retrospective analysis.)

Table 1 shows the estimated (posterior) probabilities for Arsenal, Manchester United, Liverpool and Chelsea to be the first, second and third in the final ranking. Table 1 shows also the observed rank and the number of points achieved. Manchester United had probability 0.433 of winning the league whereas Arsenal had probability 0.247. Liverpool and Chelsea have similar probabilities for being first, second and third. It seems that Arsenal were lucky to win the title from Manchester United.

Fig. 5 gives a more complete picture of the final ranking and displays the expected final rank for each team with approximate 90% (marginal) credibility intervals. The full curve shows the observed ranking. We see from the graph that Everton would have been unlucky to be relegated, and Aston Villa did better than expected. The uncertainty in the final ranking is surprisingly large, and the observed rank seems to be well within the uncertainty bounds. Aston Villa, for example, could easily have been 15th instead of seventh. The managers surely have to face a large uncertainty. It is interesting to note from the graph that the 20 teams divide themselves into four groups: the top four, the upper middle seven, the lower middle seven and the bottom two.

To study more how the top four teams differed and how their skills varied during the season, we compute the (posterior) expected value of the offensive and defensive strengths as a function of time. Fig. 6 shows the result. The differences in defensive skills of the four teams are prominent, whereas their offensive skills are more similar. Manchester United had a good and stable defence whereas their strength of attack decreased somewhat in the second half of the season. Denis Irwin was badly injured in December 1997, and this might be one reason. Later in the season both Ryan Giggs and Nicky Butt suffered from injuries and suspension, causing the

Table 1. Estimated posterior probabilities for each team being the first, second and third in the final ranking in the Premier League, 1997–1998, together with the observed rank and the number of points achieved

Team	*P(1st)*	*P(2nd)*	*P(3rd)*	*Rank*	*Points*
Arsenal	0.247	0.230	0.161	1	78
Manchester United	0.433	0.230	0.131	2	77
Liverpool	0.110	0.151	0.153	3	65
Chelsea	0.095	0.134	0.142	4	63

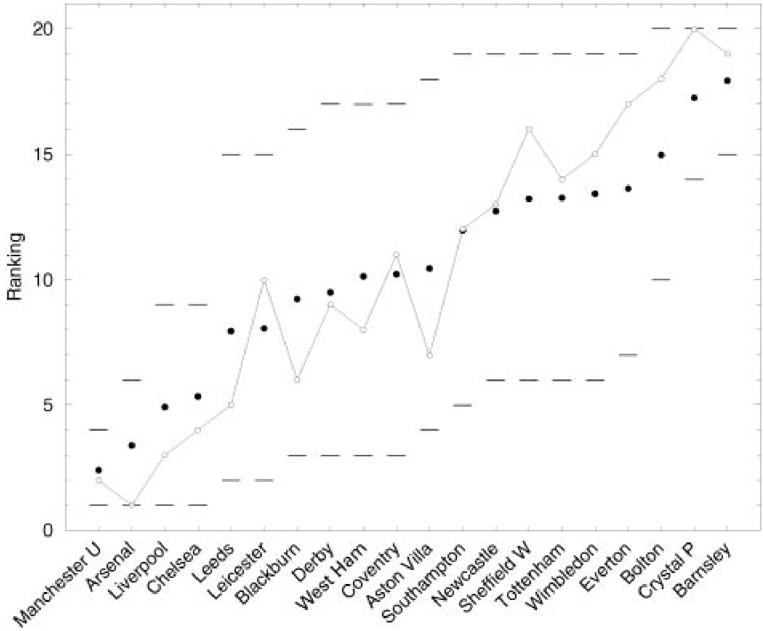


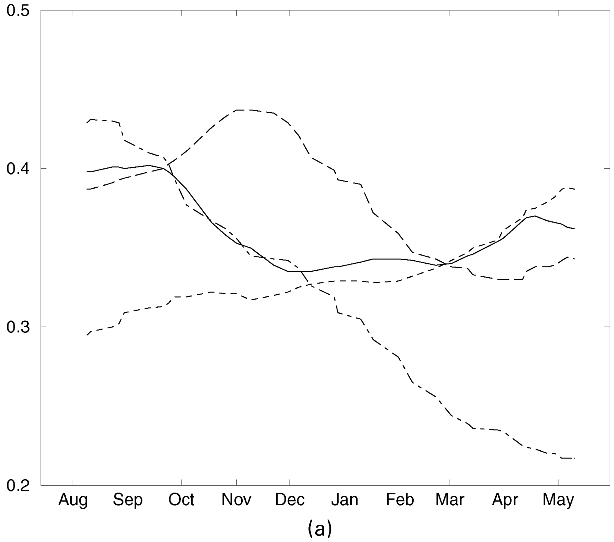
Fig. 5. Posterior expected final rank (●) for each team in the Premier League, 1997–1998, with approximate 90% marginal credibility bounds and the observed ranking (———): note the large uncertainty in the final ranking

attacking strength to decrease. The defensive skills of Arsenal improved during the season whereas their offensive skills were best at the beginning and the end of the season. Manchester United's defensive skills were superior to Arsenal's during the whole season, whereas their offensive skills were somewhat better in the period from October to March. In total, Manchester United seem to be the strongest team. Liverpool and Chelsea had similar and stable defensive qualities, whereas their offensive strength is monotone increasing for Liverpool and monotone decreasing for Chelsea. Arsenal are clearly ranked ahead of Liverpool and Chelsea mainly because of their strong defence. Liverpool are ranked ahead of Chelsea as they had both slightly better defensive and attacking properties on average. However, this is not a sufficient condition in general; which teams they meet at which time is also important.

An amusing application of the model appears when we study the posterior probability for the Bernoulli random variable attached to each match. (These variables indicate which one of the distributions on the right-hand side of equation (3) we currently use; see Appendix A for details.) We ranked each match in the Premier League after the posterior probability for these Bernoulli variables was 1. This probability has the interpretation as the (posterior) probability for that match

Table 2. Five most surprising results in the Premier League, 1997–1998, ranked according to the posterior probability for being unexplainable or an outlier

Match	*Date*	*P(outlier)*	*Result*
Liverpool–Barnsley	November 22nd, 1997	0.76	0–1
Newcastle–Leicester City	November 1st, 1997	0.66	3–3
Wimbledon–Tottenham Hotspur	May 2nd, 1998	0.61	2–6
Sheffield Wednesday–Manchester United	March 7th, 1998	0.60	2–0
Sheffield Wednesday–Arsenal	November 22nd, 1997	0.59	2–0



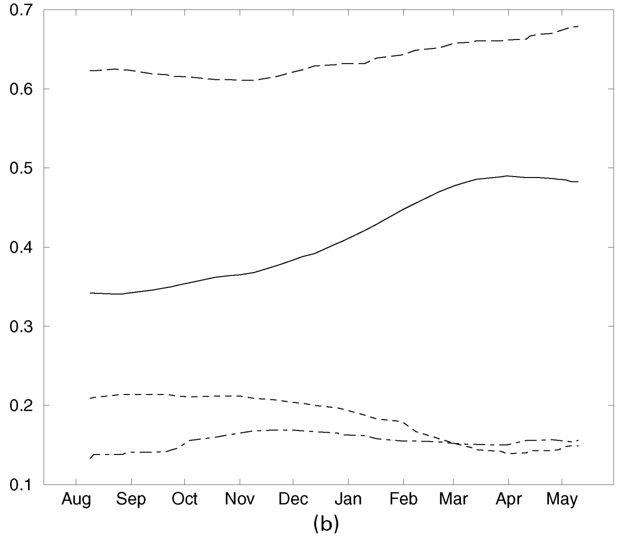


Fig. 6. Retrospective estimates of the mean strengths of (a) attack and (b) defence for the four best teams in the Premier League in the 1997–1998 season: ———, Arsenal; – · – ·, Chelsea; - - - - -, Liverpool; – – –, Manchester United

to be unexplainable or an outlier and hence gives a way of locating those matches that were most surprising, taking both the observed past and the future into account. Table 2 lists the five most surprising results in the Premier League in 1997–1998. The match between Liverpool and Barnsley on November 22nd when Barnsley won 0–1 is the clear winner as the surprise match of the season.

5. Discussion and further work

The model presented seems to capture most of the information contained in the match results and to provide reasonable predictions. The seeming stability for the γ-, τ- and ϵ-parameters across seasons is one confirmation. Although there are more variables than data, we are not in the situation of overfitting the data. We take the (posterior) dependence between the strengths of attack and defence at different time points as various ways to explain the data.

Further, the approach presented seems superior to earlier attempts to model soccer games as it

(a) allows for coherent inference of the properties between the teams in time also,
(b) easily accounts for the joint uncertainty in the variables, which is important in prediction (Draper, 1995),
(c) allows for doing various interesting retrospective analyses of a season and finally
(d) provides a framework where it is easy to change parts or the parameterization in the model.

We do not claim that our parameterization, goal and time model is optimal and cannot be improved on, but that the Bayesian approach presented with MCMC-based inference is promising for these kinds of problem.

There are several points which could and should be improved in the model.

(a) It is of major importance to include more data than just the final match result in the model, but this depends on what kind of data is (easily) available and useful. No attempts have been made along these lines as far as we are aware. This will imply a change in the model as well, but the basic ideas and framework will remain.
(b) Brownian motion is too simple a time model for the team's properties and does not include the first derivative (local trend) in the predictions. A non-stationary time model is needed to capture the *local* behaviour needed for prediction in the near future. An integrated autoregressive process might be suitable if we discretize the time, which is quite reasonable. Such a choice requires (among others) changes in move type (a) in the MCMC algorithm described in Appendix A.
(c) We assumed that all teams have a common loss-of-memory rate τ, and this is a simplification. We have not succeeded in estimating a team-specific τ or found a good way to group each team into a 'high–normal–low' loss-of-memory rate. More observation data than just the final match result are most likely needed to make progress in this direction.
(d) The goal model could be improved on. The birth process approach of Dixon and Robinson (1998) is natural and interesting, although we should estimate coherently the goal model simultaneously with the time-varying properties. Further, various parameterizations like the inclusion of the psychological effect and the idea of a mixture model need to be investigated further within their birth process framework.
(e) Each team's (constant) home ground advantage is a natural variable to include in the model. We did not find sufficient support from the match results to include this at the current stage, but hopefully more data will change this.

It seems that the statistical community is making progress in understanding the art of predicting soccer matches, which is of vital importance for two reasons:

(a) to demonstrate the usefulness of statistical modelling and thinking on a problem that many people *really* care about and
(b) to make us all rich through betting!

Acknowledgements

We are grateful to the Associate Editor and the two referees for valuable comments that helped us to improve the paper, and especially to the referee who spotted a (serious) mistake in the original presentation of the MCMC algorithm. Thanks also go to our colleagues Georg Elvebakk and Bo Lindqvist for valuable discussions on statistical aspects of soccer.

Appendix A: Details of the Markov chain Monte Carlo algorithm

This appendix contains details of our MCMC implementation sketched in Section 3.1. To ease the implementation of the mixture model in equation (6), we shall use an equivalent reformulation: define $\delta_{\mathrm{A,B}}$ as an independent Bernoulli variable which is 1 with probability ϵ, and define

$$\pi_{\mathrm{g2}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}, \delta_{\mathrm{A,B}}) = \begin{cases} \pi^*_{\mathrm{g1}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}) & \text{if } \delta_{\mathrm{A,B}} = 0, \\ \pi^*_{\mathrm{g1}}\{x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\exp(c^{(x)}), \exp(c^{(y)})\} & \text{if } \delta_{\mathrm{A,B}} = 1. \end{cases} \tag{9}$$

Then with obvious notation

$$\pi_{\mathrm{g}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}) = E_{\delta_{\mathrm{A,B}}}\{\pi_{\mathrm{g2}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}}, \delta_{\mathrm{A,B}})\}.$$

Thus, we can attach one Bernoulli variable to each match and update these variables in the MCMC algorithm also. We ignore their values in the output analysis where we consider only those components of $\boldsymbol{\theta}$ that are of interest to us. This yields a correct procedure as the marginal distribution for $\boldsymbol{\theta}$ remains unchanged when we include the Bernoulli variables.

Because of the reformulation of the mixture distribution, we need three different types of move to implement an MCMC algorithm for our model:

(a) update the properties for each team every time that there is a match,
(b) update the match result for each unobserved match and
(c) update the Bernoulli variable for each match.

In each full sweep we visit all unobserved (stochastic) variables in a random order and update each one using the appropriate type of move.

A.1. Move type (a): updating one of the properties

We describe only how we update the strengh of attack $a^{t''}_{\mathrm{A}}$ for team A at time t'' by using a Metropolis step, as the update of the strength of defence is similar. Note that all other variables remain constant when we propose an update for $a^{t''}_{\mathrm{A}}$. We assume that team A plays a match against team B at time t'' and at A's home ground, as the acceptance rate when A plays on B's home ground is similar with obvious changes. Let t' and t''' be the times of the previous and following matches for team A. We shall return to the case when there is no previous and/or following match. Denote by $(x^{t''}_{\mathrm{A,B}}, y^{t''}_{\mathrm{A,B}})$ and $\delta^{t''}_{\mathrm{A,B}}$ the (current, if not observed) number of goals in the match and the Bernoulli variable attached to that match respectively.

We sample first a new proposal for $a^{t''}_{\mathrm{A}}$ from a Gaussian (symmetric) distribution, $a^{t'',\mathrm{new}}_{\mathrm{A}} \sim N(a^{t''}_{\mathrm{A}}, \sigma^2_q)$, where σ^2_q is a fixed constant for all teams, attack and defence. For all our examples in Section 4, we used $\sigma^2_q = 0.05^2$. The new proposal is accepted with probability min(1, R), where

$$R = \frac{\pi(a^{t'',\mathrm{new}}_{\mathrm{A}}|a^{t'}_{\mathrm{A}})}{\pi(a^{t''}_{\mathrm{A}}|a^{t'}_{\mathrm{A}})} \frac{\pi(a^{t'''}_{\mathrm{A}}|a^{t'',\mathrm{new}}_{\mathrm{A}})}{\pi(a^{t'''}_{\mathrm{A}}|a^{t''}_{\mathrm{A}})} \frac{\pi_{\mathrm{g2}}(x^{t''}_{\mathrm{A,B}}, y^{t''}_{\mathrm{A,B}}|\lambda^{(x),t'',\mathrm{new}}_{\mathrm{A,B}}, \lambda^{(y),t'',\mathrm{new}}_{\mathrm{A,B}}, \delta^{t''}_{\mathrm{A,B}})}{\pi_{\mathrm{g2}}(x^{t''}_{\mathrm{A,B}}, y^{t''}_{\mathrm{A,B}}|\lambda^{(x),t''}_{\mathrm{A,B}}, \lambda^{(y),t''}_{\mathrm{A,B}}, \delta^{t''}_{\mathrm{A,B}})}; \tag{10}$$

otherwise we remain in the old state. In equation (10), $\pi(a^{t'',\mathrm{new}}_{\mathrm{A}}|a^{t'}_{\mathrm{A}})$ denotes the conditional Gaussian density for $a^{t''}_{\mathrm{A}}$ given $a^{t'}_{\mathrm{A}}$ evaluated at $a^{t'',\mathrm{new}}_{\mathrm{A}}$. Further $\lambda^{(x),t'',\mathrm{new}}_{\mathrm{A,B}}$ is computed from equation (1) using the proposed new value $a^{t'',\mathrm{new}}_{\mathrm{A,B}}$ and so on. If there is no previous match, then $\pi(a^{t'',\mathrm{new}}_{\mathrm{A}}|a^{t'}_{\mathrm{A}})$ and $\pi(a^{t''}_{\mathrm{A}}|a^{t'}_{\mathrm{A}})$ are replaced with the prior density for a_{A} evaluated at $a^{t'',\mathrm{new}}_{\mathrm{A}}$ and $a^{t''}_{\mathrm{A}}$ respectively. If there is no following match, then we remove $\pi(a^{t'''}_{\mathrm{A}}|a^{t'',\mathrm{new}}_{\mathrm{A}})$ and $\pi(a^{t'''}_{\mathrm{A}}|a^{t''}_{\mathrm{A}})$ in the expression for R.

We prefer this simple Metropolis step rather than the more elegant proposal density found from computing the Gaussian approximation to the conditional density, by a second-order Taylor expansion of the log-conditional-density around current values. Although the acceptance rate with a tailored Gaussian proposal increases to well over 90%, it does not seem to be worth the additional computation and implementational costs.

A.2. Move type (b): updating a match result

We update the match result using a Gibbs step, thus drawing $(x^{t''}_{\mathrm{A,B}}, y^{t''}_{\mathrm{A,B}})$ from the conditional distribution in equation (9). The modifications needed because of truncation and $\kappa(x^{t''}_{\mathrm{A,B}}, y^{t''}_{\mathrm{A,B}}|\lambda^{(x),t''}_{\mathrm{A,B}}, \lambda^{(y),t''}_{\mathrm{A,B}})$ are easily done by rejection steps. The algorithm is as follows.

Step 1: if $\delta^{t''}_{\mathrm{A,B}}$ is 1, then set

$$\lambda^{(x)} = \exp(c^{(x)})$$

and

$$\lambda^{(y)} = \exp(c^{(y)});$$

otherwise set

$$\lambda^{(x)} = \lambda^{(x),t''}_{\mathrm{A,B}}$$

and

$$\lambda^{(y)} = \lambda^{(y),t''}_{\mathrm{A,B}}.$$

Step 2: draw x from $\mathrm{Po}(x|\lambda^{(x)})$ until $x \leqslant 5$, and then draw y from $\mathrm{Po}(y|\lambda^{(y)})$ until $y \leqslant 5$.
Step 3: with probability

$$\frac{\kappa(x, y|\lambda^{(x)}, \lambda^{(y)})}{\max(1.1, 1 + 0.1\lambda^{(x)}\lambda^{(y)})}$$

set $x^{t''}_{\mathrm{A,B}} = x$ and $y^{t''}_{\mathrm{A,B}} = y$ and return; otherwise go back to step 2.

A.3. Move type (c): updating a Bernoulli variable

The Bernoulli variable attached to each match, $\delta^{t''}_{\mathrm{A,B}}$, is a random variable with probability ϵ to be 1; denote the law by $\pi(\delta^{t''}_{\mathrm{A,B}})$. We propose always to flip the current value of $\delta^{t''}_{\mathrm{A,B}}$ to $\delta^{t'',\mathrm{new}}_{\mathrm{A,B}} = 1 - \delta^{t''}_{\mathrm{A,B}}$ which is accepted with probability $\min(1, R)$,

$$R = \frac{\pi_{\mathrm{g2}}(x^{t''}_{\mathrm{A,B}}, y^{t''}_{\mathrm{A,B}}|\lambda^{(x),t''}_{\mathrm{A,B}}, \lambda^{(y),t''}_{\mathrm{A,B}}, \delta^{t'',\mathrm{new}}_{\mathrm{A,B}})}{\pi_{\mathrm{g2}}(x^{t''}_{\mathrm{A,B}}, y^{t''}_{\mathrm{A,B}}|\lambda^{(x),t''}_{\mathrm{A,B}}, \lambda^{(y),t''}_{\mathrm{A,B}}, \delta^{t''}_{\mathrm{A,B}})} \frac{\pi(\delta^{t'',\mathrm{new}}_{\mathrm{A,B}})}{\pi(\delta^{t''}_{\mathrm{A,B}})}.$$

A.4. A comment on implementation

The model is easiest to programme as a graph model (see Fig. 2) parsing matches with teams, dates, etc. from external input files. The MCMC algorithm described can be modified in several ways to achieve a significant acceleration. Our approach was to tabulate the truncated Poisson distribution for a large set of λs, and then to use a table look-up with interpolation to obtain values. The normalization constants for the joint density $\pi^*_{\mathrm{g1}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}})$ is also needed as a function of $(\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}})$, which we once more tabulate. It is also possible to tabulate $\pi_{\mathrm{g}}(x_{\mathrm{A,B}}, y_{\mathrm{A,B}}|\lambda^{(x)}_{\mathrm{A,B}}, \lambda^{(y)}_{\mathrm{A,B}})$ directly if ϵ is the same for all matches (as in our examples). In this case we can avoid the above reformulation of the mixture distribution using the Bernoulli variable attached to each match. However, this prohibits the analysis of the most surprising matches presented in Table 2.

A.5. A comment on Rao–Blackwellization

To predict a future match, A against B say, it is natural to consider the simulated result $(x_{\mathrm{A,B}}, y_{\mathrm{A,B}})$ of that match and to estimate the probability that A beats B by counting how many times $x_{\mathrm{A,B}}$ is greater than $y_{\mathrm{A,B}}$ and dividing by the total number. However, we can decrease the variance of this estimator by Rao–

Blackwellization: we compute Pr(A beats B$|\mathbf{e}_\mathrm{A}$, $\mathbf{e}_\mathrm{B}$) and use the empirical mean of this conditional probability as our estimate for the probability that A beats B. (Again, we tabulate these probabilities and use a table look-up with interpolation.) We refer to Liu *et al.* (1994) for the theoretical background of Rao–Blackwellization in this context.

A.6. A comment on block updating

The structure of the full model makes it possible to use block updates of the properties of each team within the MCMC algorithm. The first natural choice is to update the strength of attack, say, for team A at all time points $t_0, \ldots, t_{n-1}$ where team A plays a match, $\mathbf{a}_\mathrm{A} = (a_\mathrm{A}^{t_0}, \ldots, a_\mathrm{A}^{t_{n-1}})^\mathrm{T}$. *A priori*, $\mathbf{a}_\mathrm{A}$ is a (non-stationary) Gaussian vector (with length n) with a Markov property, and it is then tempting to update $\mathbf{a}_\mathrm{A}$ by using a Gaussian proposal in the MCMC algorithm found from a second-order Taylor expansion of the log-full-conditional of $\mathbf{a}_\mathrm{A}$ around the current state,

$$\log\{\pi(\mathbf{a}_\mathrm{A}|\text{the rest})\} = C + \sum_{i=0}^{n-1} [\log\{\pi(a_\mathrm{A}^{t_i}|a_\mathrm{A}^{t_{i-1}})\} + \log\{\pi(x_{\mathrm{A,B}_i}^{t_i}, y_{\mathrm{A,B}_i}^{t_i}|\lambda_{\mathrm{A,B}_i}^{(x)}, \lambda_{\mathrm{A,B}_i}^{(y)})\}]$$

$$\approx C' + \mathbf{b}^\mathrm{T}\mathbf{a}_\mathrm{A} - \tfrac{1}{2}\mathbf{a}_\mathrm{A}^\mathrm{T}\mathbf{Q}\mathbf{a}_\mathrm{A}. \tag{11}$$

Here, $\pi(a_\mathrm{A}^{t_0}|a_\mathrm{A}^{t_{-1}})$ is the prior, C and C' are (arbitrary) constants, $\mathbf{b}$ is an $n \times 1$ vector and $\mathbf{Q}$ a symmetric $n \times n$ tridiagonal (precision) matrix. We may also choose to update ($\mathbf{a}_\mathrm{A}$, $\mathbf{d}_\mathrm{A}$) or other choices simultaneously. Taylor expansion is not necessarily the best choice in terms of acceptance and the length of accepted moves, Rue (2000) suggests a variant for the univariate case (which is not immediate to use in this case). For notational convenience we supposed in equation (11) that team A played only home matches against team B_i at time t_i. Note that the computation of equation (11) requires only $\mathcal{O}(n^2)$ flops owing to the Markov property. Assume that $\mathbf{Q}$ is positive definite; then we can sample (and compute the normalization constant) from the Gaussian proposal with a Markov property using only $\mathcal{O}(n)$ flops (see Rue (2000) for details). The algorithm runs as follows. Compute the band Cholesky factorization of $\mathbf{Q} = \mathbf{L}\mathbf{L}^\mathrm{T}$, where $\mathbf{L}$ has lower bandwidth equal to 1 owing to the Markov property of $\mathbf{a}_\mathrm{A}$. Solve, by forward and backward substitution, $\mathbf{L}\mathbf{u} = \mathbf{b}$ and then $\mathbf{L}^\mathrm{T}\boldsymbol{\mu} = \mathbf{u}$, to obtain the mean $\boldsymbol{\mu}$ of the proposal density. We can now sample $\mathbf{a}_\mathrm{A}^\mathrm{new}$ from the Gaussian proposal density by first solving by backward substitution $\mathbf{L}^\mathrm{T}\mathbf{v} = \mathbf{z}$, where $\mathbf{z}$ is a vector with independent standard Gaussian variates, and then add the mean, $\mathbf{a}_\mathrm{A}^\mathrm{new} = \boldsymbol{\mu} + \mathbf{v}$. High performance implementations of all the numerical algorithms required exist in the freely available LAPACK library (Anderson *et al.*, 1995). The computation of the acceptance rate for this block update is straightforward after computing approximation (11) around $\mathbf{a}_\mathrm{A}^\mathrm{new}$ to obtain the proposal density of the reverse move from $\mathbf{a}_\mathrm{A}^\mathrm{new}$ to $\mathbf{a}_\mathrm{A}$.

Shephard and Pitt (1997) used similar ideas for non-Gaussian measurement time series with good results, and they reported simulation experiments studying the efficiency of using block updates in the MCMC algorithm. The efficiency of blocking compared with single-site updating can be large if $\mathbf{a}_\mathrm{A}$ is highly dependent. For our application, the dependence is not so high that single-site updating does not perform reasonably well. However, an experimental study using block updates of $\mathbf{a}_\mathrm{A}$ or, even better, $(\mathbf{a}_\mathrm{A}, \mathbf{d}_\mathrm{A})^\mathrm{T}$ simultaneously would be interesting. Similar linearization ideas like equation (11) could be used more frequently as (very) fast sampling algorithms (including an evaluation of the normalization constant) exist for Gaussian proposals (also on lattices) with a Markov property; see Rue (2000) for details about the algorithm.

References

Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., Croz, J. D., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S. and Sorensen, D. (1995) *LAPACK Users' Guide*, 2nd edn. Philadelphia: Society for Industrial and Applied Mathematics.

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

Dixon, M. J. and Robinson, M. E. (1998) A birth process model for association football matches. *Statistician*, **47**, 523–538.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.

Epstein, R. A. (1995) *The Theory of Gambling and Statistical Logic*. New York: Academic Press.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Kuonen, D. (1996) Statistical models for knock-out soccer tournaments. *Technical Report*. Department of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne.

Lee, A. J. (1997) Modeling scores in the Premier League: is Manchester United really the best? *Chance*, **10**, 15–19.

Liu, J. S., Wong, W. H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

Ridder, G., Cramer, J. S. and Hopstaken, P. (1994) Down to ten: estimating the effect of a red card in soccer. *J. Am. Statist. Ass.*, **89**, 1124–1127.

Rue, H. (2000) Fast sampling of gaussian markov random fields with applications. *Statistics Report 1*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.

Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.

West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.