Bayesian state-space models for the modelling and prediction of the results of English Premier League football

Gareth Ridall

Lancaster University, Lancaster, UK

E-mail: g.ridall@lancs.ac.uk

Andrew Titman

Lancaster University, Lancaster, UK

Anthony Pettitt

Queensland University of Technology and the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, Brisbane, Australia

Summary. The attraction of using state space models (SSM) is their ability to efficiently and dynamically predict in the presence of change. In this paper we formulate a Bayesian SSM capable of predicting the outcomes of football matches and the associated states, which are the attacking and defensive strengths of each side and the common home goal advantage. Our filter achieves accuracy and efficiency by exploiting conjugacy in its update step and using exact expressions to describe the evolution of the states. The presence of conjugacy enables us to use a mean field approximation (MFA) to update the states given fresh observations. The method is evaluated using the full history of the English Premier League and shown to be competitive, or superior, to weighted likelihood or score-driven time series based methods.

1. Introduction

Association football arguably has the largest following of any game in the world. Of all the football leagues, the English Premier League (EPL) is probably the most well-known. Furthermore, the EPL has kept the structure of the competition intact and fixed for over two decades. All of the data used in this paper is taken from the URL `https://football-data.co.uk/data.php`. The same site has a large number of national and international leagues recorded for each season in a standard uniform format. The structure of the relevant columns of the data for a single season is illustrated in Table 1. The methods and code that we develop in this paper will work (with minor adjustments) with any of the data-sets describing the history of leagues found on this site.

In the literature there appears to be two broad classes of models for forecasting the outcomes of football games. Firstly, there are models that describe the outcome directly looking on the data as ordinal or multinomial. Secondly there are score based models,

Table 1. An excerpt from a csv file of data taken from the football-data.co.uk website. The columns show the Dates, the home and away teams (HomeTeam and AwayTeam), the home and away full-time goals scored (FTHG and HTAG) and the full-time result (FTR) where H, A and D stand for a home win, an away win and a draw respectively.

	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR
1	12/9/2020	Fulham	Arsenal	0	3	A
2	12/9/2020	Crystal Palace	Southampton	1	0	H
3	12/9/2020	Liverpool	Leeds	4	3	H
4	12/9/2020	West Ham	Newcastle	0	2	A
5	13/09/2020	West Brom	Leicester	0	3	A
6	13/09/2020	Tottenham	Everton	0	1	A
⋮	⋮	⋮	⋮	⋮	⋮	⋮
376	23/05/2021	Liverpool	Crystal Palace	2	0	H
377	23/05/2021	Man City	Everton	5	0	H
378	23/05/2021	Sheff. United	Burnley	1	0	H
379	23/05/2021	West Ham	Southampton	3	0	H
380	23/05/2021	Wolves	Man United	1	2	A

where the outcome of the match is predicted indirectly through the predicted scores. See Egidi & Torelli (2020) for a discussion of the relative merits of these two approaches. Modelling the outcome directly can be achieved by using a generalisation of the Bradley Terry model: (Bradley & Terry, 1952). Examples include those of Rao & Kupper (1967) and Davidson (1970). This approach has been extended to the dynamic modelling of football by Cattelan et al. (2013).

Score based models that describe teams' scores typically use a Poisson distribution which is expressed in terms of an attacking strength, defensive strength, and home-ground advantages as in Maher (1982). An example of these models using a common home ground advantage is Dixon & Coles (1997) who use a bivariate Poisson model. In this model the scores in only low scoring games are correlated but this can be positively or negatively. They accommodate the possibility of dynamic prediction by implementing a simple method which discounts past observations through use of an exponentially weighted likelihood. Let T be the current time in days and t the day that the observation was made then the weights, given by $\omega_t = \exp(-k(T-t))$, are used to down-weight the log-likelihood contributions of observations from the past.

Karlis & Ntzoufras (2003) describe an alternative static bivariate Poisson model with fixed attacking and defensive strengths and with a common home ground advantage, by assuming the presence of an unobserved number of common goals for each game assumed to be driven by external factors not related to form. This approach can be looked on as the sharing of an additive random effect and is able to explain only positive correlations between scores.

Likelihood and weighted likelihood approaches have the problem that re-estimation of the whole model using all observations must be carried out using an optimiser each time new data arrives. Dynamic or state space models should not need to do this. The Kalman filter (Kalman, 1960), for instance, is able to recursively modify sufficient statistics which summarise the data collected up to the present time. There have been several examples in the literature of attempts to employ SSMs to model and predict football outcomes, see Koopman & Lit (2019) for a more comprehensive review. For example, Crowder et al. (2002) construct a state space model with time varying abilities by applying a Bayesian approach. Initially they used the traditional Bayesian method of estimation using Markov Chain Monte Carlo (MCMC) but because of the computational demands imposed they suggest using an approximation. Rue & Salvesen (2000) construct a dynamic Poisson model where the uncertainty in the state process is explained using Brownian motion. Koopman & Lit (2015) use the same bivariate distribution as Karlis & Ntzoufras (2003) but employ a state-space model to describe the evolution of the form parameters using additive Gaussian uncertainty in the state process. They model the changing attacking and defensive strengths of each team with fixed auto-regressive parameters and the use of Monte-Carlo methods for estimating the likelihood.

Koopman & Lit (2019) introduced a computationally simpler approach by expressing the attacking and defensive strengths of each time as a score-driven time series/generalised auto-regressive score model (Creal et al. , 2013). Following each round of matches, the likelihood score (first derivative of the log-likelihood of the match result is calculated with respect to the attacking and defensive strengths of each featured team) evaluated at their current values, is computed and used to update the parameters. The way in which the likelihood score affects the evolution is controlled by parameters which are estimated by maximizing the likelihood using the parameter values at time $t-1$ to predict outcomes at time t over a training set. Koopman & Lit (2019) show that the quality of predictions from this method were comparable to those using the method in Koopman & Lit (2015), but at a fraction of the computational cost.

In this paper, we use an alternative parameterisation and are able to devise sequential updates expressed using closed form expressions. Following previous approaches, we assume the number of goals scored by a team depends on their attacking strength (ability to score goals), their opponent's defensive strength (ability to restrict goals) and a common home goal advantage, where we allow all of these parameters to be dynamic. Like Karlis & Ntzoufras (2003) we can explain positive correlations between the scores of the home and away teams, but do so by employing a multiplicative random effect rather than an additive one. We use a state space model to update the dynamic parameters. The state space model employs a similar measurement equation to Karlis & Ntzoufras (2003) and Dixon & Coles (1997) and conjugate Gamma distributions for the evolving distributions of states for the attacking, the defensive strengths and the home ground advantage.

The main strength of our model is its tractability which is achieved by expressing both the state and measurement equations as the products of conjugate distributions and by using closed form expressions for the evolution of the state process. The dependence between the state variables is resolved using a mean field approximation (Jordan et al., 1999), a type of variational Bayes (VB), where each of the states maintain Gamma distributions. Evolution of these Gamma dynamic variables is induced through the application of multiplicative noise through scaled beta distributions which result in simple expressions dependent on "forgetting' parameters. The forgetting parameters are optimized over a training set and then assumed static over the test set. This results in a highly efficient algorithm with relatively good predictive properties.

The remainder of the paper is organised as follows. In Section 2 we present a static version of the proposed Bayesian football model. In Section 3 we describe the ingredients of a state space model and formulate our Bayesian dynamic football model which is illustrated through application to dynamic prediction of the last fourteen seasons of the EPL. In Section 4 we illustrate the ability of the SSM to filter and smooth the evolving state variables. Finally, in Section 5 we present our conclusions and suggest extensions and further work.

2. Static models for modelling football scores.

The initial three seasons of the EPL involved 22 teams with 462 Games. Whereas, since the 1995/1996 season it has consisted of 20 teams which play each other twice, once at home and once away in a season consisting of 380 games or 38 rounds. A round consists of 10 games and involves all 20 teams. After each season three teams are relegated and three teams are promoted from the next league down, except after 1994/95 where four were relegated and two promoted. In general let N_T be the number of teams, N_G be the number of games in a season, and N_R be the number of rounds of a season. Let the games within a season, in chronological order, be labeled as $t = 1, \ldots, \mathrm{N_G}$.

We let $i \in \{1, 2, \ldots, \mathrm{N_T}\}$ denote the home team at game t and $j \in \{1, 2, \ldots, \mathrm{N_T}\}$ denote the away team at the same game. Let x_t denote the number of home goals scored in game t and y_t be the number of away goals scored in the same game.

2.1. A univariate Poisson model

The univariate Poisson model is one of the simplest of all score based models and is described for example in Davison (2003). Despite its simplicity, we find that this model makes surprisingly good predictions. Many of the other models including ours are generalisations of this idea. Let λ_t^H be the expected home side score and λ_t^A be the expected away side score. Let α_i be the attacking strength of the home team for game t and β_j be the defensive strength of the away team. Note that a strong defensive team has a low value of β so we use its reciprocal, $\phi = \beta^{-1}$, in some of our plots to depict defensive strength but retain β in our expressions to help simplify the notation. Note also that γ is used to represent the common home ground advantage. The home and away scores, X_t and Y_t are modelled as conditionally independent Poisson given by

$$\begin{aligned} X_t \mid \lambda_t^H &\sim \text{Poisson}\left(\lambda_t^H\right), & \lambda_t^H &= \alpha_i \beta_j \gamma, \\ Y_t \mid \lambda_t^A &\sim \text{Poisson}\left(\lambda_t^A\right), & \lambda_t^A &= \alpha_j \beta_i, \end{aligned} \tag{2.1}$$

where $\beta_1 = 1$ is the identifiability constraint.

2.2. Bivariate Poisson models

Potentially, the assumption of independent Poisson goals may be overly restrictive and more general bivariate models that account for the possibility of dependence may be necessary.

We point to two examples of bivariate models from the football literature. Dixon & Coles (1997), assume that a non-zero correlation between goals only exists for low scoring games where both teams score fewer than 2 goals. Karlis & Ntzoufras (2003) describe the bivariate Poisson model which explains positive correlation by assuming the existence of a common latent variable (not related to form) with a Poisson distribution truncated by the minimum of the two scores.

In this section we also present a third bivariate model, theoretically capable of modelling both positive correlation and over-dispersion. This bivariate model assumes the existence of a shared Gamma multiplicative random effect, ϵ_t, of prior expectation one. This can be contrasted with Karlis & Ntzoufras (2003) who use an additive random effect. The multiplicative random effect idea was used by Arbous & Kerrich (1951) who proposed this model in the context of modelling accidents. Our bivariate model for football just involves a minor modification of Equations (2.1),

$$\begin{aligned} \epsilon_t &\sim \text{Gamma}\left(\kappa, \kappa\right), & \kappa &> 0, t = 1, 2, \ldots, N_G, \\ X_t \mid \lambda_t^H, \epsilon_t &\sim \text{Poisson}\left(\epsilon_t \lambda_t^H\right), & \lambda_t^H &= \alpha_i \beta_j \gamma, \\ Y_t \mid \lambda_t, \epsilon_t^A &\sim \text{Poisson}\left(\epsilon_t \lambda_t^A\right), & \lambda_t^A &= \alpha_j \beta_i, \end{aligned} \tag{2.2}$$

where $\beta_1 = 1$ is the identifiability constraint. The likelihood marginalised over the random effect from the joint is

$$\begin{aligned} f(x_t, y_t \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \kappa) &= \int f(x_t, y_t, \epsilon_t \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \kappa) p(\epsilon_t \mid \kappa) d\epsilon_t \\ &= \frac{\Gamma(\kappa + x_t + y_t)}{\Gamma(\kappa)\Gamma(x_t + 1)\Gamma(y_t + 1)} P_t^{x_t} Q_t^{y_t} (1 - P_t - Q_t)^{\kappa} \end{aligned} \tag{2.3}$$

where $P_t = \frac{\lambda_t^H}{\kappa + \lambda_t^H + \lambda_t^A}$ and $Q_t = \frac{\lambda_t^A}{\kappa + \lambda_t^H + \lambda_t^A}$. This is a bivariate, negative binomial model which is able to accommodate to some extent over-dispersion and positive correlation. Univariate negative Binomial distributions are common for football data (Baxter & Stevenson, 1988) or more generally count data to treat for overdispersion. The marginal variance, covariance and correlations of the bivariate distribution can be derived by using standard identities.

$$\mathrm{Var}(X_t) = \lambda_t^H + \frac{(\lambda_t^H)^2}{\kappa}, \quad \mathrm{Cov}(X_t, Y_t) = \frac{\lambda_t^H \lambda_t^A}{\kappa}, \quad \rho_t = \frac{\lambda_t^H \lambda_t^A}{\sqrt{(\kappa\lambda_t^H + (\lambda_t^H)^2)(\kappa\lambda_t^A + (\lambda_t^A)^2)}}. \tag{2.4}$$

See Appendix A for a proof.

2.3. The Bayesian sequential static model

In this subsection, we introduce a Bayesian sequential static model. We model the attacking and defensive strengths and the home ground advantages using conjugate Gamma priors. After each game we update each relevant dynamic parameters only once using a close approximation to the full conditional posterior distributions. Inference can then be carried out by manipulating the hyper-parameters. For identifiability reasons, we set $\beta_1 = 1$, so that all defensive strengths are measured relative to that of Arsenal. The observation model for the goals scored over a particular season are for games $t = 1, 2, \ldots, N_G$ is kept as Equation 2.2 and the following priors are added.

$$\begin{aligned} \alpha_{1:N_T}, \gamma &\sim \mathrm{Gamma}\,(\delta, \delta), \\ \beta_1 = 1, \beta_{2:N_T} &\sim \mathrm{Gamma}\,(\delta, \delta), \end{aligned} \tag{2.5}$$

where $\delta = 10$.

2.4. Sequential updating

In our static sequential Bayesian model we denote our estimates for game $t = 1, 2, \ldots, N_G$ of the attacking strengths of each club by $\boldsymbol{\alpha}_t = \{\alpha_{1,t}, \alpha_{2,t}, \ldots, \alpha_{N_T,t}\}$, the defensive strengths by $\boldsymbol{\beta}_t = \{\beta_{1,t}, \beta_{2,t}, \ldots, \beta_{N_T,t}\}$ and the common home ground advantage by γ_t. We denote our estimates of the parameters at game t as $\boldsymbol{\theta}_t = \{\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \gamma_t\}$, For the purpose of this section, we assume that these parameters are static and that it is only our estimates of the parameters which change with time as more games are played.

After a game is completed five parameters need to be updated and some of these parameters have strong dependencies (see Figure 1). The aim of this section is to show how single updates to the state hyper-parameters can be used to closely approximate the posterior distribution of each of these parameters taking advantage of the fact that the attacking strength or defensive strengths of each team do not change dramatically from game to game.

Let the prior and posterior distributions for the attacking strength (AS), the defensive strength (DS), and the home goal advantage (HGA) for game $t = 1, 2, \ldots, N_G$ be defined sequentially as

$$\alpha_{a,t} \sim \text{Gamma}\left(\tilde{p}^{\alpha}_{a,t}, \tilde{q}^{\alpha}_{a,t}\right), \quad \beta_{d,t} \sim \text{Gamma}\left(\tilde{p}^{\beta}_{d,t}, \tilde{q}^{\beta}_{d,t}\right), \quad \gamma_t \sim \text{Gamma}\left(\tilde{p}^{\gamma}_{t}, \tilde{q}^{\gamma}_{t}\right). \tag{Priors}$$

$$\alpha_{a,t} \sim \text{Gamma}\left(p^{\alpha}_{a,t}, q^{\alpha}_{a,t}\right), \quad \beta_{d,t} \sim \text{Gamma}\left(p^{\beta}_{d,t}, q^{\beta}_{d,t}\right), \quad \gamma_t \sim \text{Gamma}\left(p^{\gamma}_{t}, q^{\gamma}_{t}\right). \tag{Posteriors}$$

where $a = 1, 2, \ldots, N_T$ and $d = 2, \ldots, N_T$. In the static sequential model, the priors for each of the parameters are just the posteriors derived from the previous observation.

$$\begin{aligned}
\tilde{p}^{\alpha}_{a,t} &\leftarrow p^{\alpha}_{a,t-1}, & \tilde{q}^{\alpha}_{a,t} &\leftarrow q^{\alpha}_{a,t-1},\\
\tilde{p}^{\beta}_{d,t} &\leftarrow p^{\beta}_{d,t-1}, & \tilde{q}^{\beta}_{d,t} &\leftarrow q^{\beta}_{d,t-1},\\
\tilde{p}^{\gamma}_{t} &\leftarrow p^{\gamma}_{t-1}, & \tilde{q}^{\gamma}_{t} &\leftarrow q^{\gamma}_{t-1},\\
p^{\gamma}_{0}, p^{\alpha}_{a,0}, p^{\beta}_{d,0} &\leftarrow \delta, & q^{\gamma}_{0}, q^{\alpha}_{a,0}, q^{\beta}_{d,0} &\leftarrow \delta.
\end{aligned}$$

We let i denote the home side and j the away side. The posterior distribution of the parameters describing a single game is given by

$$\begin{aligned}
\pi(\boldsymbol{\theta}_t \mid \mathbf{x}_{1:t}, \mathbf{y}_{1:t}) \propto\; & \underbrace{\exp - \left[\epsilon_t(\alpha_{i,t}\beta_{j,t}\gamma_t + \alpha_{j,t}\beta_{i,t})\right] \times \left[\epsilon_t\alpha_{i,t}\beta_{j,t}\gamma_t\right]^{x_t} \times \left[\epsilon_t\alpha_{j,t}\beta_{i,t}\right]^{y_t}}_{\text{Sampling distribution}}\\
& \times \underbrace{\alpha_{i,t}^{\tilde{p}^{\alpha}_{i,t}-1}\exp(-\tilde{q}^{\alpha}_{i,t}\alpha_{i,t})\;\alpha_{j,t}^{\tilde{p}^{\alpha}_{j,t}-1}\exp(-\tilde{q}^{\alpha}_{j,t}\alpha_{j,t})}_{\text{Prior Attacking Strengths}}\\
& \times \underbrace{\beta_{i,t}^{\tilde{p}^{\beta}_{i,t}-1}\exp(-\tilde{q}^{\beta}_{i,t}\beta_{i,t})\;\beta_{j,t}^{\tilde{p}^{\beta}_{j,t}-1}\exp(-\tilde{q}^{\beta}_{j,t}\beta_{j,t})}_{\text{Prior Defensive Strengths}}\\
& \times \underbrace{\gamma_t^{\tilde{p}^{\gamma}_{t}-1}\exp(-\tilde{q}^{\gamma}_{t}\gamma_t)}_{\text{Prior HGA}} \times \underbrace{\epsilon_t^{\kappa-1}\exp(-\kappa\epsilon_t),}_{\text{Random effect}}
\end{aligned} \tag{2.6}$$

where $\mathbf{x}_{1:t}$ and $\mathbf{y}_{1:t}$ denote the home and away goals that have been scored up to this point in time. The updates to the five dynamic parameters can be formulated by examining the full conditional posterior distributions of each parameter. Because of the strong dependencies shown in Figure 1, single closed form updates for single parameters do not exist. However, an approximation, can be made by using the expectation of the parameter from the last time this parameter was updated.

The updates to the hyper-parameters are then:

$$
\begin{array}{rlll}
p_t^{\epsilon} \leftarrow \kappa + x_t + y_t, & q_t^{\epsilon} \leftarrow \kappa + \hat{\alpha}_{i,t}\hat{\beta}_{j,t}\hat{\gamma}_t + \hat{\alpha}_{j,t}\hat{\beta}_{i,t}, & \hat{\epsilon}_t = {}^{p_t^{\epsilon}}\!/_{q_t^{\epsilon}}, & \text{(R.E.)} \\
p_{i,t}^{\alpha} \leftarrow \tilde{p}_{i,t}^{\alpha} + x_t, & q_{i,t}^{\alpha} \leftarrow \tilde{q}_{i,t}^{\alpha} + \hat{\gamma}_t\hat{\beta}_{j,t}\hat{\epsilon}_t, & & \text{(AS home)} \\
p_{j,t}^{\alpha} \leftarrow \tilde{p}_{j,t}^{\alpha} + y_t, & q_{j,t}^{\alpha} \leftarrow \tilde{q}_{j,t}^{\alpha} + \hat{\beta}_{i,t}\hat{\epsilon}_t, & & \text{(AS away)} \\
p_{i,t}^{\beta} \leftarrow \tilde{p}_{i,t}^{\beta} + y_t, & q_{i,t}^{\beta} \leftarrow \tilde{q}_{i,t}^{\beta} + \hat{\alpha}_{j,t}\hat{\epsilon}_t, & & \text{(DS home)} \\
p_{j,t}^{\beta} \leftarrow \tilde{p}_{j,t}^{\beta} + x_t, & q_{j,t}^{\beta} \leftarrow \tilde{q}_{j,t}^{\beta} + \hat{\gamma}_t\hat{\alpha}_{i,t}\hat{\epsilon}_t & & \text{(DS away)} \\
p_t^{\gamma} \leftarrow \tilde{p}_t^{\gamma} + x_t, & q_t^{\gamma} \leftarrow \tilde{q}_t^{\gamma} + \hat{\alpha}_{i,t}\hat{\beta}_{j,t}\hat{\epsilon}_t. & & \text{(HGA)}
\end{array}
\tag{2.7}
$$

It is important that the above update for the random effect must be calculated first and updates of the other parameters condition on this estimate: $\epsilon_t = {}^{p_t^{\epsilon}}\!/_{q_t^{\epsilon}}$. The reason for updating this parameter first is that this is the only hyper-parameter that changes abruptly from observation to observation. When making the rest of updates shown in Equations(2.7) we assume that the expectations of the others are fixed and unchanged from the previous game. For example: $\hat{\alpha}_{i,t} = \frac{p^{\alpha}_{i,t-1}}{q^{\alpha}_{i,t-1}}$, $\hat{\beta}_{i,t} = \frac{p^{\beta}_{i,t-1}}{q^{\beta}_{i,t-1}}$ and $\hat{\gamma}_t = \frac{p^{\gamma}_{t-1}}{q^{\gamma}_{t-1}}$.

This approximation is equivalent to performing the first iteration of the coordinate ascent algorithm necessary to find the mean field approximation, a type of variational approximation (Jordan et al., 1999). The full variational approximation is outlined in Appendix C.

Section S1 of the Supplementary Materials gives an overview of existing static models for football, and in Section S2 the performance of the proposed Bayesian model is compared to those methods for 20 seasons of the EPL. The univariate Poisson Bayes model is shown to give better within-season predictions compared to the other models when estimated using maximum likelihood. Moreover, the single step method of updating given in (2.7) produces very similar estimates to those using the full variational method. To further validate our approach in Section S3 of the Supplementary Materials the estimates for the single step method are compared to full Gibbs sampling.

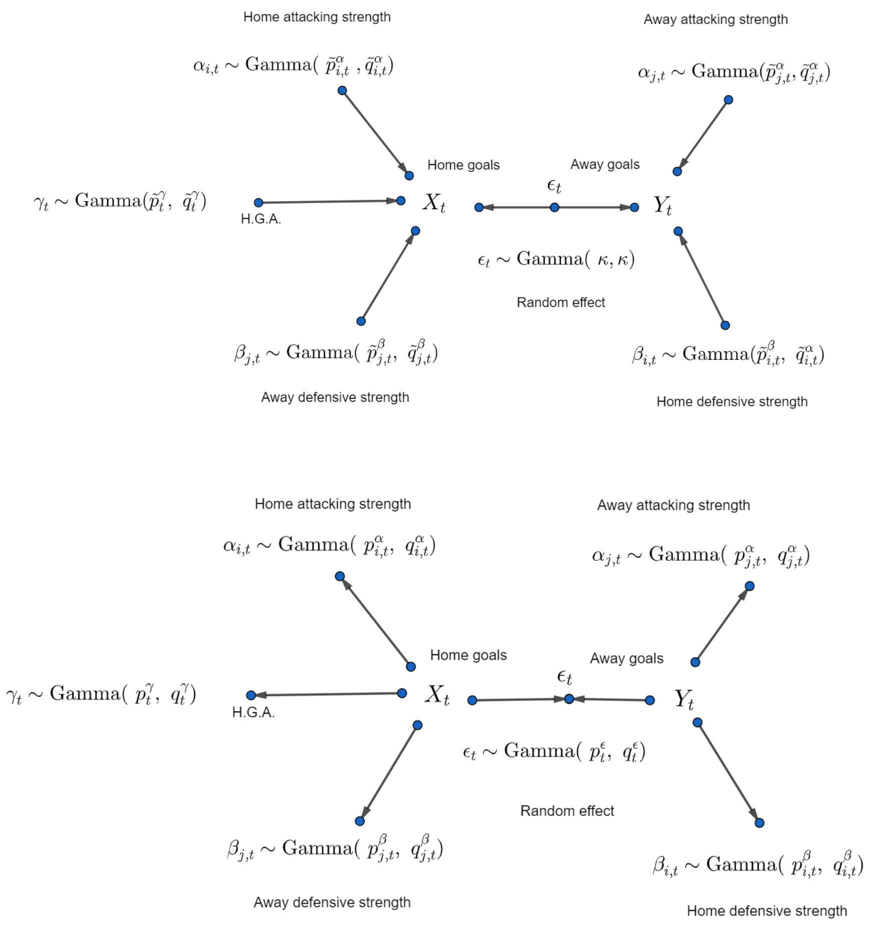

Home attacking strength
$\alpha_{i,t} \sim \mathrm{Gamma}(\tilde{p}^{\alpha}_{i,t}, \tilde{q}^{\alpha}_{i,t})$
Away attacking strength
$\alpha_{j,t} \sim \mathrm{Gamma}(\tilde{p}^{\alpha}_{j,t}, \tilde{q}^{\alpha}_{j,t})$
Home goals
Away goals
$\gamma_t \sim \mathrm{Gamma}(\tilde{p}^{\gamma}_{t}, \tilde{q}^{\gamma}_{t})$
H.G.A.
X_t
ϵ_t
Y_t
$\epsilon_t \sim \mathrm{Gamma}(\kappa, \kappa)$
Random effect
$\beta_{j,t} \sim \mathrm{Gamma}(\tilde{p}^{\beta}_{j,t}, \tilde{q}^{\beta}_{j,t})$
Away defensive strength
$\beta_{i,t} \sim \mathrm{Gamma}(\tilde{p}^{\beta}_{i,t}, \tilde{q}^{\alpha}_{i,t})$
Home defensive strength
Home attacking strength
$\alpha_{i,t} \sim \mathrm{Gamma}(p^{\alpha}_{i,t}, q^{\alpha}_{i,t})$
Away attacking strength
$\alpha_{j,t} \sim \mathrm{Gamma}(p^{\alpha}_{j,t}, q^{\alpha}_{j,t})$
Home goals
Away goals
$\gamma_t \sim \mathrm{Gamma}(p^{\gamma}_{t}, q^{\gamma}_{t})$
H.G.A.
X_t
ϵ_t
Y_t
$\epsilon_t \sim \mathrm{Gamma}(p^{\epsilon}_{t}, q^{\epsilon}_{t})$
Random effect
$\beta_{j,t} \sim \mathrm{Gamma}(p^{\beta}_{j,t}, q^{\beta}_{j,t})$
Away defensive strength
$\beta_{i,t} \sim \mathrm{Gamma}(p^{\beta}_{i,t}, q^{\beta}_{i,t})$
Home defensive strength

Fig. 1. Directed acyclic graphs (DAGs) showing the dependence structure of the variables in our football model. The top panel is a DAG showing the model before the update step and the lower panel shows the posterior distribution after the update step, Equations (2.7), have been carried out.

3. Dynamic modelling of the football data.

In this section we formulate a fully dynamic model where a team's form is allowed to vary between and within seasons. The updates to each of the hyper-parameters are the same as those for the static model and are given by Equations (2.7). We formulate the problem as a state space model and introduce uncertainty to the state processes through the introduction of forgetting factors common to all sides. We first give a general introduction to state space models before using the state space formulation to adapt the update steps of the static model.

3.1. State space models

State space models provide an intuitive approach toward predicting unknown and changing quantities of interest (called states) which are only indirectly observed. With our football application the states are the changing form of each the sides and the common home goal advantage. Bayesian state space models (SSMs) are an important tool for a forecaster because they allow us to predict and sequentially update our beliefs about an unknown state in real time. SSMs have two essential components: Firstly, a state or transition equation which describes the evolution of the parameter of interest within a Markov chain: $\theta_t \sim \pi(\cdot \mid \theta_{t-1}, y_{1:t-1})$. Secondly, there is an observation or measurement equation which expresses the observations as conditionally independent given the current states: $Y_t \mid \theta_t \sim f(\cdot \mid \theta_t)$. The dependence structure of the SSM is illustrated in Figure 2.

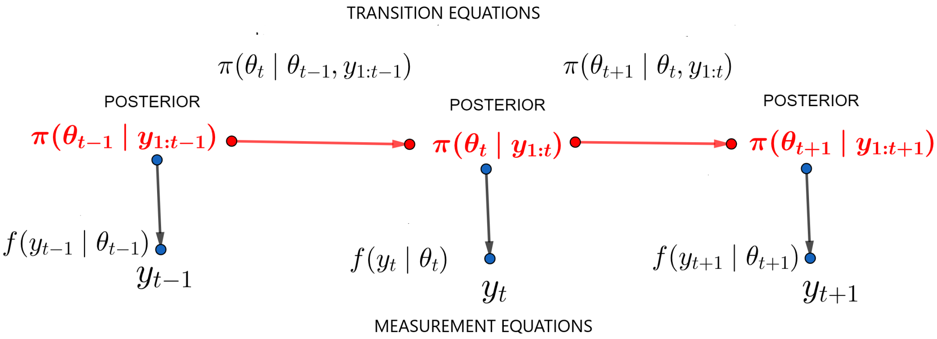

TRANSITION EQUATIONS
$\pi(\theta_t \mid \theta_{t-1}, y_{1:t-1})$
$\pi(\theta_{t+1} \mid \theta_t, y_{1:t})$
POSTERIOR
POSTERIOR
POSTERIOR
$\pi(\theta_{t-1} \mid y_{1:t-1})$
$\pi(\theta_t \mid y_{1:t})$
$\pi(\theta_{t+1} \mid y_{1:t+1})$
$f(y_{t-1} \mid \theta_{t-1})$
y_{t-1}
$f(y_t \mid \theta_t)$
y_t
$f(y_{t+1} \mid \theta_{t+1})$
y_{t+1}
MEASUREMENT EQUATIONS

Fig. 2. Dependence structure of a SSM. The transition equation for the parameter or state is denoted by $\theta_t \sim \pi(\cdot \mid \theta_{t-1}, y_{1:t-1})$, and the downward vertical arrows show the conditional sampling distribution of the observations, or the measurement equation $f(y_t \mid \theta_t)$. These two components can be used in a Bayesian filter to recursively update the evolving posterior distribution of states, $\theta_t \sim \pi(\cdot \mid y_{1:t})$.

3.2. The filter and smoother

A Bayesian filter is a two step recursive algorithm which updates the posterior distribution of states, as a new observation arrives, given all the data collected so far. The first step is known as a "predict", "evolution" or "extension" step where the next state is predicted in the form of a prior for the next update. This is done by combining the posterior distribution from the last observation with the state transition equation in what is known as a convolution. The ability, to sequentially update the parameters using only the last observation and then predict the ahead are important improvements that the SSM offers over weighted likelihood models such as the model used by Dixon & Coles (1997).

In weighted likelihood methods, the result of a match between team A and team B can indirectly affect the estimate of a third team C. This is because the likelihood implicitly assumes the same parameter values held through time hence team A beating team B will alter the impact of all encounters team C had with either A and B in the past. This is in stark contrast to the SSM where only team A and B's abilities can be impacted by a game between A and B.

In the extend step there is a loss of information or memory which we refer to as "forgetting", and we use a parameter $\omega \in (0, 1]$ to describe it where $\omega = 1$ indicates that no information is lost. The second step of the filter is the update step where the prior distribution generated from the extend step is updated using the current observation through an application of Bayes theorem.

Perhaps the most widely used of all filters is the Kalman filter (Kalman, 1960), which is used for prediction and control and where both the state and observation equation are assumed to be linear with additive Gaussian innovations. Its advantage of speed is made possible by the derivation of tractable closed form expressions for both steps. A Bayesian interpretation of the Kalman filter is described in Meinhold & Singpurwalla. (1983) and extended by West et al. (1985) to a class of models known as the dynamic generalised linear model (DGLM) or exponential family dynamic models. Uncertainty to the posterior is induced by the addition of Gaussian noise to the state process with a log link leading to an extension or prediction step which is an approximation.

Instead of using additive Gaussian uncertainty in the state process, we express state uncertainty by the product of Gamma distributions with multiplicative noise provided by the scaled beta distribution. Unlike West et al. (1985), our extension equation step is tractable and exact. Our update step exploits conjugacy and to deal with the problem of updating several dependent states we apply one step of the mean field approximation.

3.2.1. A simple one parameter Poisson-Gamma SSM

First we illustrate how the Poisson-Gamma SSM works with a one-dynamic parameter example. Our football model is just a generalisation of the simple model described here.

The extension and measurement equations are given by

$$\begin{aligned} Y_t \mid \theta_t &\sim \text{Poisson}\,(\theta_t) && \text{(Measurement equation)} \\ \theta_t \mid y_{1:t-1}, \omega &\sim \text{Gamma}\,(\omega p_{t-1}, \omega q_{t-1}), && \text{(Extension of previous posterior)} \end{aligned}$$

where p_{t-1} and y_{t-1} are the hyper-parameters of the previous posterior and $\omega \in (0, 1]$ is a forgetting factor which adds uncertainty to the state process. In the next section we outline how the extension and the Bayesian update steps are implemented within our filter through single changes to the hyper-parameters of the dynamic Gamma states.

3.2.2. The extend or predict step of the filter

The extend step involves applying a convolution of transition equation to the previous posterior, described using the Chapman-Kolmogorov equation (Ross , 2014, Section 4.2) given by:

$$\underbrace{\pi(\theta_t \mid y_{1:t-1}, \omega)}_{\text{New prior}} = \int \underbrace{\pi(\theta_t \mid \theta_{t-1}, y_{1:t-1}, \omega)}_{\text{Transition equation}} \underbrace{\pi(\theta_{t-1} \mid y_{1:t-1}, \omega)}_{\text{Old Posterior}} \, d\theta_{t-1}. \tag{3.1}$$

Thus in the case of a Gamma state equation the extension is derived from the previous posterior by multiplying each hyper-parameter by ω conserving its mean of the dynamic parameter but increasing its variance.

$$\underbrace{\theta_{t-1} \mid y_{1:t-1} \sim \text{Gamma}\,(p_{t-1}, q_{t-1})}_{\text{Old Posterior}} \longrightarrow \underbrace{\theta_t \mid y_{1:t-1} \sim \text{Gamma}\,(\tilde{p}_t, \tilde{q}_t)}_{\text{New prior}} \tag{3.2}$$

The updates to the hyper-parameters in the extend step are

$$\tilde{p}_t \leftarrow \omega p_{t-1}, \quad \tilde{q}_t \leftarrow \omega q_{t-1}, \qquad \text{(Extend step)}$$

where the forgetting parameter lies in the range $\omega \in (0, 1]$.

We can prove using moment generating functions (see Appendix B) that to achieve the desired evolution (or extension), the transition equation must involve repeated multiplication by a draw from a scaled beta distribution of mean one. This can be formally stated as follows:

Theorem 1

Given the extensions given by Equations (3.2) there exists $Z_t > 0$, a scaled beta distribution, independent of θ_t, such that θ_t undergoes a multiplicative stochastic transition given by

$$\begin{aligned} W_t &\sim \text{Beta}\,(p_{t-1}\omega, p_{t-1}(1-\omega)). \\ Z_t &= \frac{W_t}{\omega}, \\ \theta_t &= \theta_{t-1} \times Z_t\ , \end{aligned} \qquad \text{(Transition Equation)} \tag{3.3}$$

where $W_t > 0$ is a random draw from the Beta distribution. The use of a scaled beta distribution for Z_t was also suggested by Gamerman et al. (2013).

3.2.3. *The update step of the filter*

In this step the extended prior is updated using the current observation. For the gamma prior and the Poisson probability of the next observation the posterior becomes

$$\begin{aligned} \pi(\theta_t \mid y_{1:t}, \omega) &\propto \pi(\theta_t \mid y_{1:t-1}, \omega)\ f(y_t \mid \theta_t) \\ &\propto \theta_t^{\tilde{p}_{t-1}+y_t-1} \exp(-\theta_t(\tilde{q}_{t-1}+1)), \\ \theta_t \mid y_{1:t}, \omega &\sim \text{Gamma}\,(\tilde{p}_{t-1}+y_t, \tilde{q}_{t-1}+1). \end{aligned}$$

Thus updates to the hyper-parameters are then:

$$\implies p_t = \tilde{p}_{t-1} + y_t, \quad q_t = \tilde{q}_{t-1} + 1. \qquad \text{(Update step)}$$

In this way conjugacy enables us to express the posterior distribution of the states given the current observations through manipulation of the hyper-priors. The predictive distribution can also be derived as

$$Y_t \mid y_{1:t-1}, \omega \sim \text{Negative-Binomial}\left(\tilde{p}_t, \tfrac{1}{\tilde{q}_t+1}\right). \tag{3.4}$$

Once the observation has been made the cumulative evidence for ω is updated as

$$Z_t(\omega) = f(y_{2:t} \mid y_1, \omega) = \prod_{t^*=2}^{t} f(y_{t^*} \mid y_{1:t^*-1}, \omega). \tag{3.5}$$

The evidence can be used to evaluate the predictions and carry out model selection.

3.2.4. *The backward smoothing recursion*

We have shown that filter can be expressed using the forward recursion for $t = 1, 2, \ldots, T$

$$\begin{aligned} \theta_t \mid y_{1:t}, \omega &\sim \text{Gamma}\,(p_t, q_t), \\ p_t &= \omega p_{t-1} + y_t, \\ q_t &= \omega q_{t-1} + 1. \end{aligned}$$

The posterior distribution, $\pi(\theta_t \mid y_{1:T}, \omega)$ can expressed using the following factorisation:

$$\begin{aligned} p\left(\boldsymbol{\theta} \mid y_T, \omega\right) &= p\left(\theta_T \mid y_T, \omega\right) \prod_{t=1}^{T-1} p\left(\theta_t \mid \theta_{t+1}, \ldots, \theta_t, y_t, \omega\right) \\ &= p\left(\theta_T \mid y_T\right) \prod_{t=1}^{T-1} p\left(\theta_t \mid \theta_{t+1}, y_t, \omega\right). \end{aligned}$$

Gamerman et al. (2013) show that how this factorisation can be used to sample from the posterior distribution using the filtered hyper-priors $\mathbf{p} = \{p_1, p_2, \ldots, p_T\}$ and $\mathbf{q} = \{q_1, q_2, \ldots, q_T\}$ in the backward recursion:

$$\begin{aligned} \theta_t \mid \mathbf{p}, \mathbf{q}, \omega &\sim \text{Gamma}\left(p_t, q_t\right), & t &= T \\ &\sim \text{Gamma}\left((1-\omega)p_t, q_t\right) + \omega\theta_{t+1}, & t &= T-1, \ldots, 1. \end{aligned}$$

In section 4 we illustrate how this recursion can be used to to plot the change in the posterior distribution of the HGA over the history of the EPL.

3.3. Evolution of the states in the Bayesian sequential football model

We now generalise the one parameter model of the previous section to create a filter for a multi-parameter model needed to describe changes to the state variable that describe the history of scores in a football league. In Section 2.4 we have shown how the update component of the filter can be carried out for such a model. The evolution step of our Bayesian SSM resembles that of the previous section but now we need to use several forgetting factors differentiating between the type of state and between the within season and between season forgetting.

3.3.1. Variation within seasons

In this section we modify the model of Section 2.3 to incorporate forgetting or the evolution of the states.

$$\begin{aligned} \begin{pmatrix} X_t \\ Y_t \end{pmatrix} &\sim \text{Poisson}\begin{pmatrix} \epsilon_t \alpha_{i,t} \beta_{j,t} \gamma_t, \\ \epsilon_t \alpha_{j,t} \beta_{i,t} \end{pmatrix}, && \text{(Observation Equation)} \\ \boldsymbol{\alpha}_R^s &\sim \frac{1}{\omega} \boldsymbol{\alpha}_{R-1}^s \text{Beta}\left(\omega \mathbf{p}_{R-1}^{\alpha}, (1-\omega)\mathbf{p}_{R-1}^{\alpha}\right), && \text{(Transition equation AS)} \\ \boldsymbol{\beta}_R^s &\sim \frac{1}{\omega} \boldsymbol{\beta}_{R-1}^s \text{Beta}\left(\omega \mathbf{p}_{R-1}^{\beta}, (1-\omega)\mathbf{p}_{R-1}^{\beta}\right), && \text{(Transition equation DS)} \\ \gamma_t &\sim \frac{1}{\omega_h} \boldsymbol{\gamma}_{t-1} \text{Beta}\left(\omega_h p_{t-1}^{\gamma}, (1-\omega_h) p_{t-1}^{\gamma}\right), && \text{(Transition equation HGA)} \\ \epsilon_t &\sim \text{Gamma}\left(\kappa, \kappa\right), && \text{(Random Effect)} \end{aligned} \tag{3.6}$$

where $t = 1, 2, \ldots, N_G$ is the game number and R is the round. The attacking strengths, $\boldsymbol{\alpha}_1^s$ and defensive strengths $\boldsymbol{\beta}_1^s$ for season, s, for the surviving sides are related to the last round of the previous season. From the above prior state equations we can derive an extend step in the same manner as in Section 3.2.2 in order to add uncertainty to the previous posterior. In this way, the extended posterior becomes the prior for the new observation. This is achieved by manipulating the hyper-priors of the dynamic parameters:

$$
\begin{aligned}
\boldsymbol{\alpha}_R &\sim \text{Gamma}\,(\tilde{\mathbf{p}}_R^{\alpha}, \tilde{\mathbf{q}}_R^{\alpha}), & \tilde{\mathbf{p}}_R^{\alpha} &= \omega \mathbf{p}_{R-1}^{\alpha}, & \tilde{\mathbf{q}}_R^{\alpha} &= \omega \mathbf{q}_{R-1}^{\alpha}, & &\text{(AS)}\\
\boldsymbol{\beta}_R &\sim \text{Gamma}\left(\tilde{\mathbf{p}}_R^{\beta}, \tilde{\mathbf{q}}_R^{\beta}\right), & \tilde{\mathbf{p}}_R^{\beta} &= \omega \mathbf{p}_{R-1}^{\beta}, & \tilde{\mathbf{q}}_R^{\beta} &= \omega \mathbf{q}_{R-1}^{\beta}, & &\text{(DS)}\\
\gamma_t &\sim \text{Gamma}\,(\tilde{p}_t^{\gamma}, \tilde{q}_t^{\gamma}), & \tilde{p}_t^{\gamma} &= \omega_h p_{t-1}^{\gamma}, & \tilde{q}_t^{\gamma} &= \omega_h q_{t-1}^{\gamma}. & &\text{(Home advantage)}
\end{aligned}
\tag{3.7}
$$

We construct an online model on the fourteen most recent seasons using the estimates obtained from the first 18 seasons of the EPL. These parameters are estimated in Section 3.3.3. In exploratory analysis we find that the within-season forgetting parameter varies between seasons leading us to believe that some seasons are more volatile than others. In addition there is also a possibility that the forgetting factor which expresses the volatility of the season also changes within the season.

3.3.2. *Variation between seasons*

To explain the dependence on the previous season we need to introduce several new (nuisance) parameters. Each season, for the surviving teams, we use a between season forgetting factor ω_b to describe the increase in uncertainty caused by the break between the seasons. We also assume that there is forgetting, both within the season and between the season in the influence of the HGA. For each of the three newly promoted sides we need four hyper-parameters, two to describe their initial attacking strengths (p^{α} and q^{α}) and two to describe their initial defensive strengths (p^{β}) and q^{β}). Exploratory analysis has indicated that the performance of a newly promoted side does not have a strong dependence on the promoted side's history in the lower EFL Championship league.

$$
\begin{aligned}
\boldsymbol{\alpha}_1^s &\sim \text{Gamma}\left(\tilde{\mathbf{p}}_1^{\alpha,s}, \tilde{\mathbf{q}}_1^{\alpha,s-1}\right), & \tilde{\mathbf{p}}_1^{\alpha,s} &= \omega_b \mathbf{p}_{N_R}^{\alpha,s-1}, & \tilde{\mathbf{p}}_0^{\alpha,s} &= \omega_b \mathbf{p}_{N_R}^{\alpha,s}, & &\text{(Surviving teams)}\\
& & &= p^{\alpha}, & &= q^{\alpha}, & &\text{(Promoted sides)}\\
\boldsymbol{\beta}_R^s &\sim \text{Gamma}\left(\tilde{\mathbf{p}}_R^{\beta,s}, \tilde{\mathbf{q}}_R^{\beta,s}\right), & \tilde{\mathbf{p}}_1^{\beta,s} &= \omega_b \mathbf{p}_{N_R}^{\beta,s-1}, & \tilde{\mathbf{q}}_1^{\beta,s} &= \omega_b \mathbf{q}_{N_R}^{\beta,s-1}, & &\text{(Surviving teams)}\\
& & &= p^{\beta}, & &= q^{\beta}, & &\text{(Promoted teams)}\\
\gamma_1^s &\sim \text{Gamma}\,(\tilde{p}_t^{\gamma,s}, \tilde{q}_1^{\gamma,s}), & \tilde{p}_1^{\gamma,s} &= \omega_{hb} p_{N_G}^{\gamma,s-1}, & \tilde{q}_1^{\gamma,s} &= \omega_{hb} q^{\gamma,s-1}.
\end{aligned}
$$

3.3.3. Offline estimation of nuisance parameters for the standard model

We use the first 18 seasons to estimate the following nuisance parameters so that online inference is possible for the seasons:

(a) The within season forgetting factor, assumed fixed over all seasons: ω.

(b) The common between season forgetting factor for the surviving teams: ω_b.

(c) The within season forgetting factor for the home ground advantage: ω_h.

(d) The between season forgetting factor for the home ground advantage: ω_{hb}.

(e) The correlation and over-dispersion parameter: κ.

(f) The hyper-priors for the initial attacking and defensive strengths over the three newly promoted sides $p^\alpha, q^\alpha, p^\beta, q^\beta$ where the initial attacking strength is given by $\alpha \sim \text{Gamma}\,(p^\alpha, q^\alpha)$ and the initial defensive strength is described by $\beta \sim \text{Gamma}\left(p^\beta, q^\beta\right)$.

We find the value of plausible values for these parameters by maximising the out-of-sample multinomial log-likelihood discussed in Section 3.4 for the first 18 seasons of the EPL. This calculation of the nuisance parameter on the training set took about 20 minutes on a desktop PC. The values of these parameters in Table 2 were used to test the accuracy of our predictions. We found that it was not necessary to re-impose the identifiability constraint of $\beta_1 = 1$ after the first season, and dropping the constraint produced better predictions.

3.4. Evaluation of out-of-sample predictive performance

Scoring methods are an established way of evaluating probabilistic predictions. The theory has its origin in meteorology (Brier, 1950). We use them to quantify the ability of our model to make out-of-sample predictions of the outcome of the next game. If we wished to predict the match score rather than the outcome we would have used a score of $-\log(f(x_t, y_t \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \kappa))$ given by Equation (2.3). We now outline three scoring methods for measuring the accuracy of predictions of outcomes within the seasons using data only from that season. Let $\mathbf{P}_t = [P_{1,t}, P_{2,t}, P_{3,t}]^T$ denote the estimated probability of the three possible outcomes which are denoted by a row of the matrix below

$$\mathbf{Z}_t \in \begin{pmatrix} (1, & 0, & 0), & \text{HW} \\ (0, & 1, & 0), & \text{DR} \\ (0, & 0, & 1), & \text{AW} \end{pmatrix}.$$

We compute the Brier Score (BS), (Brier, 1950), the out-of-sample multinomial log-likelihood or log-score (LS) and the ranked probability score (RPS), (Murphy, 1970). The only score that takes into account the ordering of the outcomes is the RPS and the only one that is local (in the sense that it takes the probability of an event into account) is the log score (LS). A low score from any of these statistics indicates a model that predicts the outcome well. Our experience is that in the case of football these scoring methods show a high degree of agreement for this application.

$$\mathrm{BS} = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{3}(z_{j,t} - P_{j,t})^2$$

$$\mathrm{LS} = -\frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{3} z_{j,t}\log(P_{j,t}) \tag{3.8}$$

$$\mathrm{RPS} = \frac{1}{2n}\sum_{t=1}^{n}\sum_{k=1}^{2}\left(\sum_{j=1}^{k}(z_{j,t} - P_{j,t})\right)^2 \tag{3.9}$$

In this section we compare the predictions of the proposed state space models, the dynamic score-driven time series model of Koopman & Lit (2019) and different models using weighted likelihood.. For the weighted likelihood models we estimate the decay parameter of the weights by finding the value that minimizes the RPS of predictions for the univariate Poisson model over the seasons 1996/97 to 2009/10, inclusive. This resulted in a decay constant of $k = .001824$ corresponding to a half-life of 380 days.

Details of our implementation of Koopman & Lit (2019) are given in Section S4 of the Supplementary Materials.

With the Bayesian models we did not retain the data from the relegated sides to be used again when they are promoted again whereas in the weighted likelihood models we did. The offline estimates of the nuisance parameters for the bivariate Bayesian sequential model are given in Table 2.

We compare six weighted likelihood methods (where the predictions are recalculated for every round for every season apart from the first 5 rounds) with a score driven model, and three versions of our Bayesian state space models. We calculate the cumulative RPS (relative to the predictions implied by the bookmakers' average odds) which are given the following abbreviations as

UP : Weighted univariate Poisson model (Section 2.1).

BNB : Weighted bivariate negative binomial model (Section 2.2).

DC : Weighted model of Dixon & Coles (1997).

BP : Weighted bivariate Poisson (Karlis & Ntzoufras, 2003).

DAV.HGA : Weighted Davidson model including HGA (Davidson, 1970).

RK.HGA : Weighted Rao Kupper model including HGA (Rao & Kupper, 1967).

KL : The Koopman and Lit dynamic model (Koopman & Lit, 2019).

Table 2. The maximum likelihood estimates (MLE) of the nuisance parameters of the three Bayesian models displayed at the end of Section 3.4 made by maximising the log-score of Equation (3.8) using the training set of the first 18 seasons of the EPL. To speed up the training of the bivariate models we reused the estimates of the last 4 parameters obtained from the univariate model and treated these as fixed.

	ω_h	ω	ω_b	ω_{hb}	κ	p^α	q^α	p^β	q^β
Bayes.UV.VB SSM	.999	.988	.770	.865		19.3	23.9	30.0	26.4
Bayes.BV.Ax SSM	.999	.985	.795	.860	6.783	19.3	23.9	30.0	26.4
Bayes.BV.VB SSM	.999	.987	.737	.911	6.323	19.3	23.9	30.0	26.4

Bayes.BV.VB : The dynamic bivariate variational Bayes model.

Bayes.UV.VB : The dynamic univariate variational Bayes model.

Bayes.BV.Ax : Approximation using one step of the dynamic bivariate variational Bayes model.

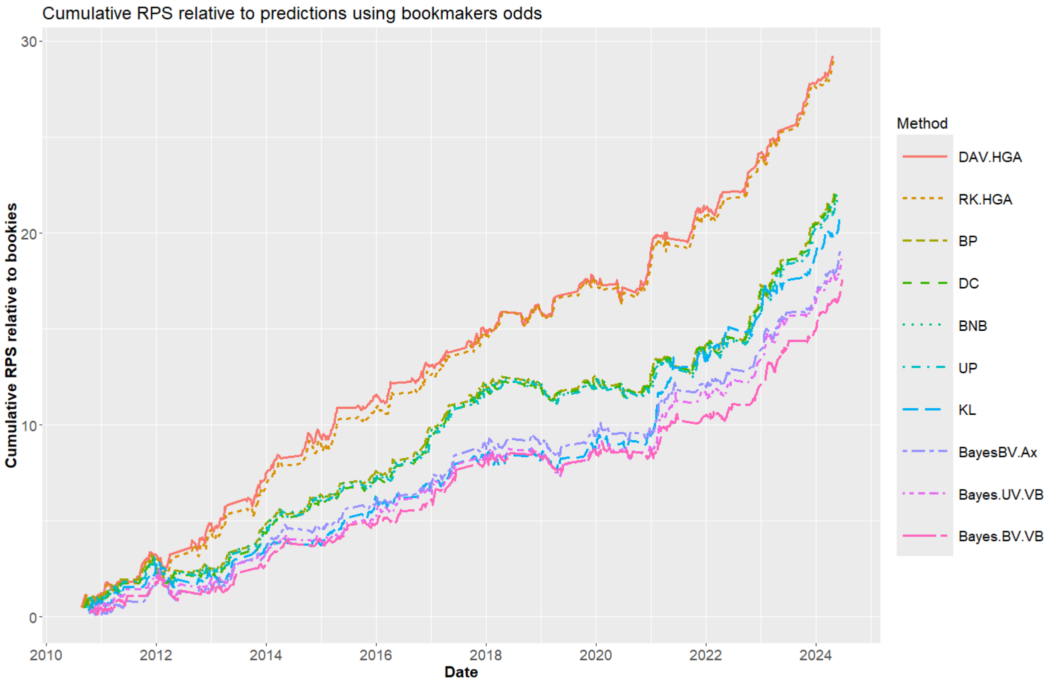


Fig. 3. A comparison of the cumulative RPS measure of the prediction accuracy (relative to those obtained using the the odds offered by the bookmakers) of three dynamic models, a score driven model and six others that use a weighted likelihood (where the predictions are recalculated for every round for every season apart from the first 5 rounds.) The training set were the 1992/93 to 2009/10 seasons and the test set 2010/11 to 2023/24.

Table 3. The cumulative RPS scores of the models displayed at the end of Section 3.4 relative to the bookmakers odds over the test set; 2010/11 to 2023/24. The dashed horizontal lines separate the methods into groups of similar accuracy of predictions. In the third group, the weighted likelihood score based models, gave barely distinguishable differences.

Method	Cumulative RPS
Bayes.BV.VB	17.55
Bayes.UV.VB	18.64
Bayes.BV.Ax	19.16
KL	20.71
UP	22.06
BNB	22.12
DC	22.22
BP	22.24
RK.HGA	28.96
DAV.HGA	29.23

Figure 3 shows a comparison of the online predictions measured by RPS over multiple seasons. Over the 14 year test set none of the models outperformed the predictions made by using the bookmakers odds so all of our RPS scores were calculated relative to the match probabilities implied by the bookmakers' average odds. The final cumulative RPS scores over the test set are shown in Table 3. The raw RPS scores to 5 decimal places for the each season from the test set are shown in Table S1 of the Supplementary Materials.

The scores fall into roughly four groups of similar strength of predictions The three Bayesian SSMs appear to be the best predictors followed by the score-driven time series method of Koopman & Lit (2019) which gave the next best predictions. We note that this model did particularly poorly in the 2020/21 season, which appears to be due to the method's assumption of a fixed HGA. Restrictions due to Covid-19 led to matches being played without fans in attendance which led to HGA falling dramatically (See Figure 6). The third group are the four weighted likelihood methods based on the scores of the matches and gave almost identical predictions. The last group are the weighted likelihood outcome based models which gave the least favourable predictions of all, These consisted of the models of Davidson (1970), generalised to include a HGA, and the similar model based on Rao & Kupper (1967), Each of the weighted models took a substantial amount of computational time to train because each time a round of games was observed the numerical optimiser had to be run to convergence at each round on an expanded data-set. None of the Bayesian methods nor the model of Koopman & Lit (2019) need to do this and are consequently a lot more efficient.

A further approach used the assess the predictive ability of a football model is to consider long term predictions of the final league positions. In Section S6 of the Supplementary Materials this is implemented for the one step approximation method using the 2016-17 season.

3.4.1. Calibration of model

We examine the ratio of predicted and observed outcomes over the games that have been played over the history of the EPL. Table 4 shows that the model slightly under-predicts the number of home wins in the test set (≈-3%) and over-predicts the number of draws (≈8%). In the training set there is also some mis-calibration, but to a much smaller degree.

Table 4. Ratio of predicted outcomes by season (home losses, draws and home wins) divided by the total of the actual outcomes of the games of the same season.

Season	Outcome.ploss	Outcome.pdr	Outcome.pwin
Mean.train	1.0426	1.0195	0.9748
Mean.test	0.9716	1.0832	0.9931

3.5. Residual diagnostics of the standard Bayesian models

Two components of the diagnostics: prediction and calibration, have been discussed in Sections 3.4. We now carry out diagnostic tests to identify any possible deficiencies in the proposed model in Section 3.3. The first problem is the existence of extra-Poisson-variation evident in the occurrence of matches where a team scores a large number of goals. Over-dispersion can be brought on by games where there is a mismatch of abilities. Lopsided scores can occur, for instance, when one of the sides has been reduced to 10 men after a red card has been issued (Titman et al., 2015). The second problem is that of dependence between the scores which can be explored by examining the correlations between the residuals on the univariate model. In order to resolve these problems we define various types of residuals for both the univariate and bivariate sequential models. The only difference between these SSMs is in the measurement equations which are Equations (2.1) for the univariate model and Equations (2.2) for the bivariate model.

3.5.1. Residuals from the univariate model

The standardised Pearson residuals for this model needed for checking the model can be expressed as

$$\mathcal{R}_t^U = \left(\frac{x_t - \lambda_t^H}{\sqrt{\lambda_t^H}}, \frac{y_t - \lambda_t^A}{\sqrt{\lambda_t^A}} \right)^T . \tag{3.10}$$

A residual designed to identify runaway or lopsided scores can be made by using the concept of a predictive p-value (ppv)(Gelman, 2013). We do this by calculating the probability that the absolute difference in scores would exceed the actual difference, using the Skellam distribution (Skellam, 1946), which are constructed from conditionally independent Poisson distributions given the predicted scores for the home and away side:

$$\text{ppv}_t = \mathbb{P}\{|X_t - Y_t| > |x_t - y_t| \mid \lambda_t^H, \lambda_t^A\}.$$

Because the ppvs are often very small we re-express the them in terms of surprise in the difference of scores which we define as

$$\mathcal{R}_t^S = -\log(\mathrm{ppv_t}). \tag{3.11}$$

3.5.2. Residuals from the bivariate model

The residuals for the bivariate model can be expressed as

$$\mathcal{R}_t^B = \left(\frac{x_t - \lambda_t^H}{\sqrt{\lambda_t^H + \frac{(\lambda_t^H)^2}{\kappa}}}, \frac{y_t - \lambda_t^A}{\sqrt{\lambda_t^A + \frac{(\lambda_t^A)^2}{\kappa}}} \right)^T . \tag{3.12}$$

Yet another residual of the bivariate model, suitable for ordering the outliers can be defined as

$$\mathcal{R}_t^C = \begin{bmatrix} x_t - \lambda_t^H \\ y_t - \lambda_t^A \end{bmatrix}^T \begin{bmatrix} \lambda_t^H + \frac{(\lambda_t^H)^2}{\kappa} & \frac{\lambda_t^H \lambda_t^A}{\kappa} \\ \frac{\lambda_t^H \lambda_t^A}{\kappa} & \lambda_t^A + \frac{(\lambda_t^A)^2}{\kappa} \end{bmatrix}^{-1} \begin{bmatrix} x_t - \lambda_t^H \\ y_t - \lambda_t^A \end{bmatrix}, \tag{3.13}$$

where the $\mathcal{R}_t^C$ should have an approximate χ_2^2 distribution if the model is correctly specified.

3.5.3. Comparisons between the univariate and bivariate models

In this section we illustrate improvements in the problem of overdispersion can be made when using the bivariate model (Equation 3.12) rather than the univariate model (Equation 3.10). These improvements in the residuals for the 2024 season are displayed as arrows in Figure 5 and are ordered by their combined residual of Equation (3.13). The same ordering of the 20 most outlying scores and the corresponding details of the teams, dates and scores are shown in Table 4. This table compares the out of sample residuals from the sequential univariate model relative to those from the sequential bivariate model. The order of size of the residuals using Equation (3.13) are also shown in table 5 together with the residuals formulated in Section 2.2.

3.5.4. Variation of correlation and overdispersion season by season

The residual plots in Figure 5 illustrate the two problems with the univariate Poisson model. The lower panel shows that the correlation of the standardised residuals with an average of about $\rho \approx .05$. These correlations are almost always positive. The 2019/20 season stands out because in that case there was a substantial negative correlation. The upper panel illustrates the problem of over-dispersion. It depicts counts of the number of outlying scores where the standardised residuals were greater than 3. We refer to these scores as outliers which potentially have too much influence on the estimates for the model. There are an average of around 5 outlying observations per season. (If the standardised residuals did have the standard normal we would expect only two games from 380 games to have over dispersed observations.)

Table 5. The table compares the out of sample residuals after running the sequential univariate and bivariate models $\mathcal{R}_t^U$ are the home and away univariate residuals calculated using Equations (3.10). $\mathcal{R}_t^S$, an indicator of lopsided or runaway scores, is the surprise at the difference of scores given the teams abilities calculated using Equation (3.11). The columns headed $\mathcal{R}_t^B$ are the home and away bivariate residuals calculated using Equation (3.12). The observations are ordered by a bivariate combined residual, $\mathcal{R}_t^C$, which is calculated using Equation (3.13).

	Date	Home	Away	HG	AG	$\mathcal{R}_t^U$		$\mathcal{R}_t^B$		$\mathcal{R}_t^C$	$\mathcal{R}_t^S$
1	2023-09-24	Sheffield Utd	Newcastle	0	8	-1.013	5.366	-.866	4.196	29.445	9.646
2	2023-09-30	Aston Villa	Brighton	6	1	3.577	-.238	3.243	-.283	18.748	6.263
3	2023-11-12	Chelsea	Manchester City	4	4	2.969	1.810	2.827	1.380	17.381	1.470
4	2024-02-03	Newcastle	Luton	4	4	1.136	3.122	.818	3.038	17.073	1.117
5	2024-04-15	Chelsea	Everton	6	0	2.927	-.931	2.709	-.890	15.555	6.247
6	2023-12-27	Brentford	Wolves	1	4	-.461	3.363	-.443	3.070	15.537	4.116
7	2024-02-11	West Ham	Arsenal	0	6	-1.146	3.269	-1.030	2.644	15.333	5.792
8	2024-02-04	Chelsea	Wolves	2	4	.189	3.443	.163	2.902	14.791	2.895
9	2024-05-19	Crystal Palace	Aston Villa	5	0	3.021	-1.179	2.607	-1.064	14.156	6.828
10	2024-02-01	Wolves	Manchester Utd	3	4	1.779	2.477	1.455	2.225	13.995	.940
11	2023-12-02	Burnley	Sheffield Utd	5	0	2.718	-1.083	2.599	-.961	13.990	6.082
12	2024-04-21	Crystal Palace	West Ham	5	2	2.968	.674	2.560	.564	13.980	3.759
13	2024-04-20	Luton	Brentford	1	5	-.440	2.670	-.362	2.432	12.473	3.617
14	2023-12-10	Fulham	West Ham	5	0	2.639	-1.143	2.300	-1.034	12.271	6.038
15	2023-12-03	Liverpool	Fulham	4	3	.900	2.573	.719	2.376	12.266	.789
16	2023-08-12	Newcastle	Aston Villa	5	1	2.812	-.042	2.379	-.040	12.214	4.714
17	2023-12-05	Luton	Arsenal	3	4	2.183	1.358	2.159	1.010	11.986	1.015
18	2023-11-06	Tottenham	Chelsea	1	4	-.544	2.694	-.543	2.481	11.866	3.516
19	2023-12-06	Fulham	Nott'ham Forest	5	0	2.480	-.990	2.208	-.916	11.619	5.487
20	2024-01-30	Luton	Brighton	4	0	2.428	-1.283	2.302	-1.152	11.364	5.961

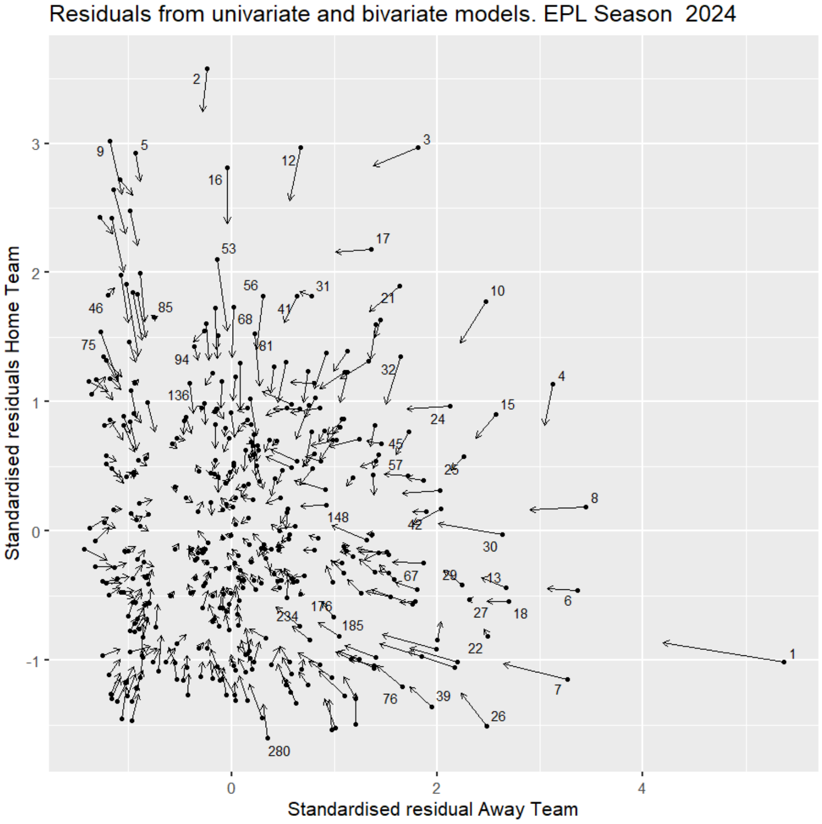


Fig. 4. The diagram shows the improvement in standardised residuals of the sequential bivariate model relative to those of the sequential univariate model. The dots represent the univariate residuals calculated using Equation (3.10) and the tips of the arrows show the bivariate residuals calculated using Equations (3.12). The numbers (next to the point) shows the ordering of the size of the residuals calculated using Equation (3.13).

Our bivariate model allows for both for a degree of overdispersion and positive correlation between the home and away scores. The magnitude of these effects is governed by the parameter, κ, which we assume as constant over all seasons and estimated over the training set using variational Bayes as $\hat{\kappa} = 6.232$. Figure 5 sheds some light on that assumption. A rule of thumb that can be deduced from this the overdispersion of the univariate model where $PV_{CS} < .05$ when the variance of the univariate residuals $VAR_U > 1.1$. For these season the variance of the bivariate model $VAR_B \approx 1$ and the overdispersion problem is reduced appropriately.

The table displays some of the diagnostics from bivariate SSM model using the MLE parameters obtained from the variational Bayes method from Table 2. The columns $NOUT_U$ and $NOUT_B$ shows the number of outlying scores in the univariate and bivariate models where the standardised residuals (calculated using Equations (3.10) and (3.12))

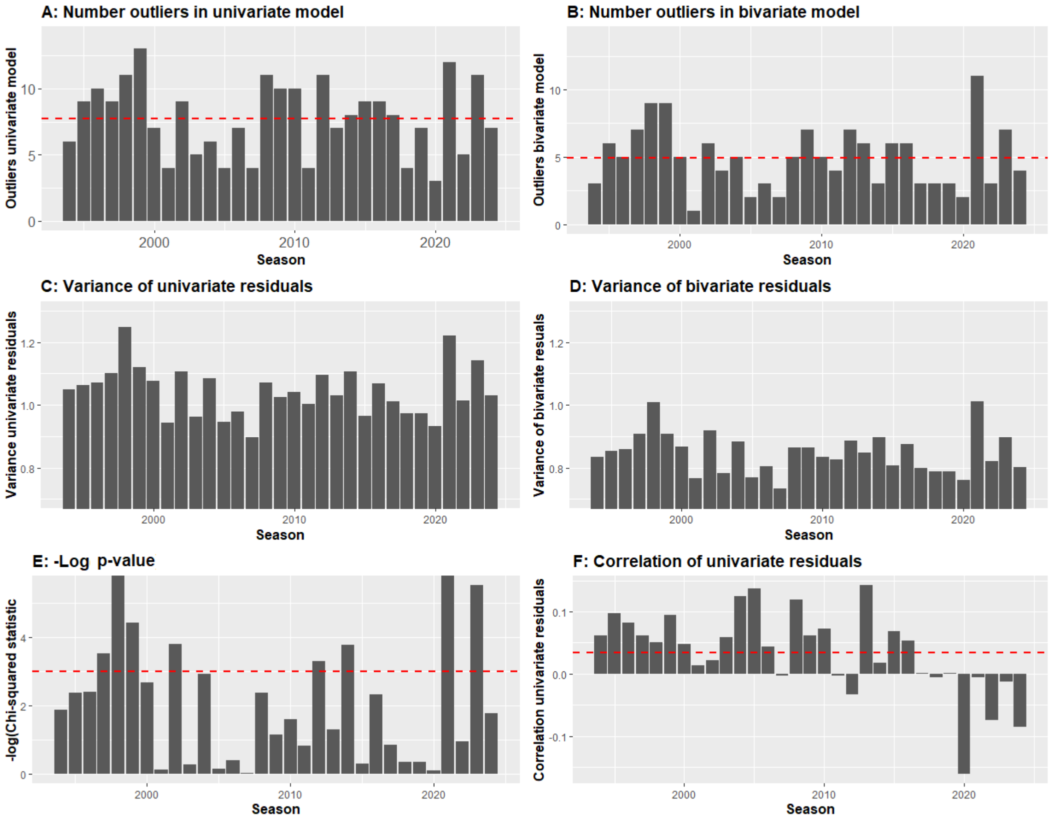
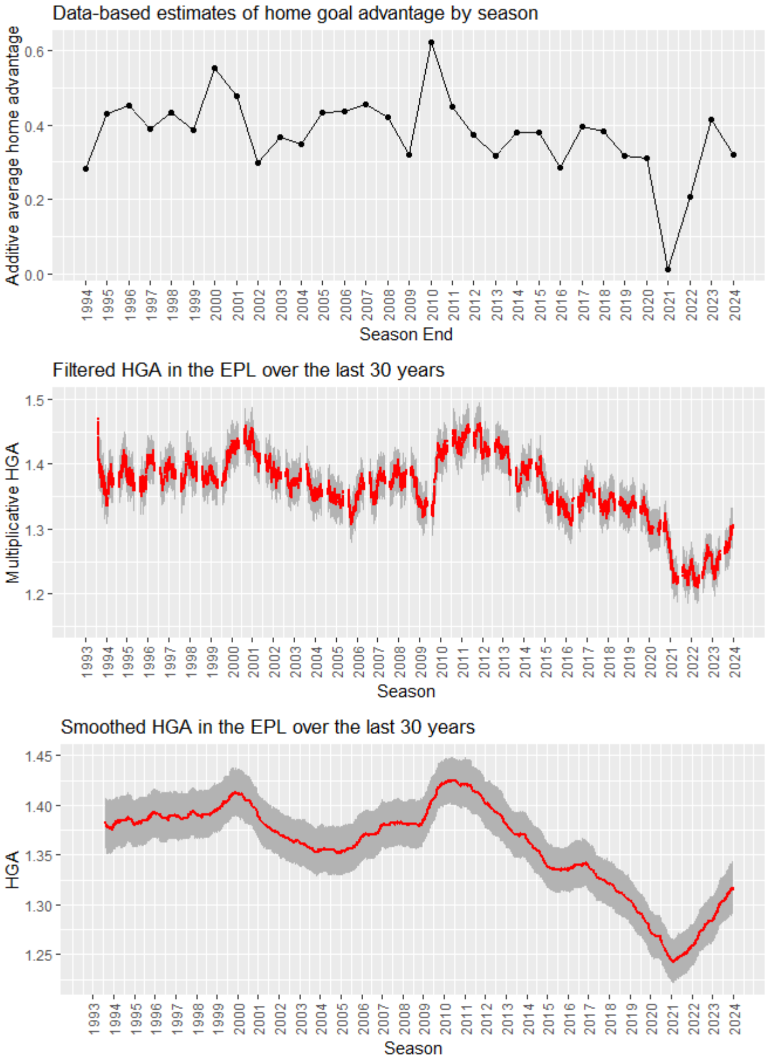


Fig. 5. Panels A, C and E show the problems of overdispersion in the univariate model. Panel F shows the correlation between the univariate residuals. Panels B and D show that the problem of overdispersion has been reduced in the bivariate model. The correlation between the residuals of the bivariate model is unchanged from Panel F. The horizontal dashed lines in panels A, B and F are averages taken over all seasons. The horizontal line in Panel E is the level above which the Chi-squared statistic for extra-Poisson dispersion is significant at the 0.05 level.

were greater than 3. The variance of the standardised residuals of the univariate model is shown under the column name VAR_U and the p-values of a Chi-squared statistic for extra-Poisson dispersion is shown as PV_{CS}. The column VAR_B shows the variance of the bivariate residuals derived from Equation (3.12) and COR_U shows the correlation of the home and away univariate residuals.

4. The evolution of the teams abilities and HGA

A strength of a Bayesian SSM is its ability to efficiently filter and smooth the states of the model. In our case the states are the defensive and attacking strength of each side and the home goal advantage (with measures of their uncertainty) both within and between seasons. See Figures 7 and Figure 8 for examples. A backward recursion can then be applied to these filtered states and the posterior distribution of the states can be sampled from using the method of Gamerman et al. (2013) and summarised in Section 3.2.4. The bottom panel of Figure 6 illustrates how this backward smoothing recursion can be used to trace the changing posterior distribution of the HGA. The data based estimates of the HGA and the filtered HGAs are shown in the top two panels for comparison. The data based estimate is just the ratio of the average number of goals scored at home over the average number of goals scored away. In particular, notice the drop in home advantage over the 2020/21 season which was a season mostly played without crowds. From Figure 3 it can be observed that the rate of increase in cumulative RPS in 2021 and 2022 was smaller for the methods that accommodated a changing HGA compared to the method of Koopman & Lit (2019).



Data-based estimates of home goal advantage by season

Filtered HGA in the EPL over the last 30 years

Smoothed HGA in the EPL over the last 30 years

Fig. 6. The top panel shows data-based estimate of the HGA (top panel) by calculating the difference between the home and away goal average for each season. The middle panel show the 75 % quantiles of the filtered model-based home ground advantage. The last panel shows a smoothed 75% credible interval to represent the posterior distribution of the HGA using the method of Gamerman et al. (2013) explained in Section 3.2.4.

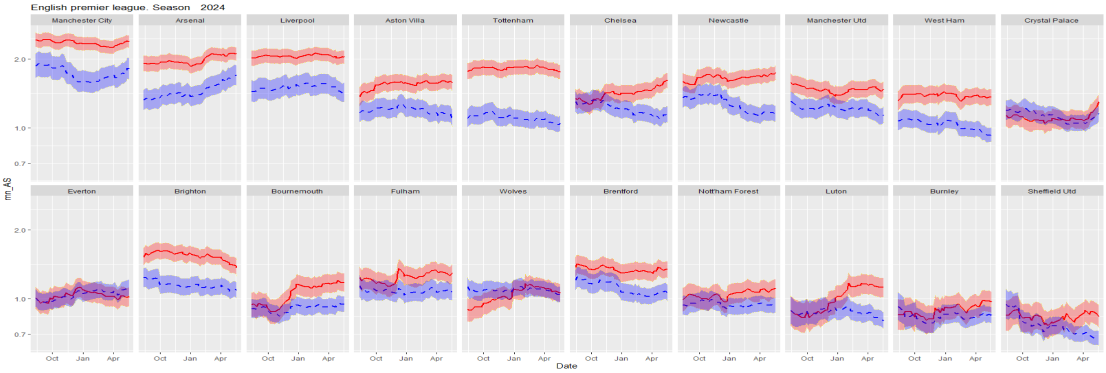


Fig. 7. 75% credible intervals showing the the filtered attacking strengths (solid line) and defensive strengths(dashed lines) of all teams in the 2023-4 season of the EPL.

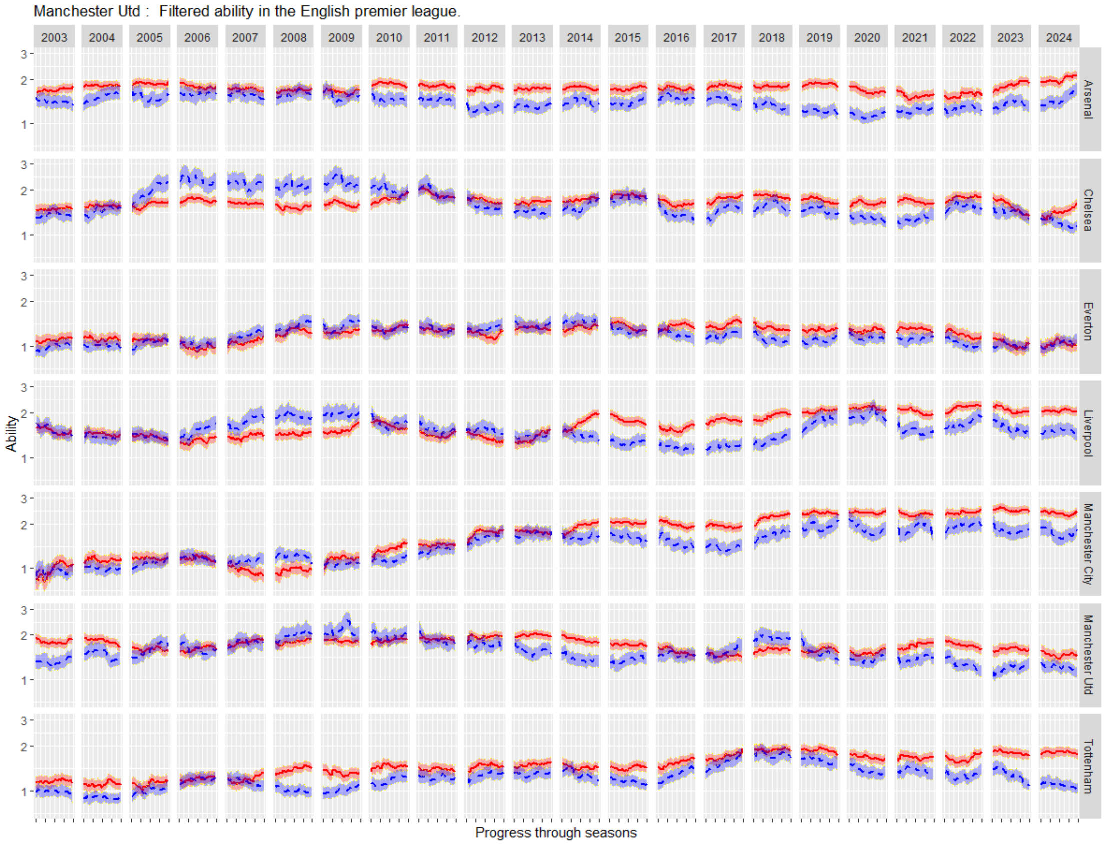


Fig. 8. 75% credible intervals showing the the filtered attacking strengths (Solid line) and the defensive strengths(Dashed lines) of the top clubs over the last 22 years of the EPL.

5. Conclusion

In this paper, we model the outcome of a football game using a Bayesian SSM and update and predict using a Bayesian filter. The choice of a SSM and prediction from it using a filter should be an attractive one for a sports modeller particularly because their primary purpose is usually to predict ahead in the presence of the inevitable changes that occur within and between seasons. Every time the result of a game has been observed, the relevant dynamic parameters for the model are updated using only that observation. Unlike the likelihood or a weighted likelihood approach the rest of the data does not need to be reprocessed each time the dataset is enlarged. If these steps can be accurately and economically carried out then a SSM should always out-perform a weighted likelihood model in terms of accuracy and economy of prediction.

The Kalman filter (Kalman, 1960) is a method of predicting and controlling for a particular type of SSM model when the measurement and state models are both linear and use Gaussian distributions for state and measurement noise, each with known and unchanging variances. The two steps of the Kalman filter can be expressed using closed formed expressions and their calculation is almost immediate. Glickman & Stern (1998) adapt the Kalman filter to create a Bayesian SSM to predict outcomes in the American National Football League and use Gibbs' sampling (Gelfand & Smith, 1990) to sample from the parameters of their expanded model.

Adopting a bivariate negative binomial, rather than independent Poisson, model gave slightly better predictions on the basis of RPS. Interestingly, for the weighted likelihood methods there was virtually no difference between the two models and the maximum weighted likelihood estimate of κ was much higher than that estimated for the Bayesian SSM. One possible explanation for this discrepancy is that in the former case, κ was directly chosen to minimize RPS. Whereas in the latter, only k (the parameter to determine the exponential weights) was optimized with respect to out-of-sample RPS, while the model parameters including κ were fitted using the likelihood of the exact match scores. Choosing a lower κ may improve RPS by shrinking the estimated probabilities of home or away wins away from 1 in cases where the match is expected to be one-sided. This may indirectly account for the uncertainty in the estimated model parameters.

Our method has similarities to that of Koopman & Lit (2019) in being a dynamic model with computationally tractable updates. In the EPL example, our method outperforms theirs but this appears to be mainly due to their assumption of a static home ground advantage, and it would be relatively straightforward to relax that assumption within their framework. Nevertheless, our approach differs in retaining the structure of a full state-space model which means that, in addition to forecasting future match outcomes, credible intervals can also be calculated for dynamic parameters such as a team's ability at a given point in time t, either based only on data up to time t (the filter) or using all data to some end time $T > t$ (the smooth).

A problem with sequential football models that use the attacking and defensive strengths to model the score is the high degree of dependence between these quantities. Given a score, the attacking strengths of the home side cannot be updated without first considering the defensive strengths of the opposing sides and the magnitude of the HGA. Furthermore, the joint update for these dependent parameters does not have a closed form expression. A traditional Bayesian approach to updating is to use MCMC simulations from the full conditional posteriors of each variable until convergence is reached. However, this would also induce a significant and an unacceptable computational load to our algorithm. Instead we use the mean field approximation (Jordan et al., 1999) which avoids the need for simulation. Moreover, we make a further approximation which avoids the need for convergence by making just one step single updates of the MFA by exploiting the fact the dynamic variables, form and HGA, are not going to change radically from previous games. Effectively this means when updating one variable we use the prior expectations of the other variables to condition on.

A computationally-efficient state-space model is achieved by representing the states via multiplicative Gamma distributions exploiting conjugacy, by assuming an extend step that produces a simple 'forgetting' structure and by assuming a mean field approximation for the form of the posterior distribution. Each of these steps potentially limits the general applicability of our approach to some degree.

The use of multiplicative Gamma distributions on the mean response is not qualitatively different from the more standard approach of using additive normal distributions with respect to the log-mean response as adopted by Koopman & Lit (2015), particularly once the parameters are reasonably large. However, while binary or categorical covariates, such as home ground advantage, can be included in the model the approach is not readily extendable to continuous covariates.

In order to achieve the simple 'forgetting' extend step, it is necessary for the multiplicative error in the transition process to be partially dependent on the observed data rather than independent. Specifically, the scaled beta distribution used in Appendix A depends on the current shape parameter, p_t, of the respective Gamma distribution, which depends on the number of previous goals observed. Effectively, this means the variability in the transition process decreases with more data. However, since $p_t = \omega p_{t-1} + X_t$, where ω is the forgetting parameter and X_t is the relevant goal variable, then assuming the X_t are generated from a stationary process such that $\lim_{t->\infty} t^{-1} \sum_{i=1}^{t} X_i = \mu^*$, $p_t \rightarrow \mu^*/(1-\omega)$ as $t \rightarrow \infty$. Hence the transition process is approximately independent of the data observed once a reasonable amount of data has been collected.

The mean field approximation was shown to match the marginal distributions of the parameters obtained through full Gibbs sampling, and the method gave good predictions of match outcomes. However, inference on the joint posterior distribution of parameters will not tend to be accurate since the assumption of posterior independence will not necessarily hold.

Data Availability

All of the data used in this paper is taken from the URL `https://football-data.co.uk/data.php`.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

No external funding.

Appendix A. The bivariate negative binomial model.

Conditional on the shared random effect, ϵ_t, the teams' scores X_t and Y_t are independent Poisson with rates $\epsilon_t \lambda_t^H$ and $\epsilon_t \lambda_t^A$. As a consequence, the (marginal) match outcome probabilities can be computed by numerically integrating the conditional match probabilities (implied by the Skellam distribution) over the Gamma distribution of ϵ_t.

The marginal variance, covariance and correlations of the bivariate negative binomial distribution are given by

$$
\begin{aligned}
\operatorname{Var}(X_t) &= \mathrm{E}\left[\operatorname{Var}\left[X_t \mid \lambda_t^H, \epsilon_t\right]\right] + \operatorname{Var}\left[\mathrm{E}\left[X_t \mid \lambda_t^H, \epsilon_t\right]\right] \\
&= \mathrm{E}\left[\epsilon_t \lambda_t^H\right] + \operatorname{Var}\left[\epsilon_t \lambda_t^H\right] \\
&= \lambda_t^H + \frac{(\lambda_t^H)^2}{\kappa}. \\
\operatorname{Cov}(X_t, Y_t) &= \mathrm{E}\left[\operatorname{Cov}[X_t, Y_t \mid \lambda_t^H, \lambda_t^A, \epsilon_t]\right] + \operatorname{Cov}\left[\mathrm{E}\left[X_t, Y_t \mid \lambda_t^H, \lambda_t^A \epsilon_t\right]\right] \\
&= \mathrm{E}\left[\operatorname{Cov}[X_t, Y_t \mid \lambda_t^H, \lambda_t^A, \epsilon_t]\right] + \operatorname{Cov}\left[\mathrm{E}\left[X_t \mid \lambda_t^H, \epsilon_t\right], E[Y_t \mid \lambda_t^A, \epsilon_t]\right] \\
&= 0 + \operatorname{Cov}[\epsilon_t \lambda_t^H, \epsilon_t \lambda_t^A] \\
&= \lambda_t^H \ \lambda_t^A \operatorname{Var} \epsilon_t \\
&= \frac{\lambda_t^H \lambda_t^A}{\kappa}. \\
\operatorname{Cor}(X_t, Y_t) &= \frac{\operatorname{Cov}(X_t, Y_t)}{\sqrt{\operatorname{Var}(Y_t)\operatorname{Var}(X_t)}} \\
&= \frac{\lambda_t^H \lambda_t^A}{\sqrt{(\kappa \lambda_t^H + (\lambda_t^H)^2)(\kappa \lambda_t^A + (\lambda_t^A)^2))}}.
\end{aligned}
$$

Appendix B. Derivation of the transition equation

THEOREM 1

If the posterior of the previous observation and its extended posterior are defined by

$$\begin{aligned} \lambda_{t-1} &\sim \text{Gamma}\,(p,q) && \text{(Previous posterior)} \\ \lambda_t &\sim \text{Gamma}\,(p\omega,q\omega) && \text{(Extension)} \end{aligned}$$

for $p, q > 0$ and $0 < \omega \leq 1$. Then $\exists\ Z_t > 0$ independent of λ_t such that λ undergoes a multiplicative stochastic transition process given by

$$\lambda_t = \lambda_{t-1} \times Z_t, \quad Z_t > 0, \tag{5.1}$$

where $Z_t = \frac{W_t}{\omega}$ and $W_t \sim \text{Beta}\,(p\omega, p(1-\omega))$.

PROOF

From Equation (5.1)

$$\begin{aligned} \log(\lambda_t) &= \log(\lambda_{t-1}) + \log(Z_t) \\ \Longrightarrow \mathrm{M}_{\log\lambda_t}(u) &= \mathrm{M}_{\log\lambda_{t-1}}(u)\ \ \mathrm{M}_{\log Z_t}(u), \end{aligned}$$

where $\mathrm{M}_V(u) = \mathrm{E}\,[e^{Vu}]$ denotes the moment generating function of V. We now derive the moment generating function of the distribution of $\log(\lambda_{t-1})$. Let $x \sim \text{Gamma}\,(p,q)$ then

$$\begin{aligned} \mathrm{M}_{\log x}(u) &= E(e^{u\log x}) = E(x^u) \\ &= \frac{\Gamma(p+u)}{\Gamma(p)q^u} \end{aligned}$$

$$\begin{aligned} \mathrm{M}_{\log Z_t}(u) &= \frac{\mathrm{M}_{\log\lambda_t}(u)}{\mathrm{M}_{\log\lambda_{t-1}}(u)} \\ &= \frac{\Gamma(p\omega+u)\,\Gamma(p)}{\Gamma(p\omega)\,\Gamma(p+u)\,\omega^u}. \end{aligned}$$

Next we show that this is the MGF of a scaled beta distribution. We let $W \sim \text{Beta}\,(\alpha,\beta)$ and let $V = \log W$

$$\begin{aligned} \mathrm{E}\,(e^{u\log W}) &= \mathrm{E}\,(W^u) \\ &= \frac{B(a+u,b)}{B(a,b)}. \end{aligned}$$

Setting $T = V + k$ for some constant k, gives

$$\begin{aligned} \mathrm{M}_T(u) &= \mathrm{M}_v(u)e^{ku} \\ &= \frac{B(a+u,b)e^{ku}}{B(a,b)}. \end{aligned}$$

Setting $a = p\omega$, $a + b = p$, and $k = \log(\omega)$ gives

$$\begin{aligned}\mathrm{M}_{T}(u) &= \frac{B(p\omega + u, p(1-\omega))}{B(p\omega, p(1-\omega))B(a,b)\omega^u} \\ &= \frac{\Gamma(p\omega + u)\Gamma(p)}{\Gamma(p+u)\Gamma(p\omega)\omega^u} \\ &= \mathrm{M}_{\log Z_t}(u).\end{aligned}$$

Hence $e^T = \frac{W}{\omega}$ where $W \sim \mathrm{Beta}\,(p\omega, p(1-\omega))$, as required.

Appendix C. Variational Bayes approximation

In this section we show that the proposed updates in Section 2.4 is approximately the same as making a mean-field variational approximation to the joint distribution at each update.

Let X_t and Y_t represent the home and away goals scored in a given match. In addition, let $\theta_t = (\alpha_t^H, \alpha_t^A, \beta_t^H, \beta_t^A, \gamma_t, \epsilon_t)$ represent the relevant parameters governing (X_t, Y_t). We can further note that $\mathcal{R}_1 = \{1, 4, 5, 6\}$ and $\mathcal{R}_2 = \{2, 3, 6\}$ give the set of indices of parameters relevant to the conditional distributions of X_t and Y_t, respectively.

A mean field variational approximation assumes that the joint distribution of the parameters and the observed data can be approximated by the product of individual densities. We wish to obtain an approximation of $f(\theta_t, X_t, Y_t)$ of the form $Q(\theta_t) = \prod_{j=1}^{6} Q_j(\theta_{tj})$ where $Q_j(.)$ is the pdf of a Gamma distribution and θ_{tj} is the jth component of θ_t. Gamma parameters are therefore sought which minimize the Kullback-Leibler divergence between $f(\theta_t, X_t, Y_t)$ and $Q(\theta_t)$. In general the optimal functions satisfy the relationship

$$\log Q_j(\theta_t j) = \mathbb{E}_{i \neq j}[\log f(\theta_t, X_t, Y_t)] + \text{const.} \tag{5.2}$$

where $\mathbb{E}_{i \neq j}$ denotes expectation with respect to all the components of θ_t except θ_{tj} (Bishop , 2006, p465–466). Finding the expectations in (5.2) typically requires an iterative algorithm because of the dependency of the jth function on the expectations of the other distributions $i \neq j$.

Since X_t and Y_t are independent conditional on $\theta_{t6} = \epsilon_t$, and that each component of θ_t has $\theta_{tj} \sim \mathrm{Gamma}(p_j, q_j)$ for $j = 1, \ldots, 6$, then

$$\begin{aligned}\log f(\theta_t, X_t, Y_t) =& X_t \sum_{k \in \mathcal{R}_1} \log \theta_{tk} - \prod_{k \in \mathcal{R}_1} \theta_{tk} + Y_t \sum_{k \in \mathcal{R}_2} \log \theta_{tj} - \prod_{k \in \mathcal{R}_2} \theta_{tk} \\ &+ \prod_{j=1}^{6} \{(p_j - 1)\log \theta_{tj} - q_j \theta_{tj}\} + \text{const.}\end{aligned}$$

and hence

$$\log Q_j(\theta_{tj}) = \begin{cases} \{X_t + (p_j - 1)\} \log \theta_{tj} \left\{ q_j + \prod_{k \in \mathcal{R}_1 \backslash j} \mathbb{E}(\theta_{tk}) \right\} \theta_{tj} + \text{const.} & j = 1, 4, 5 \\ \{Y_t + (p_j - 1)\} \log \theta_{tj} - \left\{ q_j + \prod_{k \in \mathcal{R}_2 \backslash j} \mathbb{E}(\theta_{tk}) \right\} \theta_{tj} + \text{const.} & j = 2, 3 \\ \{X_t + Y_t + (p_j - 1)\} \log \theta_{tj} & \\ \qquad - \left\{ q_j + \prod_{k \in \mathcal{R}_1 \backslash j} \mathbb{E}(\theta_{tk}) + \prod_{k \in \mathcal{R}_2 \backslash j} \mathbb{E}(\theta_{tk}) \right\} \theta_{tj} + \text{const} & j = 6. \end{cases}$$

Let $\bar{p}_j$ and $\bar{q}_j$ represent the updated Gamma parameters associated with θ_{tj} such that $Q_j(\theta_{tj})$ is the density of a Gamma$(\bar{p}_j, \bar{q}_j)$ random variable. Then the updated parameters satisfy $\bar{p}_j = p_j + I(j \in \mathcal{R}_1) X_t + I(j \in \mathcal{R}_2) Y_t$ and

$$\bar{q}_j = q_j + \begin{cases} \prod_{k \in \mathcal{R}_1 \backslash j} \bar{p}_k / \bar{q}_k & j = 1, 4, 5 \\ \prod_{k \in \mathcal{R}_2 \backslash j} \bar{p}_k / \bar{q}_k & j = 2, 3 \\ \prod_{k \in \mathcal{R}_1 \backslash j} \bar{p}_k / \bar{q}_k + \prod_{k \in \mathcal{R}_2 \backslash j} \bar{p}_k / \bar{q}_k, & j = 6 \end{cases}$$

where the $\bar{q}_j$ can be found via iteration.

Hence in comparison to (2.7), our proposed approximate updates for p_j coincide with the variational approximation, while for q_j our proposed updates can be thought of as representing a first iteration of the variational update where q_6 is updated using the prior expectations of the other θ_{tj} and then the others are updated using the updated expectation of ϵ_t and the prior expectations of the other θ_{tj}'s. Provided the amount of past data is relatively large in relation to the degree of forgetting, there will be little difference in the updates.

References

Arbous, A. G. and Kerrich, J. E. (1951). Accidental Statistics and the concept of Accident-Proneness. *Biometrics.* **7** (4). 414.

Baxter, M. and Stevenson, R. (1988). Discriminating between the Poisson and negative binomial distributions: An application to goal scoring in association football. *Journal of Applied Statistics.* **15**(3), 347–354.

Bradley, R. A. and Terry, M. E. (1952). The Method of Paired Comparisons. *Biometrika* **39** (3), 324-345.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Brier, G.W.(1950). Verification of Forecasts Expressed in Terms of Probability. *Mon. Wea. Rev.* **78** (1), 1-–3.

Cattelan, M., Varin, C. and Firth, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments *J R Stat Soc Ser C Appl Stat*, **62**, 135–15.

Creal, D., Koopman S.J. and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, **28** (5), 777–795.

Crowder, M, Dixon, M., Ledford, A. and Robinson, M. (2002). Dynamic modeling and prediction of English Football League matches for betting. *J. R. Stat. Soc. D Stat.* **51**, 157–168.

Davidson, R. R. (1970). On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *J Am Stat Assoc* **65** (329), 317-328.

Davison, A. C. (2003). Statistical modelling. *Cambridge University Press* 498–499.

Dixon, M. and Coles, S. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *J R Stat Soc Ser C Appl Stat*, **46**, 265–28.

Egidi, L. and Torelli, N. (2020). Comparing Goal-Based and Result-Based Approaches in Modelling Football *Social Indicators Research*, **156**, 801-813.

Gamerman. D., dos Santos, T.R. and Franco, C.F. (2013). A non-Gaussian family of state space models with exact marginal likelihood. *J. Time Ser. Anal.*, **34** , 625–645.

Gelman, A.(2013). Understanding posterior p-values. *Electron. J. Stat.* **4** 2595-2602.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *J Am Stat Assoc* **85**, 972–985.

Glickman, M. E. and Stern, H. L.(1998). A State-Space Model for National Football League Scores. *J Am Stat Assoc* **93**, 441. 25–35

Jordan, M. I., Ghahramani, Z. Jaakkola, T. and Saul, L. (1999). Introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.*. **82**: (35).

Karlis, M. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc. D Stat.*, **52**, 381–393.

Meinhold, R. J. & Singpurwalla, N. D. (1983). Understanding the Kalman Filter, *The American Statistician.* **37** (2), 123:127.

Koopman, S.J. and Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *J R Stat Soc Ser A Stat Soc.*, **178** (1), 167–186.

Koopman, S. J. and Lit, R. (2019). Forecasting football match results in national league competitions using score-driven time series models. *Int. J. Forecast.*, **35** (2), 979–809.

Maher, M. J.(1982). Modelling association football scores. *Stat Neerl*, **36**(3), 109–118.

Murphy, A. H.(1970). The Ranked Probability Score and the Probability Score: A Comparison. *Mon. Wea. Rev.*, **98** (12), 917-–924.

Rao, P. V., and Kupper, L. L. (1967). Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. *J Am Stat Assoc* **62** (317), 194–204.

Rue, H and Salvesen, O. (2000), Prediction and retrospective analysis of soccer matches in a league. *The Statistician* **49** (3), 399-418.

Ross, S. M.(2014). Introduction to Probability Models (11th ed.). Academic Press, Amsterdam, Netherlands.

Skellam, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *J R Stat Soc Ser A Stat Soc.* **109** (3), 26.

Titman, A. C., Costain, D. A., Ridall, P. G. and Gregory, K. (2015) Joint modelling of goals and bookings in association football. *J R Stat Soc Ser A Stat Soc.* **178** (3) 659-683.

Weigel, A. P., Liniger, M A., and Appenzeller, C. (2007) The Discrete Brier and Ranked Probability Skill Scores. *Mon. Wea. Rev.* **135** (1), 118—124.

West, M, Harrison, P. J. and Migon H.S. (1985). Dynamic Generalized linear models for Bayesian forecasting. *J Am Stat Assoc*, **80** (389), 67–86.

Figure Legends

Figure 1: Directed acyclic graphs (DAGs) showing the dependence structure of the variables in our football model. The top panel is a DAG showing the model before the update step and the lower panel shows the posterior distribution after the update step, Equations (2.7), have been carried out..

Figure 2: Dependence structure of a SSM. The transition equation for the parameter or state is denoted by $\theta_t \sim \pi(\cdot \mid \theta_{t-1}, y_{1:t-1})$, and the downward vertical arrows show the conditional sampling distribution of the observations, or the measurement equation $f(y_t \mid \theta_t)$. These two components can be used in a Bayesian filter to recursively update the evolving posterior distribution of states, $\theta_t \sim \pi(\cdot \mid y_{1:t})$.

Figure 3: A comparison of the cumulative RPS measure of the prediction accuracy (relative to those obtained using the the odds offered by the bookmakers) of three dynamic models, a score driven model and six others that use a weighted likelihood (where the predictions are recalculated for every round for every season apart from the first 5 rounds.) The training set were the 1992/93 to 2009/10 seasons and the test set 2010/11 to 2023/24.

Figure 4: The diagram shows the improvement in standardised residuals of the sequential bivariate model relative to those of the sequential univariate model. The dots represent the univariate residuals calculated using Equation (3.10) and the tips of the arrows show the bivariate residuals calculated using Equations (3.12). The numbers (next to the point) shows the ordering of the size of the residuals calculated using Equation (3.13).

Figure 5: Panels A, C and E show the problems of overdispersion in the univariate model. Panel F shows the correlation between the univariate residuals. Panels B and D show that the problem of overdispersion has been reduced in the bivariate model. The correlation between the residuals of the bivariate model is unchanged from Panel F. The horizontal dashed lines in panels A, B and F are averages taken over all seasons. The horizontal line in Panel E is the level above which the Chi-squared statistic for extra-Poisson dispersion is significant at the 0.05 level.

Figure 6: The top panel shows data-based estimate of the HGA (top panel) by calculating the difference between the home and away goal average for each season. The middle panel show the 75 % quantiles of the filtered model-based home ground advantage. The last panel shows a smoothed 75% credible interval to represent the posterior distribution of the HGA using the method of Gamerman et al. (2013) explained in Section 3.2.4.

Figure 7: 75% credible intervals showing the the filtered attacking strengths (solid line) and defensive strengths(dashed lines) of all teams in the 2023-4 season of the EPL.

Figure 8: 75% credible intervals showing the the filtered attacking strengths (Solid line) and the defensive strengths(Dashed lines) of the top clubs over the last 22 years of the EPL.