The Statistician (2002)
51, *Part* 2, *pp.* 157–168

Dynamic modelling and prediction of English Football League matches for betting

Martin Crowder,

Imperial College of Science, Technology and Medicine, London, UK

Mark Dixon

City University, London, UK

and Anthony Ledford and Mike Robinson

Man Investment Products Ltd, London, UK

[Received July 2001. Revised November 2001]

Summary. We focus on modelling the 92 soccer teams in the English Football Association League over the years 1992–1997 using refinements of the independent Poisson model of Dixon and Coles. Our framework assumes that each team has attack and defence strengths that evolve through time (rather than remaining constant) according to some unobserved bivariate stochastic process. Estimation of the teams' attack and defence capabilities is undertaken via a novel approach involving an approximation that is computationally convenient and fast. The results of this approximation compare very favourably with results obtained through the Dixon and Coles approach. We note that the full model (i.e. the model before the above approximation is made) may be implemented using Markov chain Monte Carlo procedures, and that this approach is vastly more computationally expensive. We focus on the probabilities of home win, draw or away win because these outcomes constitute the primary betting market. These probabilities are estimated for games played between any two of the 92 teams and the predictions are compared with the actual results.

Keywords: Attack and defence strengths; Football results; Poisson models; State space modelling

1. Team talk

Consider the match between team i (home) and team j (away) played at time t, in which the goals scored are x_{it} (by team i) and y_{jt} (by team j). According to the independent Poisson model, the probability of this result is

$$\{\exp(-\eta\alpha_{it}\beta_{jt})(\eta\alpha_{it}\beta_{jt})^{x_{it}}/x_{it}!\}\{\exp(-\alpha_{jt}\beta_{it})(\alpha_{jt}\beta_{it})^{y_{jt}}/y_{jt}!\}, \tag{1.1}$$

where α_{it} and α_{jt} represent the attack rates (strengths) of teams i and j at time t, β_{it} and β_{jt} their defence rates (weaknesses) and η is the home advantage factor. With increasing η, α_{it} and β_{jt}, more goals will tend to be scored by team i against team j; likewise, there will be more goals at the other end when $\alpha_{jt}\beta_{it}$ increases. Though the model describes goals scored by the

Address for correspondence: Martin Crowder, Department of Mathematics, Huxley Building, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London, SW7 2BZ, UK.
E-mail: m.crowder@ic.ac.uk

 0039–0526/02/51157

teams, this is simply a route to estimating the probabilities of win, draw or lose for the primary football betting market.

Maher (1982) took the αs and βs to be constant over time, so each team is assigned a pair of fixed parameters, α_i and β_i. Dixon and Coles (1997) adjusted the model to reflect a dependence between x_{it} and y_{jt} at the lower end (when x_{it} and y_{jt} are both less than 2). Also, with prediction in mind, they tapered the likelihood function to give greater weight to more recent results. This is in recognition of the probable need to allow α_i and β_i to vary through time, not to be constrained to stay constant.

Recently, three papers have appeared with a theme that is similar to ours. Where we take the goals scored as the data, the observations in Knorr-Held (2000) and Koning (2000) comprise only the win, lose or draw results. In addition, their models are based on a single team strength, rather than separating attack and defence, and are based on thresholds in cumulative distribution functions, in the usual way for ordered categorical data. Both Knorr-Held (2000) and Koning (2000) derived point estimates for the parameters, Koning applying maximum likelihood and Knorr-Held using an extended Kalman filter together with an *ad hoc* method for a variance parameter.

Rue and Salvesen (2000) used a modified independent Poisson model, as proposed by Dixon and Coles (1997). They allowed for separate attack and defence strengths but incorporated these in a further-modified way that involves a psychological factor to reflect the overall difference in the opposing teams' strengths. Also, they truncated the numbers of goals at 5, so a score of 9–6 is interpreted as 5–5, for example. In a final modification, a mixture model was proposed in which the previously described form is mixed with a similar form based on average scores. There was no allowance for a home ground advantage, however.

In both Knorr-Held (2000) and Rue and Salvesen (2000) the team strength parameters were allowed to vary over time according to random walks with independent normal increments. This aspect is similar to our formulation, given in Section 3, but in other respects our treatment and aims are quite different from those of Knorr-Held and Rue and Salvesen.

2. Constraints on team behaviour

The αs and βs only appear in the likelihood function (1.1) in combinations $\alpha_{it}\beta_{jt}$. In consequence, there is a lack of identifiability; for example, the αs could be doubled and the βs halved without altering their products. A standard approach in such cases is to apply one or more constraints to the parameters. However, any such constraint has the undesirable effect of tying their values together in some way. Thus, if one team's (α, β) changes, as a result of some local event, say, other teams' (α, β)s must change to preserve the constraints, even though the event in question does not affect them. This consideration suggests that any particular set of (α, β)s does not itself necessarily constitute the 'true' team strengths: the values are merely the projections of the strengths onto the constraint surface along trajectories of constant likelihood.

It is not immediately obvious how many constraints will be needed on the (α, β)s to obtain identifiability, though it is clearly at least one because of the scaling ambiguity mentioned above. With m teams, there are $2m$ αs and βs, and $m(m-1)$ combinations $\alpha_i\beta_j$. Thus, with enough data, i.e. sufficient hypothetical replication of the matches, to fix the combinations, there are $m(m-1)$ equations in $2m$ unknowns. For example, with $m = 4$ this is 12 equations in eight unknowns, which would seem to imply inconsistency even without any constraints. However, the 12 equations are not necessarily independent, and we must identify the extent of this dependence to calculate the number of constraints required. It is shown in Appendix A.1 that a single constraint will suffice for identifiability.

We shall take the single constraint to be

$$\sum_{i=1}^{m} \log(\alpha_{it}\beta_{it}) = 0, \tag{2.1}$$

i.e. that the geometric mean of the $(\alpha_{it}\beta_{it})$ is 1. Note that $\alpha_{it}\beta_{it}$ is the average scoring rate of team i's attack against its own defence per game. Constraint (2.1) will be imposed by expressing the α_{it} and β_{it} in terms of a more basic set of unconstrained parameters. Let these underlying parameters be $\gamma_{it} = (\gamma_{\alpha_{it}}, \gamma_{\beta_{it}})^{\mathrm{T}}$, and take

$$\begin{aligned} \log(\alpha_{it}) &= \gamma_{\alpha_{it}} - \bar{\gamma}_t, \\ \log(\beta_{it}) &= \gamma_{\beta_{it}} - \bar{\gamma}_t, \end{aligned} \tag{2.2}$$

where

$$\bar{\gamma}_t = (2m)^{-1} \sum_{j=1}^{m} (\gamma_{\alpha_{jt}} + \gamma_{\beta_{jt}}).$$

Thus, constraint (2.1) is automatically satisfied.

3. A stochastic process tactic

Dixon and Coles (1997) recognized the probable need for the αs and βs to vary over time. One way of tackling such a drift, alternative to theirs, is to model the processes explicitly. The changes over time may be small, for consistently performing teams, or larger, reflecting more substantial events, planned or not. We shall implement this by adopting an autoregressive AR(1) process for $\gamma_{it} = (\gamma_{\alpha_{it}}, \gamma_{\beta_{it}})^{\mathrm{T}}$:

$$\gamma_{it} - \gamma_{i0} = R(\gamma_{i,\,t-1} - \gamma_{i0}) + u_{it}. \tag{3.1}$$

Thus, the stochastic process moves in the γ-space, of $2m$ dimensions, and then, by equation (2.2), this is projected into the (α, β) space, of $2m - 1$ dimensions. In equation (3.1) R is a 2×2 matrix of autoregression parameters,

$$R = \begin{pmatrix} \rho_{\alpha\alpha} & \rho_{\alpha\beta} \\ \rho_{\beta\alpha} & \rho_{\beta\beta} \end{pmatrix},$$

the u_{it} are independent $N_2(0,\Sigma)$ innovations (independent of all previous outcomes and all other u_{jt}s) and γ_{i0} is a base-line value towards which γ_{it} is drawn if R is small in some sense (e.g. $\mathrm{tr}(R^{\mathrm{T}}R) < 1$, which ensures that $|Rx| < |x|$ for all x).

Because of the independence of the u_{it}, the γ_{it}-processes for different teams evolve over time independently of one another. However, dependence between the αs and βs for different teams is introduced via equation (2.2). These processes are not directly observable, but their effects are seen in the match results. Special cases of equation (3.1) include the following. If the γ_{i0} are all 0, the base-line attraction aspect is absent and the processes are purely autoregressive. If R and Σ are both diagonal the processes $\gamma_{\alpha_{it}}$ (attack) and $\gamma_{\beta_{it}}$ (defence) evolve independently. If $R = I$, the unit matrix, the γ_{it}-process is just a two-dimensional random walk with steps u_{it}. If $R = I$ and $\Sigma = 0$, the zero matrix, γ_{it} is constant over time as in the Maher (1982) model.

Even if we take the γ_{i0} that appear explicitly in equation (3.1) as 0 we still need to specify initial values because for $t = 1$ the formula involves γ_{i0}. These values can be defined with reference to the previous season's results, or perhaps as random effects from some suitable distribution, or retained as fixed effects parameters to be estimated.

An overall likelihood function can be constructed for matches up to and including time t as follows. The vector of unknown parameters is $\theta = (\eta, R, \Sigma)$, of dimension $1 + 4 + 3 = 8$. Let G_t comprise the goal results of all the matches played at time t, i.e. all the x_{it} and y_{jt} for i and j ranging over all the teams involved, and let Γ_t represent the set of unobserved γ_{it} at time t for these teams. Then

$$p(\Gamma_1, \ldots, \Gamma_t | \Gamma_0, \theta) = \prod_{s=1}^{t} p(\Gamma_s | \Gamma_{s-1}, \theta) = \prod_{s=1}^{t} \prod_{i} p(\gamma_{is} | \gamma_{i,\, s-1}, \theta) \tag{3.2}$$

and

$$p(G_1, \ldots, G_t | \Gamma_0, \ldots, \Gamma_t, \theta) = \prod_{s=1}^{t} p(G_s | \Gamma_s, \eta) = \prod_{s=1}^{t} \prod_{ij} p(x_{is}, y_{js} | a_{is}, a_{js}, \eta) \tag{3.3}$$

where $a_{is} = (\log(\alpha_{is}), \log(\beta_{is}))^{\mathrm{T}}$ and the product Π_{ij} in equation (3.3) is over all the games played at time s. It has been assumed in equation (3.3) that, given $(\alpha_{is}, \beta_{is}, \eta)$, the result (x_{is}, y_{js}) is independent of all others. Hence, the likelihood function for θ based on the matches up to time t, conditional on Γ_0, is

$$\begin{aligned} L_t(\theta) &= p(G_1, \ldots, G_t | \Gamma_0, \theta) \\ &= \int p(G_1, \ldots, G_t, \Gamma_1, \ldots, \Gamma_t | \Gamma_0, \theta)\, \mathrm{d}\Gamma_1 \ldots \mathrm{d}\Gamma_t \\ &= \int p(G_1, \ldots, G_t | \Gamma_0, \ldots, \Gamma_t, \theta)\, p(\Gamma_1, \ldots, \Gamma_t | \Gamma_0, \theta)\, \mathrm{d}\Gamma_1 \ldots \mathrm{d}\Gamma_t \\ &= \int \prod_{s=1}^{t} \{ \prod_{ij} p(x_{is}, y_{js} | a_{is}, a_{js}, \eta) \prod_{i} p(\gamma_{is} | \gamma_{i,\, s-1}, \theta) \}\, \mathrm{d}\Gamma_1 \ldots \mathrm{d}\Gamma_t, \end{aligned} \tag{3.4}$$

using equations (3.2) and (3.3). The two core ingredients in the integrand of equation (3.4) are straightforward to compute: $p(x_{is}, y_{js} | a_{is}, a_{js}, \eta)$ is given as the Poisson product in expression (1.1) and $p(\gamma_{is} | \gamma_{i,\, s-1}, \theta)$ as the density of $N\{\gamma_{i0} + R(\gamma_{i,\, t-1} - \gamma_{i0}), \Sigma\}$, from equation (3.1). However, the dimension of the integration is $2mt$, m being the number of teams involved. For instance, with $m = 92$, the number of teams in the English Football League, and $t = 40$, roughly the number of games played in one season, $2mt = 7360$. This makes equation (3.4) difficult to deal with directly.

Markov chain Monte Carlo (MCMC) methods were initially applied to the present problem but, for brevity, they are not reported here. In short, a form of MCMC sampling was employed using a Gibbs-type method with the components, γ_{is} and γ_{js} of $(\Gamma_1, \ldots, \Gamma_t, \theta)$, being updated one at a time via the usual Metropolis acceptance formula. By virtue of the Markov property of the γ_{it}-processes, seen in equation (3.2), and of the dependence of (x_{is}, y_{js}) on $(\Gamma_1, \ldots, \Gamma_t, \theta)$ only through (a_{is}, a_{js}, η), γ_{is} and γ_{js} each appear in only three individual factors of the full posterior expression. Thus, a significant short-cut can be made in computing the acceptance probability for proposal values of γ_{is}. On this basis the computations for predictions for a single week become feasible. However, for comparisons over whole seasons, as required for the present study, the MCMC method is still too slow to be used routinely. For this reason a faster method was sought, and such an approach is described next.

4. An approximation

The model here is essentially a non-normal, non-linear state space model in which the states are the Γ_{it} and the observed quantities are the x_{it} and y_{it}. Many references have presented

approximate methods for filtering and smoothing for such models, e.g. Kitagawa (1987), Carlin *et al*. (1992), Carter and Kohn (1994) and Shephard and Pitt (1997). Some of the methods proposed are based on numerical integration and others on simulation, but what they all have in common is the retention of the underlying model formulation. The approach to be outlined in this section differs in that the original model is replaced by a derived model that is easier to handle. The attitude here is that no model is likely to be 'true', however carefully constructed, and the only true worth of any model lies in its predictive performance: in the present context this is its ability to predict the outcomes of matches so that useful bets can be identified and placed.

It is shown in Appendix A.1 that, in consequence of equations (3.1) and (2.2), the $(\alpha_{it}, \beta_{it})$ process is given by

$$a_{it} - a_{i0} = R(a_{i,t-1} - a_{i0}) + (RJ - JR)(\gamma_{t-1} - \gamma_0) + u_{it} - (2m)^{-1} \sum_{j=1}^{m} Ju_{jt}, \tag{4.1}$$

where $a_{it} = (\log(\alpha_{it}),\ \log(\beta_{it}))^{\mathrm{T}}$ and J is a 2×2 matrix of 1s. Note that

$$\begin{aligned} RJ - JR &= \begin{pmatrix} \rho_{\alpha\alpha} + \rho_{\alpha\beta} & \rho_{\alpha\alpha} + \rho_{\alpha\beta} \\ \rho_{\beta\alpha} + \rho_{\beta\beta} & \rho_{\beta\alpha} + \rho_{\beta\beta} \end{pmatrix} - \begin{pmatrix} \rho_{\alpha\alpha} + \rho_{\beta\alpha} & \rho_{\alpha\beta} + \rho_{\beta\beta} \\ \rho_{\alpha\alpha} + \rho_{\beta\alpha} & \rho_{\alpha\beta} + \rho_{\beta\beta} \end{pmatrix} \\ &= \begin{pmatrix} \rho_{\alpha\beta} - \rho_{\beta\alpha} & \rho_{\alpha\alpha} - \rho_{\beta\beta} \\ \rho_{\beta\beta} - \rho_{\alpha\alpha} & \rho_{\beta\alpha} - \rho_{\alpha\beta} \end{pmatrix}. \end{aligned}$$

Thus, $RJ - JR$ is 0 if $R = I$, as for the random walk model. However, in that case equation (3.1) would represent a non-stationary autoregressive model. More generally, $RJ - JR$ is 0 if R is symmetric about both diagonals, i.e. $\rho_{\alpha\alpha} = \rho_{\beta\beta}$ and $\rho_{\alpha\beta} = \rho_{\beta\alpha}$, and a stationary autoregressive model obtains if $|R| < 1$, i.e. $\rho_{\alpha\alpha}^2 + \rho_{\beta\beta}^2 + \rho_{\alpha\beta}^2 + \rho_{\beta\alpha}^2 < 1$.

Suppose that $RJ - JR = 0$, and omit the $O_p(m^{-1})$ contribution in equation (4.1). Then we obtain an autoregressive model for the a_{it}:

$$a_{it} - a_{i0} = R(a_{i,t-1} - a_{i0}) + u_{it}. \tag{4.2}$$

This model might have been adopted at the outset, but it does not incorporate the constraint explicitly. However, it can lead to a useful approximate approach as follows.

Note that equation (4.2) gives

$$p(a_{it}|a_{i0}, a_{i1}, \ldots, a_{i,t-1}) = \det(2\pi\Sigma)^{-1/2} \exp\{-\tfrac{1}{2}(a_{it} - m_{it})^{\mathrm{T}}\Sigma^{-1}(a_{it} - m_{it})\}, \tag{4.3}$$

with

$$m_{it} = a_{i0} + R(a_{i,t-1} - a_{i0}). \tag{4.4}$$

Also, the appropriate distribution for updating a_{it} and a_{jt}, as a result of the unobserved changes from time $t-1$ together with the evidence from the match result (x_{it}, y_{jt}), is

$$\begin{aligned} &p(a_{it}, a_{jt}|a_{i,t-1}, a_{j,t-1}, x_{it}, y_{jt}) \\ &\quad = p(x_{it}, y_{jt}|a_{it}, a_{jt}, a_{i,t-1}, a_{j,t-1})\, p(a_{it}, a_{jt}|a_{i,t-1}, a_{j,t-1}) / p(x_{it}, y_{jt}|a_{i,t-1}, a_{j,t-1}) \\ &\quad = p(x_{it}, y_{jt}|a_{it}, a_{jt})\, p(a_{it}|a_{i,t-1})\, p(a_{jt}|a_{j,t-1}) / p(x_{it}, y_{jt}|a_{i,t-1}, a_{j,t-1}), \end{aligned} \tag{4.5}$$

where $p(x_{it}, y_{jt}|a_{it}, a_{jt})$ is given by equation (1.1) and $p(a_{it}|a_{i,t-1})$ and $p(a_{jt}|a_{j,t-1})$ are given by equation (4.3); the denominator does not involve (a_{it}, a_{jt}).

We seek some summary of the information in equation (4.5) that will enable us to avoid the integration in equation (3.4). A natural suggestion is to adopt, for the updated versions of a_{it} and a_{jt}, their maximum probability estimators based on equation (4.5); see, for example, Fahrmeir and Kaufman (1991) and Fahrmeir (1992). These are derived in Appendix A.2 as

$$\begin{aligned} a_{it} &= m_{it} + \Sigma r_{ijt}, \\ a_{jt} &= m_{jt} + \Sigma r_{jit}, \end{aligned} \tag{4.6}$$

where

$$r_{ijt} = (x_{it} - \eta\alpha_{it}\beta_{jt}, y_{jt} - \alpha_{jt}\beta_{it})^{\mathrm{T}}$$

is the vector of 'Poisson residuals', i.e. the discrepancies between the home and away goals and their expected values; r_{jit} is just r_{ijt} with the components interchanged. The updated version of a_{it} is thus expressed as its autoregressive mean m_{it} plus an adjustment Σr_{ijt}, to take account of the match result, and likewise for a_{jt}. Unfortunately, however, r_{ijt} itself involves the components of a_{it} and a_{jt}, so equations (4.6) give a_{it} and a_{jt} implicitly. An iterative solution, described in Appendix A.2, has been implemented and works reasonably well. At least one solution to equations (4.6) exists because the probability density (4.5) must have at least one maximum. If teams i and j do not play at time t, the appropriate updating distribution is just $p(a_{it}, a_{jt}|a_{i,t-1}, a_{j,t-1})$. On the same basis as above, this leads to the reduced updating formulae

$$\begin{aligned} a_{it} &= m_{it}, \\ a_{jt} &= m_{jt}. \end{aligned} \tag{4.7}$$

We now propose that the updating formulae (4.6) and (4.7) be adopted in place of equation (4.2) as the formal model for $(\alpha_{it}, \beta_{it}, \alpha_{jt}, \beta_{jt})$. In effect, the $(\alpha_{it}, \beta_{it})$ become interlinked stochastic processes driven by the innovations (x_{it}, y_{jt}) with initial values $(\alpha_{i0}, \beta_{i0})$ and parameter set $\theta = (\eta, R, \Sigma)$. In equation (4.2) the $(\alpha_{it}, \beta_{it})$ evolve independently over time and information would just be gathered about their progress from the match results. Now, in equations (4.6) and (4.7), the processes $(\alpha_{it}, \beta_{it})$ are actually driven by the match results and are thereby interdependent.

The nature of the approximation made in this section can be described as follows: the original process (4.1) has been replaced by the approximation (4.2), and then equation (4.2) has been replaced by a different, but closely related, process.

A likelihood function based on this scheme can be constructed for the matches up to and including time t as

$$L_t(\theta) = p(G_1, \ldots, G_t|A_0, \theta) = \prod_{s=1}^{t} p(G_s|G_1, \ldots, G_{s-1}, A_0, \theta),$$

where $A_t = (a_{1t}, \ldots, a_{mt})$. Assume that $(G_1, \ldots, G_{s-1}, A_0)$ and $(A_0, \ldots, A_{s-1})$ are equivalent, i.e. that expression (4.6) has a unique solution for a_{it} and a_{jt}. Then,

$$\begin{aligned} p(G_s|G_1, \ldots, G_{s-1}, A_0, \theta) &= p(G_s|A_0, \ldots, A_{s-1}, \theta) \\ &= \int p(G_s|A_0, \ldots, A_s, \theta)\, p(A_s|A_0, \ldots, A_{s-1}, \theta)\, \mathrm{d}A_s \\ &= p(G_s|A_0, \ldots, A_{s-1}, A_s = M_s, \theta) = p(G_s|A_s = M_s, \theta) \end{aligned} \tag{4.8}$$

because, according to equations (4.7), $p(A_s|A_0, \ldots, A_{s-1}, \theta)$ is concentrated at the single point $A_s = M_s$, where M_s is the complete set of m_{is}-values for all the teams. Hence,

$$L_t(\theta) = \prod_{s=1}^{t} p(G_s|A_s = M_s, \theta) = \prod_{s=1}^{t} \prod_{ij} p(x_{is}, y_{js}|a_{is} = m_{is}, a_{js} = m_{js}, \theta), \tag{4.9}$$

where Π_{ij} is the product over all the matches played at time s and $p(x_{is}, y_{js}|a_{is} = m_{is}, a_{js} = m_{js}, \theta)$ is given by equations (1.1) and (4.4). Note that $L_t(\theta)$ in equation (4.9) is implicitly conditioned on Γ_0 because the m_{i1} depend on the a_{i0}. This likelihood function is very simply computed, unlike that in equation (3.4).

5. Prediction

Because of the steadily accruing information on the a_{it} we might hope that predictions would become more useful as time goes on. However, the inherent variability of low count Poisson variables limits the attainable accuracy of forecasting match results. Even if the model were correct, and the parameters precisely known, there would still be considerable uncertainty in the predictions.

For prediction one step ahead, the relevant distribution is $p(G_t|G_1, \ldots, G_{t-1}, \Gamma_0)$. This gives the probabilities of various scores (x_{it}, y_{jt}) for particular games, and so the probabilities of home or away wins can be computed. It may be evaluated as

$$p(G_t|G_1, \ldots, G_{t-1}, \Gamma_0) = \int p(G_t|G_1, \ldots, G_{t-1}, \Gamma_0, \theta)\, p(\theta|G_1, \ldots, G_{t-1}, \Gamma_0)\, \mathrm{d}\theta, \tag{5.1}$$

where $p(\theta|G_1, \ldots, G_{t-1}, \Gamma_0)$ is the posterior distribution of θ given the match results up to time $t - 1$.

For the approximation method we have, from equation (4.8),

$$p(G_t|G_1, \ldots, G_{t-1}, \Gamma_0, \theta) = p(G_t|A_t = M_t, \theta),$$

which is given by equations (1.1) and (4.4). Provided that $p(\theta|G_1, \ldots, G_{t-1}, \Gamma_0)$ is sufficiently peaked, the integral (5.1) can be well approximated by a plug-in estimate, $p(G_t|A_t = M_t, \theta = \hat{\theta}_{t-1})$, where $\hat{\theta}_{t-1}$ is the maximum likelihood estimate based on $(G_1, \ldots, G_{t-1}, \Gamma_0)$. The integral could be evaluated by simulation, e.g. by MCMC sampling from the posterior, but we have not pursued this alternative because our focus is on a computational approach that avoids MCMC methods: see the discussion in Section 7.

6. Full results round-up

The methods described above were applied to the Football Association League fixtures over the years 1992–1997, and home win, draw and away win probabilities were calculated accordingly. It is these probabilities that are of chief concern in the primary betting market. Some summary statistics examining the performance of each approach were evaluated and are now presented.

Table 1 shows summaries of the actual outcomes given the most likely predicted outcome for the approximation method and the Dixon and Coles (1997) model. If a model is performing well then the elements in the leading diagonal of these tables should dominate. In practice though, since a draw is so seldom predicted as the most likely outcome, only the results for a home win and an away win are of significant interest. The two approaches appear to have roughly the same predictive ability for home wins (in the first row of the table), with just under 50% of games

Table 1. Comparison of predicted and actual outcomes

Most likely predicted outcome	*Approximation results for the following actual outcomes:*			*Dixon–Coles results for the following actual outcomes:*		
	Home win	*Draw*	*Away win*	*Home win*	*Draw*	*Away win*
Home win	0.48	0.28	0.24	0.49	0.28	0.23
Draw	—	—	—	0.33	0.33	0.33
Away win	0.33	0.28	0.39	0.33	0.31	0.36

predicted as home wins actually resulting in home wins. For away wins, the methods perform less well, with the approximation method slightly better.

The indicators examined above, although informative, are prone to considerable variability since, for example, a home win could be the most likely outcome for any probability greater than $\frac{1}{3}$. As an alternative, consider the predictive ability of each method when only strong favourites are considered. Table 2 gives a summary of the empirical outcomes conditional on strong favourites, 'strong' in the sense that the predicted probability is 0.5 or more. Again, for both home and away wins the two approaches perform very similarly.

Tables 1 and 2 provide summary information averaged over the entire time horizon of the study. Of potentially greater interest is the performance of the approaches through time: do the methods become better at predicting the outcomes of games as time goes on? Fig. 1 depicts a quantity that is related to this issue, namely, a rolling predictive likelihood defined as

$$\mathrm{RPL}(s,t) = -N(s,t)^{-1} \sum_{r=s}^{t} \sum_{ij} \log\{\mathrm{pr}(o_{ijr})\},$$

where o_{ijr} denotes the outcome of the match between teams i and j played at time r, Σ_{ij} denotes summation over those games and $N(s,t)$ denotes the total number of games played at times $s, s+1, \ldots, t$; $\mathrm{pr}(o_{ijr})$ is the predicted probability of outcome o_{ijr}. A good performance is indicated by a small value of this statistic. Fig. 1 shows a plot, for each method, of $\mathrm{RPL}(t-69,t)$ *versus* t; the window length, 70 time periods, is roughly one season. The Dixon–Coles method starts poorly but improves rapidly and betters the approximation around week 210. However, the approximation, which is a more general, likelihood-based approach, is at least competitive throughout.

7. Conclusion

The purpose of the work described here is to develop a fast computational approach for predicting the results of football matches with a view to betting in the primary market on win, draw

Table 2. Comparison of predicted and actual outcomes

Approximation	0.51	0.43
Dixon–Coles	0.52	0.49

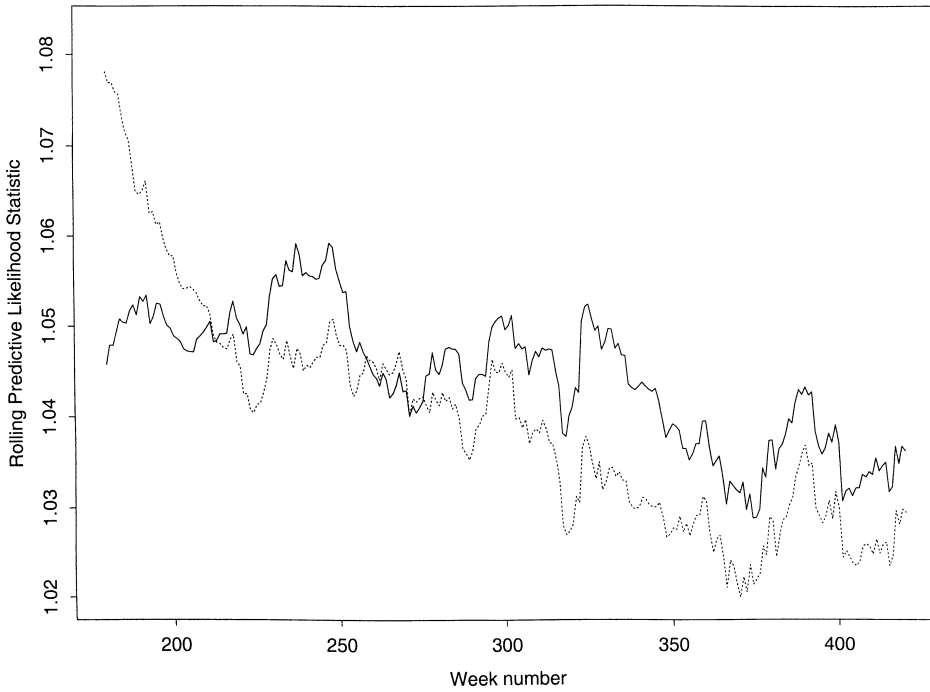

Fig. 1. Rolling predictive likelihood statistic for the Dixon–Coles method (· · · · · · ·) and the approximation (———)

or lose outcomes. For this, a novel method has been developed in which the original stochastic process model is replaced by an approximation that yields more tractable computation than MCMC methodology in this case, but which does not lose too much predictive power. The approximation method clearly has wider potential for tractable computations in the context of non-linear state space models.

Acknowledgements

We thank the Joint Editor and referees for their helpful comments.

Appendix A

A.1. The identifiability constraint

Suppose that $\alpha_i\beta_j$ is fixed as $\exp(\nu_{ij})$, i.e. $\log(\alpha_i) + \log(\beta_j) = \nu_{ij}$. Then the whole set of equations can be written as $Ax = \nu$, where x is the $2m \times 1$ vector

$$(\log(\alpha_1), \ldots, \log(\alpha_m), \log(\beta_1), \ldots, \log(\beta_m))^{\mathrm{T}},$$

ν is the $m(m-1) \times 1$ vector of $\nu_{ij}s$ and A is the appropriate $m(m-1) \times 2m$ incidence matrix. With $m = 4$, for instance,

$$A = \begin{pmatrix} 1\,0\,0\,0 & 0\,1\,0\,0 \\ 1\,0\,0\,0 & 0\,0\,1\,0 \\ 1\,0\,0\,0 & 0\,0\,0\,1 \\ 0\,1\,0\,0 & 1\,0\,0\,0 \\ 0\,1\,0\,0 & 0\,0\,1\,0 \\ 0\,1\,0\,0 & 0\,0\,0\,1 \\ 0\,0\,1\,0 & 1\,0\,0\,0 \\ 0\,0\,1\,0 & 0\,1\,0\,0 \\ 0\,0\,1\,0 & 0\,0\,0\,1 \\ 0\,0\,0\,1 & 1\,0\,0\,0 \\ 0\,0\,0\,1 & 0\,1\,0\,0 \\ 0\,0\,0\,1 & 0\,0\,1\,0 \end{pmatrix}.$$

The identifiable parametric functions are the elements of Ax, and so the number of independent identifiable functions is rank(A). Thus, the appropriate number of constraints will be $2m - \text{rank}(A)$. We can determine the dimension of the kernel of A quite easily by seeing that the only solutions of $Ay = 0$ are y-vectors proportional to $(1, \ldots, 1, -1, \ldots, -1)$; this can be checked for the example with $m = 4$ given above, and thence generalized. Note that this y-vector corresponds to adding a constant to each $\log(\alpha_i)$ and subtracting the same constant from each $\log(\beta_i)$, which amounts to precisely the rescaling of the αs and βs referred to in Section 2. Thus, $\dim\{\ker(A)\} = 1$, so $\text{rank}(A) = 2m - 1$ and so only a single constraint is needed.

The implications of the autoregressive model (3.1) for the αs and βs defined by equation (2.2) are as follows. We have

$$\begin{pmatrix} \log(\alpha_{it}) \\ \log(\beta_{it}) \end{pmatrix} = \begin{pmatrix} \gamma_{\alpha it} - \gamma_t \\ \gamma_{\beta it} - \gamma_t \end{pmatrix} = \gamma_{it} - (2m)^{-1} \sum_{j=1}^{m} J\gamma_{jt},$$

where J is a 2×2 matrix of 1s. Hence,

$$\begin{aligned} \begin{pmatrix} \log(\alpha_{it}) - \log(\alpha_{i0}) \\ \log(\beta_{it}) - \log(\beta_{i0}) \end{pmatrix} &= \gamma_{it} - \gamma_{i0} - (2m)^{-1} \sum_{j=1}^{m} J(\gamma_{jt} - \gamma_{j0}) \\ &= R(\gamma_{i,\,t-1} - \gamma_{i0}) + u_{it} - (2m)^{-1} \sum_{j=1}^{m} J\{R(\gamma_{j,\,t-1} - \gamma_{j0}) + u_{jt}\} \\ &= R(\gamma_{i,\,t-1} - \gamma_{i0}) - (2m)^{-1} \sum_{j=1}^{m} RJ(\gamma_{j,\,t-1} - \gamma_{j0}) \\ &\quad + (2m)^{-1} \sum_{j=1}^{m} (RJ - JR)(\gamma_{j,\,t-1} - \gamma_{j0}) + u_{it} - (2m)^{-1} \sum_{j=1}^{m} Ju_{jt} \\ &= R \begin{pmatrix} \log(\alpha_{i,t-1}) - \log(\alpha_{i0}) \\ \log(\beta_{i,t-1}) - \log(\beta_{i0}) \end{pmatrix} + (RJ - JR)(\gamma_{t-1} - \gamma_0) + u_{it} - (2m)^{-1} \sum_{j=1}^{m} Ju_{jt}, \end{aligned}$$

which is equation (4.1).

A.2. Equations and numerical solution for the approximation method

Equation (4.6) is derived together with a suggested numerical method of solution. The relevant part of the logarithm of equation (4.5) is

$$\begin{aligned} M = &-\gamma\alpha_{it}\beta_{jt} + x_{it}\log(\gamma\alpha_{it}\beta_{jt}) - \alpha_{jt}\beta_{it} + y_{jt}\log(\alpha_{jt}\beta_{it}) \\ &-\tfrac{1}{2}(a_{it} - m_{it})^{\mathrm{T}}\Sigma^{-1}(a_{it} - m_{it}) - \tfrac{1}{2}(a_{jt} - m_{jt})^{\mathrm{T}}\Sigma^{-1}(a_{jt} - m_{jt}), \end{aligned}$$

and this has derivatives as follows:

$$\frac{\partial M}{\partial \alpha_{it}} = -\gamma\beta_{jt} + \alpha_{it}^{-1}x_{it} - \alpha_{it}^{-1}[\{\log(\alpha_{it}) - m_{it1}\}\Sigma^{11} + \{\log(\beta_{it}) - m_{it2}\}\Sigma^{12}];$$

$$\frac{\partial M}{\partial \beta_{it}} = -\alpha_{jt} + \beta_{it}^{-1}y_{jt} - \beta_{it}^{-1}[\{\log(\beta_{it}) - m_{it2}\}\Sigma^{22} + \{\log(\alpha_{it}) - m_{it1}\}\Sigma^{12}];$$

$$\frac{\partial M}{\partial \alpha_{jt}} = -\beta_{it} + \alpha_{jt}^{-1}y_{jt} - \alpha_{jt}^{-1}[\{\log(\alpha_{jt}) - m_{jt1}\}\Sigma^{11} + \{\log(\beta_{jt}) - m_{jt2}\}\Sigma^{12}];$$

$$\frac{\partial M}{\partial \beta_{jt}} = -\gamma\alpha_{it} + \beta_{jt}^{-1}x_{it} - \beta_{jt}^{-1}[\{\log(\beta_{jt}) - m_{jt2}\}\Sigma^{22} + \{\log(\alpha_{jt}) - m_{jt1}\}\Sigma^{12}];$$

here, m_{it1} and m_{it2} are the components of m_{it}, and the Σ^{jk} are the elements of Σ^{-1}. On equating these derivatives to 0, and rearranging, we obtain

$$\{\log(\alpha_{it}) - m_{it1}\}\Sigma^{11} + \{\log(\beta_{it}) - m_{it2}\}\Sigma^{12} = x_{it} - \gamma\alpha_{it}\beta_{jt},$$

$$\{\log(\alpha_{it}) - m_{it1}\}\Sigma^{12} + \{\log(\beta_{it}) - m_{it2}\}\Sigma^{22} = y_{jt} - \alpha_{jt}\beta_{it},$$

$$\{\log(\alpha_{jt}) - m_{jt1}\}\Sigma^{11} + \{\log(\beta_{jt}) - m_{jt2}\}\Sigma^{12} = y_{jt} - \alpha_{jt}\beta_{it},$$

$$\{\log(\alpha_{jt}) - m_{jt1}\}\Sigma^{12} + \{\log(\beta_{jt}) - m_{jt2}\}\Sigma^{22} = x_{it} - \gamma\alpha_{it}\beta_{jt}.$$

These equations may be 'solved' to give

$$\{\log(\alpha_{it}) - m_{it1}\}\{\Sigma^{11}\Sigma^{22} - (\Sigma^{12})^2\} = \Sigma^{22}(x_{it} - \gamma\alpha_{it}\beta_{jt}) - \Sigma^{12}(y_{jt} - \alpha_{jt}\beta_{it}),$$

$$\{\log(\beta_{it}) - m_{it2}\}\{\Sigma^{11}\Sigma^{22} - (\Sigma^{12})^2\} = -\Sigma^{12}(x_{it} - \gamma\alpha_{it}\beta_{jt}) + \Sigma^{11}(y_{jt} - \alpha_{jt}\beta_{it}),$$

$$\{\log(\alpha_{jt}) - m_{jt1}\}\{\Sigma^{11}\Sigma^{22} - (\Sigma^{12})^2\} = -\Sigma^{12}(x_{it} - \gamma\alpha_{it}\beta_{jt}) + \Sigma^{22}(y_{jt} - \alpha_{jt}\beta_{it}),$$

$$\{\log(\beta_{jt}) - m_{jt2}\}\{\Sigma^{11}\Sigma^{22} - (\Sigma^{12})^2\} = \Sigma^{11}(x_{it} - \gamma\alpha_{it}\beta_{jt}) - \Sigma^{12}(y_{jt} - \alpha_{jt}\beta_{it}).$$

Using the identity $(\Sigma^{-1})^{-1} = \Sigma$, in component form, these equations yield equations (4.6):

$$\log(\alpha_{it}) = m_{it1} + \Sigma_{11}(x_{it} - \gamma\alpha_{it}\beta_{jt}) + \Sigma_{12}(y_{jt} - \alpha_{jt}\beta_{it}),$$

$$\log(\beta_{it}) = m_{it2} + \Sigma_{12}(x_{it} - \gamma\alpha_{it}\beta_{jt}) + \Sigma_{22}(y_{jt} - \alpha_{jt}\beta_{it}),$$

$$\log(\alpha_{jt}) = m_{jt1} + \Sigma_{12}(x_{it} - \gamma\alpha_{it}\beta_{jt}) + \Sigma_{11}(y_{jt} - \alpha_{jt}\beta_{it}),$$

$$\log(\beta_{jt}) = m_{jt2} + \Sigma_{22}(x_{it} - \gamma\alpha_{it}\beta_{jt}) + \Sigma_{12}(y_{jt} - \alpha_{jt}\beta_{it}).$$

For numerical solution we first reduce these four equations to two, by adding the first and fourth, and the second and third, to obtain

$$\log(\alpha_{it}\beta_{jt}) = m_{it1} + m_{jt2} + (\Sigma_{11} + \Sigma_{22})(x_{it} - \gamma\alpha_{it}\beta_{jt}) + 2\Sigma_{12}(y_{jt} - \alpha_{jt}\beta_{it}),$$

$$\log(\alpha_{jt}\beta_{it}) = m_{it2} + m_{jt1} + 2\Sigma_{12}(x_{it} - \gamma\alpha_{it}\beta_{jt}) + (\Sigma_{11} + \Sigma_{22})(y_{jt} - \alpha_{jt}\beta_{it}).$$

We now have two non-linear equations in the two quantities, $\alpha_{it}\beta_{jt}$ and $\alpha_{jt}\beta_{it}$, so an iterative scheme for solution is called for. The first attempt was one in which the right-hand sides contained the previous estimates, but the successive values tended to oscillate. This suggested that convergence could be accelerated by updating the estimates as the average of the current and previous values, and this scheme was found to work quite well. The technique bears a strong resemblance to Aitken's δ^2 acceleration method (Dixon (1974), section 7.6). Once the process has converged, the previous set of equations is used to compute α_{it}, β_{it}, α_{jt} and β_{jt} separately.

References

Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992) A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Statist. Ass.*, **87**, 493–500.
Carter, C. K. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.
Dixon, C. (1974) *Numerical Analysis*. London: Blackie.
Dixon, M. J. and Coles, S. G. (1997) Modelling association football scores and inefficiencies in the football betting market. *Appl. Statist.*, **46**, 265–280.
Fahrmeir, L. (1992) Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalised linear models. *J. Am. Statist. Ass.*, **87**, 501–509.
Fahrmeir, L. and Kaufman, H. (1991) On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family models. *Metrika*, **38**, 37–60.
Kitagawa, G. (1987) Non-Gaussian state-space modelling of nonstationary time series (with discussion). *J. Am. Statist. Ass.*, **82**, 1032–1063.
Knorr-Held, L. (2000) Dynamic rating of sports teams. *Statistician*, **49**, 261–276.
Koning, R. H. (2000) Balance in competition in Dutch soccer. *Statistician*, **49**, 419–431.
Maher, M. J. (1982) Modelling association football scores. *Statist. Neerland.*, **36**, 109–118.
Rue, H. and Salvesen, Ø. (2000) Prediction and retrospective analysis of soccer matches in a league. *Statistician*, **49**, 399–418.
Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.