International Journal of Forecasting 35 (2019) 797–809

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Forecasting football match results in national league competitions using score-driven time series models

Siem Jan Koopman [a,b,*], Rutger Lit [a]

[a] Vrije Universiteit Amsterdam and Tinbergen Institute, The Netherlands
[b] CREATES, Aarhus University, Denmark

ARTICLE INFO

Keywords:
Bivariate Poisson
Ordered probit
Skellam
Probabilistic loss function

ABSTRACT

We develop a new dynamic multivariate model for the analysis and forecasting of football match results in national league competitions. The proposed dynamic model is based on the score of the predictive observation mass function for a high-dimensional panel of weekly match results. Our main interest is in forecasting whether the match result is a win, a loss or a draw for each team. The dynamic model for delivering such forecasts can be based on three different dependent variables: the pairwise count of the number of goals, the difference between the numbers of goals, or the category of the match result (win, loss, draw). The different dependent variables require different distributional assumptions. Furthermore, different dynamic model specifications can be considered for generating the forecasts. We investigate empirically which dependent variable and which dynamic model specification yield the best forecasting results. We validate the precision of the resulting forecasts and the success of the forecasts in a betting simulation in an extensive forecasting study for match results from six large European football competitions. Finally, we conclude that the dynamic model for pairwise counts delivers the most precise forecasts while the dynamic model for the difference between counts is most successful for betting, but that both outperform benchmark and other competing models.

1. Introduction

The forecasting of football match results is a highly popular activity. Amongst football supporters, it is common to forecast the next match result, and often the individual forecasts show a positive bias towards the team that one is supporting. However, even the pundit knows that forecasting a match result is a challenging task. The common way of betting on a football match is simply to indicate whether one expects the team to win, lose, or draw its next game. Whether the match result is a win, loss or draw depends on the difference between the numbers of goals scored by the two opposing teams in a football match. There are many factors that determine whether a goal is scored, including the attack strength of the team, the defence strength of the opposing team, the home ground advantage (when applicable), and specific events that take place during the match. We consider the use of three possible observational variables on which to base our forecast of the next match result in terms of win, loss, or draw. The first variable is two-dimensional and consists of the numbers of goals scored by the two opposing teams during a match. The second variable is the difference between the numbers of goals scored. The third variable is simply an indicator of win, loss, or draw. The informational content of these three consecutive variables is clearly decreasing. For each of the variable categories, a variety of dynamic models can be considered for the forecasting of the match result. Many contributions in the statistical literature on the modelling and forecasting of the three variables have been made. We refer to Table E.1 in the Appendix for a schematic overview

* Correspondence to: Vrije Universiteit Amsterdam School of Business and Economics, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
E-mail address: s.j.koopman@vu.nl (S.J. Koopman).

https://doi.org/10.1016/j.ijforecast.2018.10.011
0169-2070/

of the main contributions. This literature also extends to other fields such as econometrics and machine learning; two recent examples are the studies by Angelini and De Angelis (2017) and Baboota and Kaur (2018). A discussion of the statistical literature on forecasting football matches, together with their relevance to our study, follows.

Most of the contributions to the statistical literature on the modelling and forecasting of match results focus on the first variable, where the pairwise observations of the numbers of goals scored by the opposing teams are assumed to come from a bivariate distribution. The probability of a given possible match outcome is implied by the bivariate distribution, and is given formally by $P(X = x, Y = y)$, for $x, y \in \mathbb{N}_0$, where X and Y denote the numbers of goals scored by the home and away teams, respectively. Hence, we have a probability for any match outcome. The main interest usually focuses on forecasting the probabilities of a home win, draw, or away win; these are the *toto probabilities* and are given by $P(X > Y)$, $P(X = Y)$, and $P(X < Y)$, respectively. The parameters of the distribution can be expressed as a function of the competing teams' strengths of attack and defence. This procedure was first proposed by Maher (1982), who expresses the means of the double-Poisson distribution (product of two independent Poissons) as team-specific strengths of attack and defence. Dixon and Coles (1997) consider the double-Poisson distribution as well, and introduce a dependence parameter for the match results 0–0, 1–0, 0–1 and 1–1. They also propose a weighting function to down-weight the likelihood contributions of observations from the more distant past. Crowder, Dixon, Ledford, and Robinson (2002) formulate the model of Dixon and Coles (1997) as a non-Gaussian state space model with time-varying strengths of attack and defence, then develop approximating methods for parameter estimation and signal extraction, stating that an exact analysis is too expensive computationally. A bivariate Poisson distribution is also used by Karlis and Ntzoufras (2003), who show that the introduction of a parameter for dependence between the goals scored by the two teams during a match leads to a more accurate prediction of the outcome of a draw. Rue and Salvesen (2000) incorporate the framework of Dixon and Coles (1997) and develop a dynamic generalized linear model which is analysed in continuous time by Markov chain Monte Carlo methods. Owen (2011) adopts advanced Bayesian methods for the forecasting of football match outcomes. Goddard (2005) explores the inclusion of covariates in a bivariate Poisson model. Koopman and Lit (2014) show that a high-dimensional panel of weekly match results can be analysed effectively within a non-Gaussian state space framework based on the bivariate Poisson model with stochastically time-varying attack and defence strengths, and with some of the above extensions. Finally, there are many other interesting and original contributions that could be listed in this category were space not an issue. For example, Dixon and Robinson (1998) treat the numbers of goals scored by the competing teams during a match as interacting birth processes, while Boshnakov, Kharrat, and McHale (2017) deal with them within a bivariate distribution based on Weibull inter-arrival-times count processes and a copula function for their dependence.

The second category is the difference between numbers of goals in a match, and can be regarded as a team's margin of victory. In this category we let $Z = X - Y$ be the difference between the numbers of scored goals X and Y, with $Z \in \mathbb{Z}$. By modelling Z, we consider the toto probabilities as given by $P(Z > 0)$, $P(Z = 0)$, and $P(Z < 0)$ for a home win, draw, and away win, respectively. Modelling the difference in goals instead means that information is lost, since, for example, the pairs $(X = 0, Y = 1)$ and $(X = 2, Y = 3)$ produce the same values for Z. On the other hand, smaller numbers of summations are needed in order to obtain toto probabilities from Z than from the pair (X, Y). It is not clear immediately what the overall effect of modelling Z instead of (X, Y) would be on the forecasting of the toto probabilities. The reasoning behind this relates to the accumulation of modelling error, which could potentially be smaller, since a smaller number of probability components are being summed compared to the first category. A model for the differences between numbers of goals in football matches is provided by Karlis and Ntzoufras (2009), who introduce the Skellam distribution for analysing match results. This distribution was originally derived by Skellam (1946) as the difference of two independent Poisson distributions. However, Karlis and Ntzoufras (2009) show that independence is not strictly necessary, and even the Poisson assumption for the pair of variables (X, Y) is not needed. Their analysis keeps the parameters of the Skellam distribution static. Lit (2016, Ch. 4) extends the Skellam model to allow for strengths of attack and defence that evolve stochastically over time in a non-Gaussian state space framework.

Rather than modelling toto probabilities via the double or bivariate Poisson models or via the Skellam models, we can consider the modelling of the toto probabilities directly. For this third variable category, this study introduces ordered logit or ordered probit models. The modelling of match results in terms of wins, losses, and draws, rather than scores or differences in scores, leads not only to a more parsimonious model, but also to a simpler estimation procedure. Koning (2000) investigates the balance in competition in Dutch professional soccer by means of an ordered probit model with static team strengths. A selection of covariates can be introduced in both the static ordered probit regression model of Goddard and Asimakopoulos (2004) and the static ordered logit model of Forrest and Simmons (2000). Cattelan, Varin, and Firth (2013) propose a (semi-)dynamic Bradley–Terry model in which team strengths are modelled by exponentially weighted moving average processes. An early contribution is made by Fahrmeir and Tutz (1994), who introduce an ordered logit non-Gaussian state space model that incorporates random walks for the team strengths. The estimation of the parameters of this model is carried out by the Kalman filter and recursive posterior mode estimation methods. The dynamic cumulative link model of Knorr-Held (2000) has been applied to German Bundesliga data for an analysis based on the extended Kalman filter and smoother. Held and Vollnhals (2005) adopt the model of Knorr-Held (2000) for evaluating the comparative strengths of football teams in five European competitions combined in a single data set. The competitions are linked through the inclusion of

European competitions (Champions League, Winners Cup and UEFA Cup). Their study concentrates on the comparison of team strengths within Europe. Forecasting evidence for a range of models, in-sample or out-of-sample, is not presented. Finally, Hvattum and Arntzen (2010) propose an ordered logit model in which team strengths are updated over time using Elo ratings.

Our research contributes to the literature in a number of ways. First, we develop a new dynamic multivariate model for the analysis and forecasting of football match results for each of the three variable categories. The dynamic extensions of the static models are based on the class of score-driven models in which the time-varying coefficients are updated as an autoregressive process. The autoregressive updating of the time-varying parameter is driven by the score of the conditional observation probability density function; see Creal, Koopman, and Lucas (2013) for a discussion of this approach. Three features of this class of models are particularly attractive in our context:

(i) The score-driven models are observation-driven, which means that the likelihood is available in closed form. This allows for a fast estimation process despite the challenges involved with the use of high-dimensional models due to the large numbers of teams that participate in European football competitions over a number of years. The Kalman filter, which is more demanding computationally, is not required for estimation and forecasting.
(ii) The filtered estimates of the time-varying parameters in the score-driven models are locally optimal in a Kullback–Leibler sense; see Blasques, Koopman, and Lucas (2015).
(iii) The forecasting performances of the score-driven models are comparable to those of their parameter-driven counterparts; see Koopman, Lucas, and Scharth (2016).

Second, we conduct an extensive empirical study to determine which of the three variable categories leads to the most accurate forecasts. Third, as part of our empirical study we also investigate whether dynamic models with time-varying parameters show better forecasting performances relative to models with static parameters. We further verify whether the dynamic extension of the static model is achieved best by formulating a time-varying parameter model or by weighting the likelihood contributions over time, as proposed by Dixon and Coles (1997).

We have constructed time series panels of match results from six European competitions: the English Premier League, the German Bundesliga, the Spanish Primera División, the French Ligue 1, the Italian Serie A, and the Dutch Eredivisie. We have collected 17 seasons of match results, from 1999–2000 to 2015–2016, of which the first ten seasons are used for parameter estimation and the last seven are used for the forecasting study. Our out-of-sample forecasting study can be regarded as large, especially relative to other related forecasting studies in the literature. Hence, it is a considerable achievement in our forecasting study when a model turns out to be the best performer. We use the rank probability score as a loss function, and explain why this is the most suitable loss function for this exercise. The losses are evaluated using the Diebold and Mariano (1995) statistic to test for equal predictive accuracy. Furthermore, we determine whether any of the proposed models are capable of turning the forecasts into profits in a betting simulation.

The remainder of this forecasting study is organized as follows. Section 2 introduces the general statistical modelling framework, while Section 3 discusses the specific details of the score-driven football models. Section 4 provides the design of our extended forecasting study, including a data description, presents our empirical findings and discusses various aspects of our analyses. Section 5 concludes. An Online Appendix provides additional figures, tables, and many technical details related to estimation, score functions and toto probabilities for the various models that we consider in our study.

2. The distributions for the three variable categories

We develop three corresponding modelling frameworks for the three variable categories, then consider the different observational characteristics and propose their corresponding discrete mass functions. The details of maximum likelihood estimation are given in Appendix A. The dynamic extensions are developed and discussed in Section 3.

2.1. Bivariate Poisson distribution

The outcome of a football match is determined simply by the numbers of goals scored and conceded by a team. The outcome can be considered as a pair of counts (X, Y), where X is the number of goals scored by the home team and Y the number scored by the away team. We assume here that the pair of counts (X, Y) is generated by a bivariate Poisson distribution with intensities $\lambda_1, \lambda_2 > 0$ for (X, Y) and with the covariance between (X, Y) being denoted by $\lambda_3 \geq 0$. The probability mass function of the bivariate Poisson distribution is given by

$$p_{\text{BP}}(X = x, Y = y; \lambda_1, \lambda_2, \lambda_3) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{k=0}^{min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k, \tag{1}$$

see Johnson, Kotz, and Balakrishnan (1997) and Kocherlakota and Kocherlakota (1992) for more information. It can be shown that

$$\begin{aligned} &\mathbb{E}_{\text{BP}}(X) = \mathbb{V}\text{ar}_{\text{BP}}(X) = \lambda_1 + \lambda_3, \\ &\mathbb{E}_{\text{BP}}(Y) = \mathbb{V}\text{ar}_{\text{BP}}(Y) = \lambda_2 + \lambda_3, \quad \mathbb{C}\text{ov}_{\text{BP}}(X, Y) = \lambda_3, \end{aligned} \tag{2}$$

where $\mathbb{E}_p$, $\mathbb{V}\text{ar}_p$ and $\mathbb{C}\text{ov}_p$ denote the expectation, variance and covariance, respectively, with respect to density p. For $\lambda_3 = 0$, the bivariate Poisson distribution reduces to the double Poisson distribution. The covariance is a "shared component" in the intensities: a higher λ_3 leads to a higher number of equal observations $(X = Y)$, which in football are draws.

In the context of modelling match results in football, we follow the framework developed by Maher (1982), which has become the standard in the statistics literature on sports modelling. We therefore specify the intensities λ_1

and λ_2 as functions of the latent strengths of attack α and defence β of the two opposing teams, and the home ground advantage effect δ. Suppose that home team i welcomes away team j for a football match. Then, the intensity $\lambda_{1,ij}$, associated with the number of home goals X in this match of team i versus team j, and the intensity $\lambda_{2,ij}$, associated with the corresponding number of away goals Y, can be specified as

$$\lambda_{1,ij} = \exp(\delta + \alpha_i - \beta_j), \qquad \lambda_{2,ij} = \exp(\alpha_j - \beta_i), \tag{3}$$

with α_m and β_m being the attack and defence strengths, respectively, of team $m = i, j$ and $i \neq j$. The home ground advantage δ can also be made team-specific, but we restrict this effect to be equal for all teams.

2.2. Skellam distribution

The win, loss or draw of a football match is determined by the difference between the numbers of goals scored and conceded by a team. The difference between the numbers of home goals X and away goals Y can be regarded as a team's margin of victory. We can assume that this difference in the counts $Z = X - Y$ is distributed according to the Skellam distribution with intensities λ_4, λ_5. The probability mass function of the Skellam distribution is given by

$$p_{\text{Sk}}(Z = z; \lambda_4, \lambda_5) = e^{-(\lambda_4+\lambda_5)} (\lambda_4/\lambda_5)^{z/2} I_{|z|}(2\sqrt{\lambda_4\lambda_5}), \tag{4}$$

where $I_{|z|}(\cdot)$ is the modified Bessel function of order $|z|$. The mean and variance of Z are given by

$$\mathbb{E}_{\text{Sk}}(Z) = \lambda_4 - \lambda_5, \qquad \mathbb{V}\text{ar}_{\text{Sk}}(Z) = \lambda_4 + \lambda_5. \tag{5}$$

We refer to the original work of Irwin (1937) and Skellam (1946) for the derivation of the Skellam distribution based on the differences between two independent Poisson distributions. Alzaid and Omair (2010) presented higher moments and several other interesting properties of the Skellam distribution. Karlis and Ntzoufras (2009) showed that the underlying Poisson assumption is not strictly necessary and that the Skellam distribution can also be considered by itself as a distribution defined on integers.

When modelling football match results in terms of their margins of victory, we can also incorporate the framework of Maher (1982) for the Skellam distribution. The intensity λ_4 is associated with the number of home goals and the intensity λ_5 with the number of away goals. Hence, the specifications for $\lambda_{1,ij}$ and $\lambda_{2,ij}$ in Eq. (3) can apply similarly to $\lambda_{4,ij}$ and $\lambda_{5,ij}$, respectively, for home team i and away team j.

2.3. Ordered probit models

The outcome (win, loss or draw) of a football match can also be considered as an observed variable that we then model directly. In this case, the observed categorical variable C is determined simply by $C = 2$ for a home win $X > Y$ (or $Z > 0$), $C = 1$ for a draw $X = Y$ (or $Z = 0$), and $C = 0$ for a home loss $X < Y$ (or $Z < 0$). The margin of victory is not measured. The variable C can also be interpreted as the credit points for a win, draw or loss of a match, although the credit for a win is 3 points rather than 2 in all of the football competitions that we consider. In an ordered probit model, we assume that an unobserved stochastic variable C^* determines the category C probabilistically, with C^* being given by

$$C^* = \lambda_6 + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2), \tag{6}$$

where λ_6 is an unknown constant that indicates the strength of the home team relative to that of the away team, and where η is a random variable that is generated by a normal variable with mean zero and variance σ_η^2. For the ordered probit model, we assume that the variable C is generated conditional on C^* via the equations

$$C = \begin{cases} 2 & \text{if} \quad C^* \leq \kappa_1, \\ 1 & \text{if} \quad \kappa_1 < C^* \leq \kappa_2, \\ 0 & \text{if} \quad C^* > \kappa_2, \end{cases} \tag{7}$$

where the cutoff points κ_1 and κ_2 and the variance σ_η^2 are treated as unknown parameters; see for example Greene (2012) for a textbook treatment of ordered probit models, as well as for a more general treatment with more categories. Given the construction with the unobserved variable C^* and the random variable η, we cannot jointly identify the three parameters uniquely. As a result, we constrain the scale of η and set $\sigma_\eta^2 = 1$. An alternative is to set κ_1 to zero and choose σ_η^2 freely, together with κ_2. For an ordered probit model with categorical observations $C \in \{2, 1, 0\}$ and $\sigma_\eta^2 = 1$, the probability density function is given by

$$p_{\text{OP}}(C \in \{2, 1, 0\}; \lambda_6, \kappa_1, \kappa_2) = \begin{cases} \Phi(\kappa_1 - \lambda_6) & \text{if} \quad C = 2, \\ \Phi(\kappa_2 - \lambda_6) - \Phi(\kappa_1 - \lambda_6) & \text{if} \quad C = 1, \\ 1 - \Phi(\kappa_2 - \lambda_6) & \text{if} \quad C = 0, \end{cases} \tag{8}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function (cdf). We ensure that the probabilities are all positive by further restricting the parameters by $\kappa_1 < \kappa_2$.

Given the limited amount of information in the category variable of a win, loss or draw of a football match, the framework of Maher (1982) cannot be incorporated in an ordered probit model. It is also the design of the ordered probit model that does not allow the separation of the strength of a team into attack and defence strengths. Since the relative strength of the home team is represented by λ_6, and since it determines the probability of category $C \in \{2, 1, 0\}$, we have

$$\lambda_{6,ij} = \gamma_i - \gamma_j, \tag{9}$$

where γ_m is the total strength or capability of team m.

3. Score-driven time-varying parameters

The treatment of football match results using the three observational variables discussed in Section 2 is relatively straightforward. This is due partly to the assumptions that matches and teams' efforts in each round of the competition are treated as independent events. However, it is not realistic to treat the numbers of goals scored by a team in a series of matches as independent events: a football team's strength is likely to be related in part to its performance in recent matches. Teams' attack and defence strengths

also change over time as their compositions evolve through the years. Several dynamic extensions of the static models described above are considered in the statistics and econometrics literature, and Appendix C considers a selection of such existing methods for comparisons.

However, we contribute by developing an effective and computationally fast approach to the dynamic modelling of football teams' attack and defence strengths. These developments are presented in Section 3.1, which provides a short review of score-driven time series models, and 3.2, which presents the details of its implementation for the models of Section 2. The initialization of the dynamic processes are discussed in Appendix D. Section 4 presents the empirical results from European football league competitions, which include excellent forecasting results for our proposed dynamic extensions.

3.1. Score-driven time series models: a short review

We consider the class of score-driven models by Creal et al. (2013) for capturing the dynamic behavior of a parameter or a selection of parameters. We address the case of a panel of time series variables for which y_t represents the corresponding vector of observations at time t. We assume that the data are generated from a distribution with density function $p(y_t; \psi)$ for which the density functions discussed in Section 2 are examples. The observation vector y_t can include the pairs of counts for the numbers of goals of each team, the differences between the numbers of goals in a match, or the indicator of a win, loss or draw. One part of the static model parameters in ψ is assumed to be time-varying and is collected in the time-varying parameter vector f_t. The remaining static parameters are collected in the parameter vector ψ^*.

In this framework, the score-driven model is based on the predictive density function that is treated as the observation density and is given by

$$y_t \sim p\left(y_t|f_t, \mathcal{F}_t; \psi^*\right), \qquad t = 1, \ldots, T, \tag{10}$$

where $\mathcal{F}_t$ represents the information set available at time t, consisting of lagged observations $\{y_{t-1}, y_{t-2}, \ldots\}$ and past time-varying parameter vectors $\{f_{t-1}, f_{t-2}, \ldots\}$. The score-driven updating mechanism for the time-varying parameter f_t is given by

$$f_{t+1} = \omega + Bf_t + As_t, \tag{11}$$

where ω is a vector of unknown constants, the matrices A and B are unknown coefficient matrices, and s_t is the scaled score vector, defined by

$$\begin{aligned} s_t &= S_t \cdot \nabla_t, \qquad \nabla_t = \frac{\partial \log p(y_t|f_t, \mathcal{F}_t; \psi^*)}{\partial f_t}, \\ S_t &= S(f_t, \mathcal{F}_t; \psi^*), \end{aligned} \tag{12}$$

with $S(\cdot)$ being a matrix function for scaling the score vector. A score-driven model updates the factor f_{t+1} in the direction of the steepest increase of the log-density at time t given the current parameter f_t and the data history $\mathcal{F}_t$. The initialisation of Eq. (11) is discussed in Appendix D. Under correct model specification, the score vectors are a martingale sequence, since $\mathbb{E}_{t-1}(s_t) = 0$, where $\mathbb{E}_{t-1}$ denotes the expectation with respect to $p(y_t|f_t, \mathcal{F}_t; \psi^*)$. We note that the predictive density $p(y_t|f_t, \mathcal{F}_t; \psi^*)$ is conditional on the time-varying parameter at time t and has the same functional form as $p_{\mathcal{M}}(\cdot; \psi_{\mathcal{M}})$, with $\mathcal{M} \in \{\text{BP, Sk, OP}\}$.

The scaling matrix is chosen regularly to be a function of the variance of the score, so as to take into account the curvature of the log-density at time t, as summarized by the Fisher information matrix $\mathcal{I}_{t|t-1} = \mathbb{E}_{t-1}\left[\nabla_t \nabla_t'\right]$. When it is intricate or impossible to obtain the Fisher matrix analytically, we can take S_t to be the unity matrix. The score-driven updating of parameters has a theoretical foundation, since the estimates of the time-varying parameter are optimal in a Kullback–Leibler sense, see Blasques et al. (2015), and therefore it is not a heuristic method. The details of maximum likelihood estimation are given in Appendix A.

3.2. Score-driven models for football match results

Next, we adopt the score-driven time-varying parameter framework for the three densities discussed in Section 2. We consider the time-variation for a selection of parameters in ψ and provide details of its implementation for the modelling of football match results. The relevant score functions for the three densities are given in Appendix B. We obtain a flexible and effective framework for the time series analysis and forecasting of football match results in large competitions and over many seasons of a competition.

3.2.1. Bivariate Poisson distribution

The modelling of football match results via the observation pair (x_{ijt}, y_{ijt}), where x_{ijt} and y_{ijt} are the numbers of goals made by the home team i and the away team j, respectively, in round t, can be based on the bivariate Poisson distribution $p_{\text{BP}}(X = x_{ijt}, Y = y_{ijt}; \lambda_{1,ij}, \lambda_{2,ij}, \lambda_3)$ and using the approach of Maher (1982), as reflected in Eq. (3). The dynamic model lets us allow the strengths of the teams in attack and defence to be time-varying. In particular, we replace α_i and β_i in Eq. (3) with α_{it} and β_{it}, respectively. We then obtain the $2N \times 1$ time-varying parameter f_t which contains α_{it} and β_{it} for all N teams that are active in a competition, that is

$$f_t = (\alpha_{1t}, \ldots, \alpha_{Nt}, \beta_{1t}, \ldots, \beta_{Nt})', \qquad t = 1, \ldots, T. \tag{13}$$

The home ground advantage δ and the covariance λ_3 in ψ_{BP} can remain constant over time: they can be treated as static parameters and are placed in ψ^* of the score-driven model. The implication of this dynamic extension is that the intensities $\lambda_{k,ij}$ can now be treated as time-varying intensities that we denote by $\lambda_{k,ijt} = \lambda_{k,ij}(f_t)$, for $k = 1,2$, where $\lambda_{k,ij}(\cdot)$ refers to the functions in Eq. (3).

The time-varying updating equation for f_t is provided by Eq. (11). However, it is more efficient to carry out the updating at round t for each match result. We assume that the observation pair (x_{it}, y_{jt}) is generated by the bivariate Poisson and select a subset of f_t that is relevant for this match, that is

$$f_{ijt} = \left(\alpha_{it}, \alpha_{jt}, \beta_{it}, \beta_{jt}\right)' = M_{ij} f_t, \tag{14}$$

where M_{ij} is the $4 \times 2N$ selection matrix of zeros and ones, and is defined implicitly. We update the selected time-varying parameters as in Eq. (11). It reduces to the updating

$$f_{ij,t+1} = \omega_{ij} + B_{ij} f_{ijt} + A_{ij} s_{ijt}, \tag{15}$$

with a 4×1 vector of constants $\omega_{ij} = M_{ij}\omega$, 4×4 coefficient matrices $A_{ij} = M_{ij} A M_{ij}'$ and $B_{ij} = M_{ij} B M_{ij}'$, and the 4×1 scaled score vector s_{ijt} which is defined as in Eq. (12) but with the gradient with respect to the 4×1 vector f_{ijt}. This updating is then repeated for all $N/2$ matches in each round t, and hence the full vector f_t is updated effectively. From the update f_{t+1}, we can then make predictions for the match results for the next round $t+1$ of the competition. In particular, we can forecast the probabilities of a win, draw, and loss as described at the end of Section 2.1.

The dimension of vector f_t is as high as $2N$. We may therefore want to specify the $2N \times 2N$ coefficient matrices A and B in a parsimonious manner. In our empirical study, we specify them as

$$A = \begin{bmatrix} a_1 \cdot I_N & 0 \\ 0 & a_2 \cdot I_N \end{bmatrix}, \qquad B = \begin{bmatrix} b_1 \cdot I_N & 0 \\ 0 & b_2 \cdot I_N \end{bmatrix},$$

such that the attack strengths (α) rely on the updating coefficients a_1 and b_1 and the defence strengths (β) rely on the updating coefficients a_2 and b_2. The coefficient matrices A_{ij} and B_{ij} are then defined similarly, but for $N = 2$. In this specification, the coefficient matrices A and B do not have team-specific characteristics and there are no spillover effects in the dynamic specifications between the teams. Such specifications can be considered in a straightforward manner because all unknown elements of A and B are placed in ψ^*.

The specification of the scaled score $s_{ijt} = S_{ijt} \cdot \nabla_{ijt}$ in Eq. (15) for the bivariate Poisson is provided in Appendix B.2, at least for the derivation of ∇_{ijt}. Since the derivation of the Fisher matrix is intricate, we set the scaling to the unity matrix, that is $S_{ijt} = I_4$. The parameter vector ψ^* is given by

$$\psi^* = (a_1, a_2, b_1, b_2, \lambda_3, \delta)', \tag{16}$$

and is estimated by maximum likelihood. The estimation of ω_{ij} in Eq. (14) is discussed in Appendix D.

A team plays only once in each round t. Hence, the maximum number of matches at time t is $N/2$. When every team plays according to schedule, a season consists of $T = 2(N-1)$ rounds in a competition. In practice, however, some football matches are postponed due to bad weather conditions and other external events, and these matches are then played later at some convenient time, with additional rounds being inserted in the calendar of the competition for this purpose. A small set of matches are scheduled for such additional rounds. If a team does not play in round t, its score is set to zero and the updating for the strengths of attack and defence reduces to

$$\alpha_{m,t+1} = \omega_m + b_1 \alpha_{mt} \quad \text{and} \quad \beta_{m,t+1} = \omega_m + b_2 \beta_{mt},$$

respectively, for any team m that does not play in round t.

3.2.2. Skellam distribution

When the football match result is modelled in terms of the observed margin of victory, $z_{ijt} = x_{ijt} - y_{ijt}$, we can opt for the Skellam distribution in Section 2.2 with a similar dynamic extension as for the bivariate Poisson distribution. The probability mass function of the Skellam distribution is given by Eq. (4); see also Koopman, Lit, and Lucas (2017) for a discussion of its implementation in a dynamic setting. The time-varying parameter f_t has the same composition as in Eq. (13), and hence we consider the Skellam pmf $p_{\text{Sk}}(Z = z_{ijt}; \lambda_{4,ijt}, \lambda_{5,ijt})$, where $\lambda_{k,ijt} = \lambda_{k,ij}(f_t)$, for $k = 4,5$, where $\lambda_{k,ij}(\cdot)$ refers to the corresponding functions in Eq. (3). The updating of f_t can take place for each match separately using Eq. (15), where f_{ijt} is defined in Eq. (14). The derivation of the score s_{ijt} for the Skellam density is presented in Appendix B.3. We set the scaling equal to the unity matrix, that is, $S_{ijt} = I_4$. The parameter vector ψ^* is given by $\psi^* = (a_1, a_2, b_1, b_2, \delta)'$, and is estimated by maximum likelihood.

3.2.3. Ordered probit

When we record the data simply as the win, loss or draw of a given football match (that is, we observe $c_{ijt} \in \{2, 1, 0\}$), we can model the data using the ordered logit model of Section 2.3 with a dynamic extension for the overall strength (or capability) of the team. Given that we cannot separate the strength in defence and attack, we have a more parsimonious model. The strengths are made time-varying by replacing γ_i with γ_{it} and placing them in the $N \times 1$ vector f_t. We have

$$f_t = (\gamma_{1t}, \ldots, \gamma_{Nt})', \qquad t = 1, \ldots, T.$$

As a result, the coefficient $\lambda_{6,ij}$ has also become time-varying, where we define $\lambda_{6,ijt} = \gamma_{it} - \gamma_{jt}$ and this indicates the difference in strength between the home team i and the away team j for their football match in round t. This is our dynamic version of the ordered logit model for football match results of Koning (2000). The cutoff points κ_1 and κ_2 (which also represent the home ground advantage) remain static coefficients.

The updating of the time-varying team capabilities f_t can also be done for each match result separately, as is implied by Eq. (15), but now f_{ijt} is simply the 2×1 vector $(\gamma_{it}, \gamma_{jt})'$, with the 2×1 constant vector $\omega_{ij} = (\omega_i, \omega_j)'$ and the 2×2 coefficient matrices $A_{ij} = a_1 \cdot I_2$ and $B_{ij} = b_1 \cdot I_2$. The derivation of the 2×1 score vector s_{ijt} for the ordered probit pmf is provided in Appendix B.4. We set the scaling equal to the unity matrix, that is, $S_{ijt} = I_2$. The parameter vector is given by $\psi^* = \{a_1, b_1, \kappa_1, \kappa_2\}$, and is estimated by maximum likelihood.

4. Forecasting football match results in Europe

Our empirical study is a basic and straightforward exercise: we forecast all match results in the next round of a football competition, for all rounds in seven yearly competitions and for six European football competitions. Thus, this design of our study involves making almost 15,000 probabilistic forecasts for the football toto results, which are the match results in terms of wins, losses and draws for the home team. These probability forecasts are based on a

particular model. We consider all three static models from Section 2 (using the three different variable categories) and their various dynamic extensions as discussed in Section 3. This forecasting study is of an exceptional magnitude, and hence allows us to draw strong conclusions concerning which model performs best. The forecast precision measurements are based both on the average rank probability score statistics and on losses in a particular betting simulation.

4.1. Data description

We forecast the football toto results for six European football competitions: The English Premier League, German Bundesliga, Spanish Primera División, Italian Serie A, France Ligue 1, and the Dutch Eredivisie. For each competition, the total data set consist of 17 seasons of football match results. We have partitioned the data set into the in-sample seasons 1999–2009, which are used for initial parameter estimation, and the out-of-sample seasons 2009–2016, which are used for our forecasting study. After each football season, the poorest performing team(s) are relegated and new teams are promoted into the competition. Hence, the total number of teams in the data set increases with every season, since the relegated teams remain in the panel because they can re-appear in future seasons. The number of teams relegated varies between competitions and across seasons. We refer to Table 1 for some descriptive statistics of the six football competitions. The data used in our empirical study can be found at http://www.football-data.co.uk.

4.2. Estimated strengths from the score-driven model

We illustrate our proposed score-driven model empirically as a dynamic extension of our models for football match results by presenting in Fig. 1 the time-varying estimates of the attack, defence and total strengths of the two major rival teams in the Spanish Primera división: Barcelona and Real Madrid. We present the estimated strengths for the dynamic bivariate Poisson model; the implementation details for filtering and parameter estimation are discussed in Section 3.2. The graphs in Fig. 1 are all based on the values of f_t as defined in Eq. (13) and evaluated recursively by Eq. (11), where ω, A and B are replaced by their maximum likelihood estimates, which are provided in Table 2. These estimates reveal that the persistence levels for the dynamic models are very high, with most estimates of b_1 and b_2 being close to unity. However, we notice that the persistence is for a weekly frequency (approximately). For example, an estimate of 0.998 for b_1 implies a quarterly persistence of $0.998^{13} = 0.97$ and a yearly persistence of $0.998^{52} = 0.90$. Given how football teams evolve from one season to another, a yearly persistence of 0.90 is rather realistic.

The estimated toto probabilities are computed as in Eq. (A.1), but with the underlying parameters replaced by their estimates. Since the strengths are time-varying (α and β are in f_t), these probability computations are done repeatedly, before each new football round starts.

The estimated strengths of attack and defence for Barcelona and Real Madrid, as obtained from the dynamic bivariate Poisson model, reveal that the attack strengths of both teams have been competitive and increasing steadily in the 16 seasons from 2000 onwards, while the defence strength of Barcelona has been stronger overall since 2004, becoming even stronger in the more recent years. The superior strength of Barcelona over Real Madrid overall since 2004 has been small, but nevertheless clearly visible. This conclusion is also supported by the relative probabilities of a Barcelona win and a Real Madrid win. Considering now the period since 2008 rather than 2004, we find here that the probability of a Barcelona win is persistently close to 0.5, while the Real Madrid win probability is closer to 0.35 during the last seven seasons of the sample. We note that the strengths and probabilities are displayed for each match in the sample, and hence the strengths of the two teams are not presented exclusively for the Barcelona versus Real Madrid (and vice versa) matches. We focus more on those two matches in each season by indicating in each plot when these key matches took place and whether each was a win, a draw or a loss for Barcelona. During the 16 seasons in our sample, the rivals played against each other 32 times, resulting in 14 wins for Barcelona, 8 draws and 10 wins for Real Madrid. The home ground advantage effect is not accounted for in these plots in order to obtain more precise comparisons.

Graphs similar to that in Fig. 1 can be presented for the dynamic extension of the Skellam model, while only the total strengths can be presented for the dynamic ordered probit model, since the separation into attack and defence strengths cannot be identified in the latter framework. The total strengths and toto probability estimates for the dynamic extensions of the Skellam and ordered probit models are presented in Figure E.3 of Appendix E. When we compare these results amongst the three score-driven models, the paths of the estimated total strengths are clearly different, although the main patterns appear to be similar. Hence, the question as to which model is best at forecasting the toto outcome is relevant and of interest. We have already presented the results for the two rival teams from the Spanish Primera división, but, for the sake of completeness, we also present graphs similar to Fig. 1, with panels of attack, defence and overall strengths and a panel of toto probabilities, for the two main rival teams in the other national football league competitions; see Figures E.4–E.8 in Appendix E. These results merely illustrate that the estimated attack, defence and overall strengths are truly time-varying. It is interesting to view the strength increases in the last few years of our sample by teams such as Dortmund, Juventus, and Paris SG, as well as the more recent strength decreases of Manchester United.

4.3. Forecasting: study design and measurement precision

We produce probability forecasts for the toto outcomes of the next round of matches in six national league competitions and based on nine model categories, as described in the introduction and summarised in Table E.1 of Appendix E. For a description of the static models we refer to Karlis and Ntzoufras (2003), Koning (2000) and Maher (1982) for example, and for a description of the semi-dynamic models we refer to Dixon and Coles (1997). The details of

Table 1
Descriptive statistics of six football competitions.

Competition	# Teams	# Matches	Mean(H)	Mean(A)	Var(H)	Var(A)
In-sample, 1999–2009						
English Premier League	20	3800	1.503	1.092	1.649	1.181
German Bundesliga	18	3060	1.673	1.185	1.781	1.274
Spanish Primera división	20	3800	1.533	1.116	1.566	1.169
Italian Serie A	18/20	3430	1.505	1.089	1.434	1.109
French Ligue 1	18/20	3578	1.382	0.913	1.381	0.960
Dutch Eredivisie	18	3060	1.766	1.242	2.195	1.478
Out-of-sample, 2009–2016						
English Premier League	20	2660	1.573	1.171	1.752	1.312
German Bundesliga	18	2142	1.617	1.281	1.804	1.462
Spanish Primera división	20	2660	1.627	1.121	1.947	1.343
Italian Serie A	20	2660	1.496	1.120	1.497	1.174
French Ligue 1	20	2660	1.420	1.050	1.409	1.151
Dutch Eredivisie	18	2142	1.789	1.325	1.956	1.558

Notes: The table reports the in-sample and out-of-sample characteristics of the six football competitions that are considered in our forecasting study. The column '# Teams' reports the number of teams that are active in one season of the respective competition. For the Italian Serie A, the in-sample data set has 5×18 and 5×20 teams in a season. For the French Ligue 1, the in-sample data set has 3×18 and 7×20 teams in a season. The column labeled Mean$(\cdot)$ and Var$(\cdot)$ are the sample mean and sample variance of home (H) and away (A) goals.

Table 2
Maximum likelihood estimates of the parameters in ψ^*.

	a_1	a_2	b_1	b_2	δ	λ_3	κ_1	κ_2
Biv. Poisson								
Premier League	0.012	0.016	1.000	0.999	0.331	0.083		
Bundesliga	0.014	0.014	0.999	0.997	0.365	0.043		
Primera división	0.015	0.014	0.998	0.999	0.350	0.088		
Serie A	0.016	0.021	0.998	0.998	0.365	0.199		
Ligue 1	0.014	0.015	0.998	0.998	0.402	0.043		
Eredivisie	0.016	0.017	0.998	0.998	0.386	0.052		
Skellam								
Premier League	0.015	0.018	1.000	0.998	0.341			
Bundesliga	0.028	0.014	0.998	0.996	0.386			
Primera división	0.016	0.022	1.000	1.000	0.365			
Serie A	0.022	0.031	0.998	1.000	0.395			
Ligue 1	0.015	0.023	1.000	1.000	0.423			
Eredivisie	0.020	0.022	1.000	0.999	0.381			
Ordered probit								
Premier League	0.027		0.998				−0.665	0.096
Bundesliga	0.026		0.999				−0.647	0.047
Primera división	0.018		1.000				−0.646	0.064
Serie A	0.032		0.996				−0.757	0.109
Ligue 1	0.020		1.000				−0.758	0.079
Eredivisie	0.036		0.998				−0.653	0.042

Notes: The table presents the maximum likelihood estimates of the parameters in ψ^* for the dynamic bivariate Poisson, Skellam, and ordered probit models, which are defined in Section 3.2. The initialisation of f_1 is based on static estimates for one year of competition in 1999–2000; see the discussion in Appendix D. The dataset used to obtain the parameter estimates is from 2000 to 2009; see also Table 1.

our dynamic extension based on score-driven models are discussed in Section 3. The probability forecasts for the toto results are discussed in Appendix A and computed by Eqs. (A.1), (A.2), and (A.3), where the strengths of attack and defence (or overall strength) are treated as either static or time-varying.

Before computing the forecasts for round $t + 1$, all static parameters (whether in ψ or ψ^*) are re-estimated using all data up to time t. The first forecasts are for the toto probabilities of all matches in the first round of the football season 2009–2010, and are based on the parameter estimates from the data panel of the previous ten seasons 1999–2009. These computations are repeated for each model, dynamic extension and method. In the case of the score-driven model, we evaluate f_t recursively, and at the end of the estimation sample we obtain f_{t+1}, from which, together with the static parameter estimates, the probability forecasts can be computed. Given the realised match results and their forecasts, we can evaluate a loss function to measure the forecast precision; see the details below. When forecasting the next round of football matches, we re-estimate the parameter vector after including the football match results of the most recent round in our data set. Hence, we have an expanding estimation sample in order to ensure that we can utilize as much data as possible for estimation. Thus, the procedure for our forecasting study is simple. After each round of matches, we re-estimate the static parameters, filter the time-varying parameters

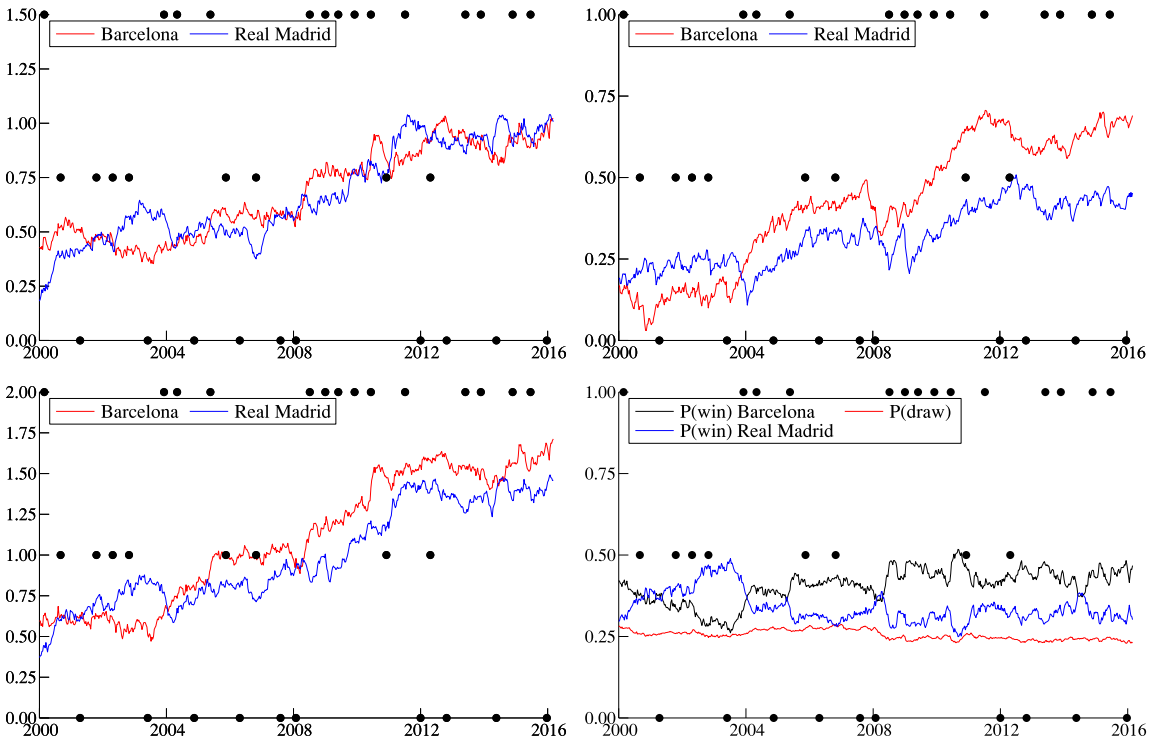

Fig. 1. Estimated strengths for Barcelona and Real Madrid. Notes: The figure displays the estimated attack, defence and total strengths from the score-driven dynamic extension of the bivariate Poisson model for the two top teams in the Spanish Primera división: Barcelona (red) and Real Madrid (blue). All panels: a dot at the top of the panel represents a win for Barcelona, a dot at the bottom is a win for Real Madrid, and a dot in the middle of the panel represents a draw. Top left panel: time series plot of extracted strengths of attack. Top right panel: time series plot of extracted strengths of defence. Bottom left panel: sum of extracted strengths of attack and defence. Bottom right panel: probability of toto results from the dynamic bivariate Poisson model: Barcelona win (red), Real Madrid win (blue) and draw (black). Note that these graphs do not take the home ground advantage into account; that is, δ is set to zero in Eq. (3), to enable better comparisons of the two teams. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(if any), and forecast the toto probabilities in the next round. We repeat these steps for each round in the seven consecutive football seasons and for each of the six football competitions.

Given the forecasted probabilities of a win, loss and draw for a match and the realised toto result for each match in a round, we can measure the precision of our forecasts for this round as follows. For example, assume that we have two rival models that produce probability forecasts for the toto outcome of a football match as follows: Model I has $P(\text{win}) = 0.50$, $P(\text{draw}) = 0.40$, and $P(\text{loss}) = 0.10$; Model II has $P(\text{win}) = 0.50$, $P(\text{draw}) = 0.30$, and $P(\text{loss}) = 0.20$. The outcome of the match is a home win. If we opt for the log-loss function, we would have $\log(0.50)$ as the loss probability for this match for both models, while the probabilities assigned to the other possible outcomes $P(\text{draw})$ and $P(\text{away})$ are ignored. The rank probability score (RPS) is a loss function that accounts for the other probabilities as well; see Epstein (1969) and, for an application to football match results, Constantinou and Fenton (2012). For the toto forecasts, the RPS statistic is given by

$$RPS = 0.5 \sum_{k=1}^{3} \left(cdf_{f,k} - cdf_{y,k} \right)^2, \tag{17}$$

where $cdf_{f,k}$ and $cdf_{y,k}$ are the cumulative density functions of the forecast and the realised outcome, respectively. In our example, we have $RPS_I = [(0.5-1)^2 + (0.9-1)^2 + (1-1)^2]/2 = 0.13$ and $RPS_{II} = [(0.5-1)^2 + (0.8-1)^2 + (1-1)^2]/2 = 0.145$ for Models I and II, respectively. Hence, the probability forecasts from Model I are more precise. We average the RPS statistic over all football matches in a round and take this average as our loss function. We facilitate model comparisons by collecting the value of this loss function, for each round, in a loss vector and using it to compute the Diebold Mariano (DM) test statistic for equal predictive accuracy; see Diebold and Mariano (1995). The DM statistic is distributed asymptotically as a standard normal random variable, and hence rejects the null hypothesis of equal predictive accuracy at the 5% significance level if the DM test statistic is smaller than -1.96 (the benchmark model performs significantly worse) or larger than 1.96 (the benchmark model performs significantly better). We also report the ARPS, which is defined as the average of the RPS statistic over all rounds in the football season, and over all seven years in our out-of-sample data set (it is simply the average of values in the loss vector).

Finally, given the forecasted probabilities, we can verify whether any of the categories home win, draw, or away win is over- or under-represented in our forecasts. We do this by summing the forecasted probabilities in each of the three categories and comparing them with the realised numbers in each category. The results are provided in Table

E.2. We conclude that the shifting of probability mass can be beneficial for some competitions and some models, see the studies by Dixon and Coles (1997) and Koopman et al. (2017).

4.4. Betting simulation

Section 4.3 considered the use of the RPS loss function for the purpose of model comparisons in terms of forecast precision. An alternative basis for comparisons of football models is to verify whether a model is able to generate a profit over the bookmakers' odds when predicting the outcome of a football match. For this purpose, we perform a betting simulation. The resulting profit or loss from this betting exercise provides an additional loss function that can be used for model comparisons. The design of the betting simulation is as follows. For each match prediction, we have the bookmakers' odds together with the forecast probabilities, as described in Section 4.3. On the basis of this information for each football match, we can calculate three expected values (EV), namely the EVs for a home win, draw, and away win. The expected value of a unit bet on an event A is given by

$$\begin{aligned}\mathrm{EV}(A) &= P(A)\{\mathrm{odds}(A) - 1\} - P(\text{not } A) \times 1 \\ &= P(A)\mathrm{odds}(A) - 1,\end{aligned}$$

where event A represents a win, a loss or a draw for the home team, $P(A)$ is the probability of event A and odds(A) is the bookmaker's odds for event A; see also Koopman and Lit (2014) for a more detailed description. Our betting strategy is basic: we bet one unit on whichever of the three categories (home win, draw, away win) has the highest EV and for which it holds that EV > 0. More advanced betting strategies, such as Kelly betting or alternative versions of it, that take into account the odds and EV in order to find the optimal stake on a wager, will outperform our simple strategy, but are considered to be beyond the scope of this paper. An interesting result of this betting simulation is that it allows us to determine whether the models are able to generate a profit over the bookmakers' odds with *only* the past match results as input.

4.5. Results of forecasting study

Tables 3, 4, and 5 present the forecast comparisons based on ARPS, DM statistics, and the losses in our betting simulation study, respectively, where we consider the three static models and their dynamic extensions with two initialization strategies. For the dynamic models, we also include the non-stationary specification of $b_1 = b_2 = 1$, which implies that the time-varying strengths are modelled as random walks. This extension allows us to investigate the differences in forecasting capabilities between the stationary and the more parsimonious non-stationary dynamic models. We report these forecasting results for the six European football competitions and present a summary of our findings. Given that we have made an average of ± 2500 forecasts for each competition, we may regard our forecasting study as extensive. We learn from the reported results that our dynamic extension, based on the score-driven model with f_1 estimated using the first season of the in-sample data set, is the best performing forecasting strategy for all six European football competitions. It is only for the Spanish Primera división that a constrained version of our model with $b_1 = b_2 = 1$ (leading to a random walk updating for the time-varying parameter f_t) performs the best for forecasting.

In regard to forecast precision, our score-driven dynamic extensions outperform both the static model and the semi-dynamic model extension significantly in almost all cases. The Dutch Eredivisie is the exception: the improvement is not strongly significant. We also learn that estimating f_1 as part of ψ^* introduces too much uncertainty into the parameter space, as the associated forecasts are significantly worse almost in all cases. The best strategy for initialization (that is, setting a value for f_1) appears to be obtained by estimating the static model using the first season in the sample.

Tables 3 and 4 provide convincing evidence that the bivariate Poisson and Skellam distributions are preferred overall. These two models are almost always preferred over the ordered probit model. The Dutch Eredivisie is the only exception, with the dynamic ordered probit model being preferred in terms of forecast precision. However, the superiority of the ordered probit model over the other dynamic models is never significant. Thus, we may conclude that the condensation of data has a negative impact on the forecast precision: when data is recorded in a more condensed manner, we lose information. The counts of the numbers of goals made by the two teams in a football match, or the difference between these two counts, contains more information than the sign of the difference (and zero). However, we could also have opted for some bivariate distribution other than the Poisson, such as the bivariate negative binomial distribution; see Famoye (2010) for all relevant details. Its dynamic extension can be implemented in a similar way to the Poisson, but the scaled score function for updating the time-varying parameters will be different. However, the reported data descriptives in Table 1 do not provide much evidence of over-dispersion in the number of goals scored, with only some possible evidence for the Dutch Eredivisie competition. Although we do not report them, we have produced forecasting results for the bivariate negative binomial model; however, we have not found any improvements when compared to the Poisson model.

The results of the betting simulation for each competition, as described in Section 4.4, are presented in Table 5. When only past match results are considered and no other explanatory variables are used, none of the models considered are able to beat the bookmaker. Thus, we may conclude that improvements in the field of sports modelling should be directed towards finding more and better explanatory variables, rather than towards developing better models that try to extract more information from noisy data. Overall, the results of the betting simulation show that the dynamic Skellam model outperforms the bivariate Poisson model. In particular, the performance of the dynamic Skellam model is relatively strong for the larger competitions of England, Germany and Spain. In accordance with the RPS results, the ordered probit models also perform poorly in a betting context.

Table 3
Rank probability scores and Diebold–Mariano test results (England, Germany, Spain).

Distribution	Type	Additional specifications	Time	# par	England		Germany		Spain	
					ARPS	DM	ARPS	DM	ARPS	DM
Biv. Poisson	Static	$\sum_{i=1}^{N}\alpha_i = 0$	31.20	$2N+2$	0.2062	6.17	0.2165	4.76	0.1980	5.98
Skellam	Static	$\sum_{i=1}^{N}\alpha_i = 0$	30.95	$2N+1$	0.2068	6.68	0.2167	4.81	0.1983	6.06
Ordered probit	Static	$\sum_{i=1}^{N}\alpha_i = 0$	6.68	$N+2$	0.2083	6.69	0.2173	5.07	0.1989	5.74
Biv. Poisson	Semi-dynamic	$\sum_{i=1}^{N}\alpha_i = 0$	116.47	$2N+2$	0.2014	4.15	0.2127	3.35	0.1942	4.28
Skellam	Semi-dynamic	$\sum_{i=1}^{N}\alpha_i = 0$	65.53	$2N+1$	0.2034	5.55	0.2156	4.47	0.1944	3.88
Ordered probit	Semi-dynamic	$\sum_{i=1}^{N}\alpha_i = 0$	40.56	$N+2$	0.2027	4.39	0.2138	3.88	0.1952	4.01
Biv. Poisson	Dynamic	$f_1 \in \psi^*$	183.30	$2N+6$	0.2024	4.69	0.2126	2.99	0.1952	4.54
Skellam	Dynamic	$f_1 \in \psi^*$	252.45	$2N+5$	0.2020	4.69	0.2120	2.54	0.1953	4.34
Ordered probit	Dynamic	$f_1 \in \psi^*$	41.65	$N+4$	0.2030	3.93	0.2140	3.27	0.1962	4.61
Biv. Poisson	Dynamic	$b_1 = b_2 = 1, f_1$ static init	1.72	4	0.1987	2.44	0.2095	2.15	0.1915	–
Skellam	Dynamic	$b_1 = b_2 = 1, f_1$ static init	1.30	3	0.1982	1.01	0.2092	0.73	0.1918	1.03
Ordered probit	Dynamic	$b_1 = 1, f_1$ static init	0.80	3	0.2001	3.06	0.2104	2.67	0.1929	3.09
Biv. Poisson	Dynamic	f_1 static init	8.44	6	0.1984	1.22	0.2089	–	0.1916	1.57
Skellam	Dynamic	f_1 static init	3.76	5	0.1981	–	0.2089	0.01	0.1918	1.03
Ordered probit	Dynamic	f_1 static init	1.30	4	0.1996	2.32	0.2103	2.51	0.1929	3.09
Biv. Poisson	Dynamic	State space model	1 h	6	0.1984	1.26	0.2099	2.10	0.1928	3.04

Notes: The table presents probability forecast losses based on the average rank probability scores (ARPS) obtained by the forecasting of seven seasons of match results for each of six European football competitions with an expanding window. The RPS are averaged over the out-of-sample window in order to get a summarizing statistic. The Diebold–Mariano (DM) statistic is based on the entire out-of-sample window with one-step-ahead forecasts. The lowest ARPS for each column is shaded (in blue). The column '# par' indicates the dimension of the parameter vector ψ or ψ^*, where N is the number of teams in the data set (all seasons combined). The DM statistic is distributed asymptotically as a standard normal random variable, and hence rejects the null hypothesis of equal predictive accuracy at the 5% significance level if the DM test statistic is either smaller than -1.96 (the benchmark model performs significantly worse) or larger than 1.96 (the benchmark model performs significantly better). The benchmark model performs the best for each competition. The column 'Time' indicates the average computer time (in seconds) taken to maximize the log-likelihood function (averaged over the six competitions). All computations are performed on a i7-2600, 3.40 GHz desktop PC using four cores. Note that the "$f_1 \in \psi^*$" specification has all elements of vector f_1 placed in the parameter vector ψ^* that is estimated by maximum likelihood, whereas "f_1 static init" has the vector f_1 estimated by static regression using the first season; see Appendix D.

Table 4
Rank probability scores and Diebold–Mariano test results (Italy, France, Netherlands).

Distribution	Type	Additional specifications	Time	# par	Italy		France		Netherlands	
					ARPS	DM	ARPS	DM	ARPS	DM
Biv. Poisson	Static	$\sum_{i=1}^{N}\alpha_i = 0$	31.20	$2N+2$	0.2042	3.44	0.2109	4.34	0.1974	2.34
Skellam	Static	$\sum_{i=1}^{N}\alpha_i = 0$	30.95	$2N+1$	0.2049	3.77	0.2113	4.56	0.1976	2.41
Ordered probit	Static	$\sum_{i=1}^{N}\alpha_i = 0$	6.68	$N+2$	0.2050	3.82	0.2126	5.23	0.1973	2.29
Biv. Poisson	Semi-dynamic	$\sum_{i=1}^{N}\alpha_i = 0$	116.47	$2N+2$	0.2025	2.78	0.2070	1.28	0.1939	0.58
Skellam	Semi-dynamic	$\sum_{i=1}^{N}\alpha_i = 0$	65.53	$2N+1$	0.2036	3.45	0.2077	2.02	0.1944	0.88
Ordered probit	Semi-dynamic	$\sum_{i=1}^{N}\alpha_i = 0$	40.56	$N+2$	0.2033	3.12	0.2086	2.66	0.1943	1.05
Biv. Poisson	Dynamic	$f_1 \in \psi^*$	183.30	$2N+6$	0.2028	2.60	0.2082	2.55	0.1949	1.17
Skellam	Dynamic	$f_1 \in \psi^*$	252.45	$2N+5$	0.2027	2.42	0.2081	2.51	0.1942	0.70
Ordered probit	Dynamic	$f_1 \in \psi^*$	41.65	$N+4$	0.2031	2.69	0.2093	3.41	0.1938	0.43
Biv. Poisson	Dynamic	$b_1 = b_2 = 1, f_1$ static init	1.72	4	0.2011	2.27	0.2066	1.76	0.1940	0.73
Skellam	Dynamic	$b_1 = b_2 = 1, f_1$ static init	1.30	3	0.2008	0.80	0.2062	1.31	0.1942	1.07
Ordered probit	Dynamic	$b_1 = 1, f_1$ static init	0.80	3	0.2023	2.89	0.2071	2.21	0.1938	1.08
Biv. Poisson	Dynamic	f_1 static init	8.44	6	0.2005	–	0.2060	0.37	0.1936	0.31
Skellam	Dynamic	f_1 static init	3.76	5	0.2007	0.66	0.2059		0.1941	1.01
Ordered probit	Dynamic	f_1 static init	1.30	4	0.2019	2.22	0.2067	1.59	0.1934	–
Biv. Poisson	Dynamic	State space model	1 h	6	0.2010	1.84	0.2066	1.62	0.1938	0.45

Note: See the notes to Table 3.

Koopman et al. (2016) argued that the forecasting performances of univariate score-driven models are comparable to those of their parameter-driven (state space model) counterparts. Our study has confirmed this conclusion, but now for a class of multivariate score-driven and state space models. In terms of forecast precision, the score-driven model produces a lower forecast loss than the dynamic state space model; in fact, we even report a significant improvement in some instances. Finally, we also report the number of seconds of computer-time that are needed for maximizing the log-likelihood function for a single model. The differences in computing-time for parameter estimation are noteworthy: the estimation of the score-driven model requires < 10 s, while that of the state space model

Table 5
Unit losses from betting on matches (all six competitions).

Distribution	Type	Additional specifications	ENG	GER	SPA	ITA	FRA	NET	Total
Biv. Poisson	Static	$\sum_{i=1}^{N} \alpha_i = 0$	−179.20	−100.28	−275.28	−171.15	−159.95	−133.30	−1019.16
Skellam	Static	$\sum_{i=1}^{N} \alpha_i = 0$	−228.60	−98.46	−318.85	−230.14	−189.40	−113.84	−1179.29
Ordered probit	Static	$\sum_{i=1}^{N} \alpha_i = 0$	−157.47	−106.21	−258.94	−238.29	−199.65	−107.34	−1067.90
Biv. Poisson	Semi-dynamic	$\sum_{i=1}^{N} \alpha_i = 0$	−173.41	−95.11	−254.35	−276.05	−167.47	−145.64	−1112.03
Skellam	Semi-dynamic	$\sum_{i=1}^{N} \alpha_i = 0$	−168.96	−130.25	−289.94	−268.87	−76.26	−122.29	−1056.57
Ordered probit	Semi-dynamic	$\sum_{i=1}^{N} \alpha_i = 0$	−186.19	−94.15	−304.19	−236.86	−162.27	−92.19	−1075.85
Biv. Poisson	Dynamic	$f_1 \in \psi^*$	−142.52	−78.00	−251.92	−152.49	−188.58	−144.78	−958.29
Skellam	Dynamic	$f_1 \in \psi^*$	−128.90	−96.52	−307.68	−220.55	−162.52	−126.55	−1042.72
Ordered probit	Dynamic	$f_1 \in \psi^*$	−178.39	−77.33	−317.47	−168.64	−173.18	−66.58	−981.59
Biv. Poisson	Dynamic	$b_1 = b_2 = 1, f_1$ static init	−141.19	−79.57	−165.05	−314.59	−210.51	−132.78	−1043.69
Skellam	Dynamic	$b_1 = b_2 = 1, f_1$ static init	−81.54	−107.87	−157.58	−355.38	−183.13	−133.44	−1018.94
Ordered probit	Dynamic	$b_1 = 1, f_1$ static init	−145.91	−105.41	−282.80	−334.27	−224.66	−106.12	−1199.17
Biv. Poisson	Dynamic	f_1 static init	−157.88	−128.91	−188.53	−268.03	−214.68	−120.59	−1078.62
Skellam	Dynamic	f_1 static init	−57.34	−43.74	−126.89	−304.06	−206.75	−136.02	−506.34
Ordered probit	Dynamic	f_1 static init	−105.81	−125.77	−262.19	−251.29	−216.84	−87.38	−1049.28

The table presents unit losses from a betting simulation based on the forecasts of seven seasons of match results for each of the six European football competitions: England (ENG), Germany (GER), Spain (SPA), Italy (ITA), France (FRA), Netherlands (NET), and for all combined (Total). The forecasts are generated over time with an expanding window. The betting simulation is based on a simple strategy where a single unit wager is placed on the match outcome with the highest expected value. The lowest loss for each column is shaded (in blue). Bookmaker odds are obtained as the average odds taken from 10 to 20 large bookmakers.

requires approximately one hour. Thus, the score-driven models clearly outperform the state space model, in terms of both forecast precision and computer-time.

5. Conclusion

We have developed a multivariate score-driven model for analysing a high-dimensional panel of football match results. We then applied this score-driven methodology to three classes of models. In the first class, a match result is treated as a pairwise observation that is assumed to come from the bivariate Poisson distribution. The second class of models assumes that the difference between the numbers of goals, or the margin of victory of a team, is generated by the Skellam distribution. In the third class of models, the probability of a win, draw, or loss of a match is modelled by an ordered probit model. These different model classes with their different variables require somewhat different statistical treatments, but can be extended with time-varying parameters using the same score-driven framework. All three approaches are able to forecast toto probabilities for football matches in a national league competition. We then used a large-scale forecasting study to investigate which of the three model classes performs best in forecasting the toto probabilities in the next round of the competition. For this purpose, we have used a large panel of match results from six European football competitions over a range of seasons. The results of the forecasting study, which included a betting simulation, show that our score-driven football models outperform a range of benchmark models in both forecast precision and lowest betting losses. The dynamic bivariate Poisson model and the dynamic Skellam model turn out to perform the best for forecasting overall, while the ordered probit model almost never produces a more precise forecast. We may conclude that the subsequent merging of data (from two counts, to the difference in counts, to the sign of the difference) leads to a decrease in forecasting performance by reducing the informational content of the data, which is key for signal extraction.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2018.10.011.

References

Alzaid, A., & Omair, M. A. (2010). On the Poisson difference distribution inference and applications. *Bulletin of the Malaysian Mathematical Sciences Society*, *33*(1), 17–45.

Angelini, G., & De Angelis, L. (2017). Parx model for football match predictions. *Journal of Forecasting*, *36*(7), 795–807.

Baboota, R., & Kaur, H. (2018). Predictive analysis and modelling football results using machine learning approach for English premier league. *International Journal of Forecasting*, *34*, (in press).

Blasques, F., Koopman, S. J., & Lucas, A. (2015). Information theoretic optimality of observation driven time series models for continuous responses. *Biometrika*, *102*(2), 325–343.

Boshnakov, G., Kharrat, T., & McHale, I. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*(2), 458–466.

Cattelan, M., Varin, C., & Firth, D. (2013). Dynamic Bradley-Terry modelling of sports tournaments. *Applied Statistics*, *62*(1), 135–150.
Constantinou, A. C., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, *8*(1), 1559-0410.
Creal, D. D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777–795.
Crowder, M., Dixon, M. J., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English football league matches for betting. *The Statistician*, *51*(2), 157–168.
Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 253–265.
Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, *46*(2), 265–280.
Dixon, M. J., & Robinson, M. E. (1998). A birth process model for association football matches. *The Statistician*, *47*(3), 523–538.
Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*, 985–987.
Fahrmeir, L., & Tutz, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association*, *89*(428), 1438–1449.
Famoye, F. (2010). On the bivariate negative binomial regression model. *Journal of Applied Statistics*, *37*(6), 969–981.
Forrest, D., & Simmons, R. (2000). Making up the results: The work of the football pools panel, 1963–1997. *The Statistician*, *49*(2), 253–260.
Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, *21*, 331–340.
Goddard, J., & Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, *23*, 51–66.
Greene, W. H. (2012). *Econometric analysis*. New York: Pearson Education.
Held, L., & Vollnhals, R. (2005). Dynamic rating of European football teams. *IMA Journal of Management Mathematics*, *16*, 121–130.
Hvattum, L. M., & Arntzen, H. (2010). Using Elo ratings for match result prediction in association football. *International Journal of Forecasting*, *26*, 460–470.
Irwin, J. O. (1937). The frequency distribution of the difference between two independent variates following the same Poisson distribution. *Journal of the Royal Statistical Society, Series A*, *100*(3), 415–416.
Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete multivariate distributions*. New York: John Wiley & Sons.
Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician*, *52*(3), 381–393.
Karlis, D., & Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, *20*, 133–145.
Knorr-Held, L. (2000). Dynamic rating of sports teams. *The Statistician*, *49*(2), 261–276.
Kocherlakota, S., & Kocherlakota, K. (1992). *Bivariate discrete distributions*. New York: Dekker.
Koning, R. H. (2000). Balance in competition in Dutch soccer. *The Statistician*, *49*(3), 419–431.
Koopman, S. J., & Lit, R. (2014). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society, Series A*, *178*(1), 167–186.
Koopman, S. J., Lit, R., & Lucas, A. (2017). Intraday stochastic volatility in discrete price changes: the dynamic Skellam model. *Journal of the American Statistical Association*, *112*, 1490–1503.
Koopman, S. J., Lucas, A., & Scharth, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics & Statistics*, *98*(1), 97–110.
Lit, R. (2016). *Time-varying parameter models for discrete valued time series. Number 642 in Tinbergen Institute research series* (Ph.D dissertation), Amsterdam: Thela Thesis and Tinbergen Institute.
Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118.
Owen, A. (2011). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, *22*, 99–113.
Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician*, *49*(3), 399–418.
Skellam, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society*, *109*(3), 296.