International Journal of Forecasting 33 (2017) 458–466

Contents lists available at ScienceDirect

International Journal of Forecasting

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

A bivariate Weibull count model for forecasting association football scores

Georgi Boshnakov[a], Tarak Kharrat[a,b], Ian G. McHale[b,*]

[a] School of Mathematics, University of Manchester, UK

[b] Centre for Sports Business, Salford Business School, University of Salford, UK

ARTICLE INFO

Keywords:
Betting
Calibration
Copula
Counting process
Soccer
Weibull

ABSTRACT

The paper presents a model for forecasting association football scores. The model uses a Weibull inter-arrival-times-based count process and a copula to produce a bivariate distribution of the numbers of goals scored by the home and away teams in a match. We test it against a variety of alternatives, including the simpler Poisson distribution-based model and an independent version of our model. The out-of-sample performance of our methodology is illustrated using, first, calibration curves, then a Kelly-type betting strategy that is applied to the pre-match win/draw/loss market and to the over–under 2.5 goals market. The new model provides an improved fit to the data relative to previous models, and results in positive returns to betting.

1. Introduction

Since the seminal paper by Maher (1982), a considerable amount of effort has been invested in modelling the probability distribution of scores in association football. Maher's model assumes that the numbers of goals scored by each team in a football match follow independent Poisson processes, and that the rates at which the two teams can expect to score goals are functions of their respective abilities in attack and defence. Subsequent efforts have enhanced the Maher model in a variety of directions. Dixon and Coles (1997) make two enhancements to Maher's model: first, they allow for dependence between the goals scored by the two teams; and second, they address the dynamic nature of teams' abilities by using a time-decay function in the likelihood, so that more recent results affect a team's estimated strength parameters more than results further in the past. Rue and Salvesen (2000) address the dynamic nature of teams' abilities in a Bayesian framework, as does Owen (2011). Karlis and Ntzoufras (2003) use a bivariate Poisson model with diagonal inflation, so that the probabilities of draw scores are calibrated better than with the simple independent Poisson model. Most recently, Koopman and Lit (2015) use a state space model to allow team strengths to vary stochastically with time.

These models all assume that the basic scoring pattern in football follows a (time-homogeneous) Poisson process. This assumption may be made more out of convenience than for any other reason, since there are surprisingly few natural alternatives, other than the negative binomial distribution.

Here, we propose the use of a count process that is derived when the inter-arrival times are assumed to follow an independent and identically distributed Weibull distribution. We refer to this model as the Weibull count distribution, and the form of the distribution for the count process generated by Weibull inter-arrival times was not known until recently. However, McShane, Adrian, Bradlow, and Fader (2008) derived this distribution, meaning that a new, more general, count process model can now be

[*] Corresponding author.
E-mail address: i.mchale@salford.ac.uk (I.G. McHale).

http://dx.doi.org/10.1016/j.ijforecast.2016.11.006
0169-2070/

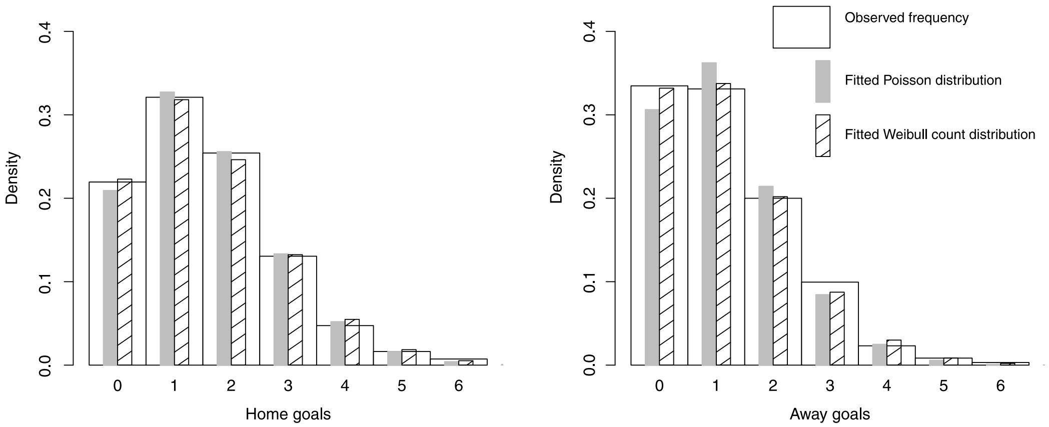

Fig. 1. Histograms of home (left) and away (right) goals with the fitted Poisson and Weibull count models. The estimated parameters for the home team are $\lambda_H = 1.50\ (0.04)$, $c_H = 1.56\ (0.03)$, while those for the away team are $\lambda_A = 1.10\ (0.03)$ and $c_A = 0.85\ (0.04)$, where the figures in parentheses are standard errors.

adopted. In addition to using a Weibull count model, we also allow for dependence between the goals scored by the two teams by employing a copula to generate a bivariate distribution allowing for positive or negative dependence.

Our objective in this paper is to build a model for the goals scored by the two teams in a football match. Our model can be used to construct the probabilities of the score-lines, and hence can be employed in betting market analysis or to study market efficiency, for example.

The computations and the graphs in the paper were done in R (R Core Team, 2016) using the Countr package (Kharrat & Boshnakov, 2016), available from the Comprehensive R Archive Network (CRAN).

The remainder of the paper is structured as follows. Section 2 presents the Weibull count distribution, our bivariate model, and provides our specification for its use when modelling the goals scored by the two teams in a football match. The results from fitting our model to data from the English Premier League are presented in Section 3, whilst the out-of-sample predictive performances, including the results of a simple Kelly-based betting strategy, are described in Section 4. We conclude with some closing remarks in Section 5.

2. A bivariate Weibull count distribution

2.1. The Weibull renewal process

McShane et al. (2008) derive the probability distribution of the number of events occurring by some time t when the inter-arrival times are assumed to be independent and identically distributed Weibull random variables (this process is also known as a Weibull renewal process). They do this by using a Taylor series expansion of the exponential in the Weibull density. They refer to the resulting count process as the '*Weibull count model*', and its probability mass function is given by

$$\Pr(X(t) = x) = \sum_{j=x}^{\infty} \frac{(-1)^{x+j}(\lambda t^c)^j \alpha_j^x}{\Gamma(cj+1)}, \tag{1}$$

where $\alpha_j^0 = \Gamma(cj+1)/\Gamma(j+1)$, $j = 0, 1, 2, \ldots$, and $\alpha_j^{x+1} = \sum_{m=x}^{j-1} \alpha_m^x \Gamma(cj - cm + 1)/\Gamma(j - m + 1)$, for $x = 0, 1, 2, \ldots$, and $j = x+1, x+2, x+3, \ldots$. In Eq. (1), λ is a 'rate' parameter and c is the 'shape' parameter of the distribution, where the observation unit is the match, which we take as having a duration of one time unit. Thus, λ is the scoring rate per match.

The use of the Weibull distribution to model the inter-arrival times allows the hazard $h(t)$ associated with the count process to vary over time. The Weibull hazard is given by

$$h(t) = \lambda c t^{c-1},$$

and may be monotonically increasing for $c > 1$, monotonically decreasing for $c < 1$, or constant (and equal to λ) for $c = 1$. Note that we recover the (time-homogeneous) Poisson process when $c = 1$. It is also interesting to note that this model handles both over-dispersed data (the mean is smaller than the variance; $c < 1$) and under-dispersed data (the mean is larger than the variance; $c > 1$) naturally, whilst the Poisson count distribution ($c = 1$) can only accommodate equi-dispersed data (the mean is equal to the variance).

Despite the somewhat intimidating appearance of Eq. (1), the computations for the Weibull count model can be performed without much trouble. For the usual values of the count (goals) observed in association football ($x \in [0, 10]$), the first 50 terms of the infinite series are sufficient to enable the probabilities to be computed accurately. For the sake of speed, we implemented these computations in C++, though McShane et al. (2008) were able to perform the computations in Microsoft Excel. We validated the computations by retrieving the Poisson case ($c = 1$) and reproducing the analysis conducted by McShane et al. (2008). All of the computations described in this paper can be reproduced using the R (R Core Team, 2016) add-on package Countr (Kharrat & Boshnakov, 2016).

Fig. 1 shows the Weibull count model and the Poisson distribution fitted to the goals scored by the home (left) and away (right) teams in matches played in the English Premier League during the five seasons from 2010–11 to