Guanhao Feng, Nicholas Polson* and Jianeng Xu

The market for English Premier League (EPL) odds

DOI 10.1515/jqas-2016-0039

Abstract: This paper employs a Skellam process to represent real-time betting odds for English Premier League (EPL) soccer games. Given a matrix of market odds on all possible score outcomes, we estimate the expected scoring rates for each team. The expected scoring rates then define the implied volatility of an EPL game. As events in the game evolve, we re-estimate the expected scoring rates and our implied volatility measure to provide a dynamic representation of the market's expectation of the game outcome. Using a dataset of 1520 EPL games from 2012–2016, we show how our model calibrates well to the game outcome. We illustrate our methodology on real-time market odds data for a game between Everton and West Ham in the 2015–2016 season. We show how the implied volatility for the outcome evolves as goals, red cards, and corner kicks occur. Finally, we conclude with directions for future research.

Keywords: English Premier League; market expectations; market odds; Skellam process; sports betting.

1 Introduction

1.1 The betting market for the EPL

Gambling on soccer is a global industry with revenues between $700 billion and $1 trillion a year (see "Football Betting - the Global Gambling Industry worth Billions." BBC Sport). Betting on the result of a soccer match is a rapidly growing market, and online real-time odds exists (Betfair, Bet365, Ladbrokes). Market odds for all possible score outcomes (0–0, 1–0, 0–1, 2–0,. . .) as well as outright win, lose and draw are available in real time. In this paper, we employ a two-parameter probability model based on a Skellam process and a non-linear objective function to extract the expected scoring rates for each team from the odds matrix. The expected scoring rates then define the implied volatility of the game.

***Corresponding author: Nicholas Polson,** Booth School of Business, University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637, USA, e-mail: nicholas.polson@chicagobooth.edu
Guanhao Feng: Booth School of Business, University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637, USA
Jianeng Xu: Department of Statistics, University of Chicago, 5747 S Ellis Avenue, Chicago, IL 60637, USA

A key feature of our analysis is to use the real-time odds to re-calibrate the expected scoring rates instantaneously as events evolve in the game. This allows us to assess how market expectations change according to exogenous events such as corner kicks, goals, and red cards. A plot of the implied volatility provides a diagnostic tool to show how the market reacts to event information. In particular, we study the evolution of the odds implied final score prediction over the course of the game. Our dynamic Skellam model fits the scoring data well in a calibration study of 1520 EPL games from the 2012–2016 seasons.

The goal of our study is to show how a parsimonious two-parameter model can flexibly model the evolution of the market odds matrix of final scores. We provide a non-linear objective function to fit our Skellam model to instantaneous market odds matrix. We then define the implied volatility of an EPL game and use this as a diagnostics to show how the market's expectation changes over the course of a game.

One advantage of viewing market odds through the lens of a probability model is the ability to obtain more accurate estimates of winning probabilities. For example, a typical market "vig" (or liquidity premium for bookmakers to make a return) is 5–8% in the win, lose, draw market. Now there is also extra information in the final score odds about the win odds. Our approach helps to extract that information. Another application of the Skellam process is to model final score outcomes as a function of characteristics (see Karlis and Ntzoufras (2003, 2009)).

The rest of the paper is outlined as follows. The next subsection provides connections with existing research. Section 2 presents our Skellam process model for representing the difference in goals scored. We then show how to make use of an odds matrix while calibrating the model parameters. We calculate a dynamic implied prediction of any score and hence win, lose and draw outcomes, using real-time online market odds. Section 3 illustrates our methodology using an EPL game between Everton and West Ham during the 2015–2016 season. Finally, Section 4 discusses extensions and concludes with directions for future research.

1.2 Connections with existing work

There is considerable interest in developing probability models for the evolution of the score of sporting events.

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

Stern (1994) and Polson and Stern (2015) propose a continuous time Brownian motion model for the difference in scores in a sporting event and show how to calculate the implied volatility of a game. We build on their approach by using a difference of Poisson processes (a.k.a. Skellam process) for the discrete evolution of the scores of an EPL game, see also Karlis and Ntzoufras (2003, 2009) and Koopman, Lit and Lucas (2014). Early probabilistic models (Lee 1997) predicted the outcome of soccer matches using independent Poisson processes. Later models incorporate a correlation between the two scores and model the number of goals scored by each team using bivariate Poisson models (see Maher (1982) and Dixon and Coles (1997)). Our approach follows Stern (1994) by modeling the score difference (a.k.a. margin of victory), instead of modeling the number of goals and the correlation between scores directly.

There is also an extensive literature on soccer gambling and market efficiency. For example, Vecer et al. (2009) estimates the scoring intensity in a soccer game from betting markets. Dixon and Pope (2004) presents a detailed comparison of odds set by different bookmakers. Fitt (2009) uses market efficiency to analyze the mispricing of cross-sectional odds and Fitt, Howls and Kabelka (2005) models online soccer spread bets.

Another line of research, asks whether betting markets are efficient and, if not, how to exploit potential inefficiencies in the betting market. For example, Levitt (2004) discusses the structural difference of the gambling market and financial markets. The study examines whether bookmakers are more skilled at game prediction than bettors and in turn exploit bettor biases by setting prices that deviate from the market clearing price. Avery and Chevalier (1999) examine the hypothesis that sentimental bettors act like noise traders and can affect the path of prices in soccer betting markets.

2 Skellam process for EPL scores

To model the outcome of a soccer game between team A and team B, we let the difference in scores, $N(t) = N_A(t) - N_B(t)$ where $N_A(t)$ and $N_B(t)$ are the team scores at time point t. Negative values of $N(t)$ indicate that team A is behind. We begin at $N(0) = 0$ and ends at time one with $N(1)$ representing the final score difference. The probability $\mathbb{P}(N(1) > 0)$ represents the ex-ante odds of team A winning. Half-time score betting, which is common in Europe, is available for the distribution of $N(\frac{1}{2})$.

We develop a probabilistic model for the distribution of $N(1)$ given $N(t) = \ell$ where ℓ is the current lead. This model, together with the current market odds can be used to infer the expected scoring rates of the two teams and then to define the implied volatility of the outcome of the match. We let λ^A and λ^B denote the expected scoring rates for the whole game. We allow for the possibility that the scoring abilities (and their market expectations) are time-varying, in which case we denote the expected scoring rates after time t by λ_t^A and λ_t^B respectively, instead of $\lambda^A(1-t)$ and $\lambda^B(1-t)$.

2.1 Implied score prediction from EPL odds

The Skellam distribution is defined as the difference between two independent Poisson variables, see Skellam (1946), Sellers (2012), Alzaid et al. (2010), and Barndorff-Nielsen and Shephard (2012). Karlis and Ntzoufras (2009) shows how Skellam distribution can be extended to a difference of distributions which have a specific trivariate latent variable structure. Following Karlis and Ntzoufras (2003), we decompose the scores of each team as

$$\begin{cases} N_A(t) = & W_A(t) + W(t) \\ N_B(t) = & W_B(t) + W(t) \end{cases} \tag{1}$$

where $W_A(t)$, $W_B(t)$ and $W(t)$ are independent processes with $W_A(t) \sim$ Poisson $(\lambda^A\, t)$, $W_B(t) \sim$ Poisson $(\lambda^B\, t)$. Here $W(t)$ is a non-negative integer-valued process to induce a correlation between the numbers of goals scored. By modeling the score difference, $N(t)$, we avoid having to specify the distribution of $W(t)$ as the difference in goals scored is independent of $W(t)$. Specifically, we have a Skellam distribution

$$\begin{aligned} N(t) &= N_A(t) - N_B(t) \\ &= W_A(t) - W_B(t) \sim \text{Skellam}(\lambda^A t, \lambda^B t). \end{aligned} \tag{2}$$

where $\lambda^A\, t$ is the cumulative expected scoring rate on the interval $[0, t]$. At time t, we have the conditional distributions

$$\begin{cases} W_A(1) - W_A(t) \sim & \text{Poisson}(\lambda^A(1-t)) \\ W_B(1) - W_B(t) \sim & \text{Poisson}(\lambda^B(1-t)) \end{cases} \tag{3}$$

Now letting $N^\star(1-t)$, the score difference of the sub-game which starts at time t and ends at time 1 and the duration is $(1-t)$. By construction, $N(1) = N(t) + N^\star(1-t)$. Since $N^\star(1-t)$ and $N(t)$ are differences of two Poisson process on two disjoint time periods, by the property of Poisson process, $N^\star(1-t)$ and $N(t)$ are independent. Hence, we can re-express equation (2) in terms of $N^\star(1-t)$, and deduce

$$N^\star(1-t) = W_A^\star(1-t) - W_B^\star(1-t) \sim \text{Skellam}(\lambda_t^A, \lambda_t^B) \tag{4}$$

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

where $W_A^\star(1-t) = W_A(1) - W_A(t)$, $\lambda^A = \lambda_0^A$ and $\lambda_t^A = \lambda^A(1-t)$. A natural interpretation of the expected scoring rates, λ_t^A and λ_t^B, is that they reflect the "net" scoring ability of each team from time t to the end of the game. The term $W(t)$ model a common strength due to external factors, such as weather. The "net" scoring abilities of the two teams are assumed to be independent of each other as well as the common strength factor. We can calculate the probability of any particular score difference, given by $\mathbb{P}(N(1) = x|\lambda^A, \lambda^B)$, at the end of the game where the λ's are estimated from the matrix of market odds. Team strength and "net" scoring ability can be influenced by various underlying factors, such as the offensive and defensive abilities of the two teams. The goal of our analysis is to only represent these parameters at every instant as a function of the market odds matrix for all scores.

To derive the implied winning probability, we use the law of total probability. The probability mass function of a Skellam random variable is the convolution of two Poisson distributions:

$$\begin{aligned}
\mathbb{P}(N(1) = x|\lambda^A, \lambda^B) &= k, \lambda^B)\mathbb{P}(W_A(1) = k|\lambda^A) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(W_B(1) = k - x|W_A(1) \\
&= \sum_{k=max\{0,x\}}^{\infty} \left\{ e^{-\lambda^B} \frac{(\lambda^B)^{k-x}}{(k-x)!} \right\} \left\{ e^{-\lambda^A} \frac{(\lambda^A)^k}{k!} \right\} \quad (5) \\
&= e^{-(\lambda^A+\lambda^B)} \sum_{k=max\{0,x\}}^{\infty} \frac{(\lambda^B)^{k-x}(\lambda^A)^k}{(k-x)!k!} \\
&= e^{-(\lambda^A+\lambda^B)} \left(\frac{\lambda^A}{\lambda^B}\right)^{x/2} I_{|x|}(2\sqrt{\lambda^A\lambda^B})
\end{aligned}$$

where $I_r(x)$ is the modified Bessel function of the first kind (for full details, see Alzaid et al. (2010)), thus has the series representation

$$I_r(x) = \left(\frac{x}{2}\right)^r \sum_{k=0}^{\infty} \frac{(x^2/4)^k}{k!\Gamma(r+k+1)}.$$

The probability of home team A winning is given by

$$\mathbb{P}(N(1) > 0|\lambda^A, \lambda^B) = \sum_{x=1}^{\infty} \mathbb{P}(N(1) = x|\lambda^A, \lambda^B). \quad (6)$$

In practice, we truncate the number of possible goals since the probability of an extreme score difference is negligible. Unlike the Brownian motion model for the evolution of the outcome in a sports game (Stern (1994), Polson and Stern (2015)), the probability of a draw in our setting is not zero. Instead, $\mathbb{P}(N(1) = 0|\lambda^A, \lambda^B) > 0$ depends on the sum and product of two parameters λ^A and λ^B and thus the odds of a draw are non-zero.

For two evenly matched teams with $\lambda^A = \lambda^B = \lambda$, we have

$$\begin{aligned}
\mathbb{P}(N(1) = 0|\lambda^A = \lambda^B = \lambda) &= e^{-2\lambda} I_0(2\lambda) \\
&= \sum_{k=0}^{\infty} \frac{1}{(k!)^2} \left(\frac{\lambda^k}{e^\lambda}\right)^2. \quad (7)
\end{aligned}$$

Figure 1 shows that this probability is a monotone decreasing function of λ and so two evenly matched teams with large λ's are less likely to achieve a draw.

Another quantity of interest is the conditional probability of winning as the game progresses. If the current lead at time t is ℓ, and $N(t) = \ell = N_A(t) - N_B(t)$, the Poisson property implied that the final score difference $(N(1)|N(t) = \ell)$ can be calculated by using the fact that

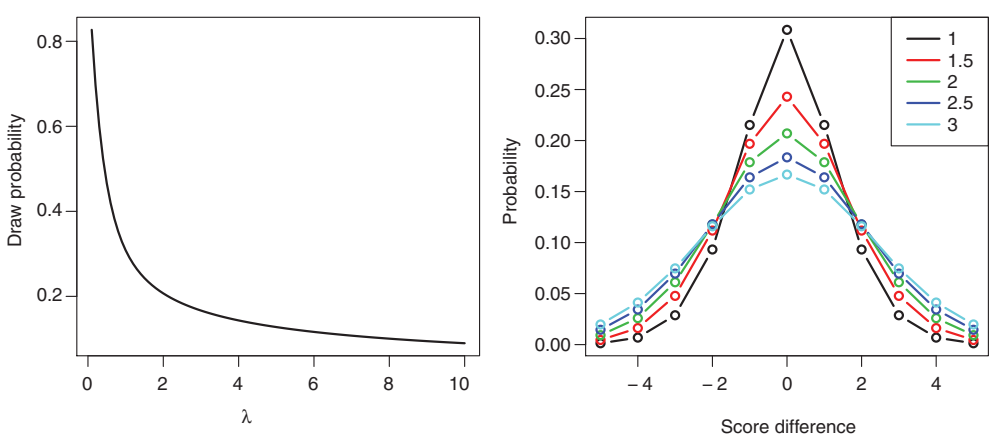

Figure 1: Left: Probability of a draw for two evenly matched teams. Right: Probability of score differences for two evenly matched teams. Λ values are denoted by different colors.

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

$N(1) = N(t) + N^\star(1-t)$ and $N(t)$ and $N^\star(1-t)$ are independent. Specifically, conditioning on $N(t) = \ell$, we have the identity

$$N(1) = N(t) + N^\star(1-t) = \ell + \text{Skellam}(\lambda_t^A, \lambda_t^B).$$

We are now in a position to find the conditional distribution $(N(1) = x|N(t) = \ell)$ for every time point t of the game given the current score. Simply put, we have the time homogeneous condition

$$\begin{aligned}\mathbb{P}(N(1) = x|\lambda_t^A, \lambda_t^B, N(t) = \ell) \\ = \mathbb{P}(N(1) - N(t) = x - \ell|\lambda_t^A, \lambda_t^B, N(t) = \ell) \\ = \mathbb{P}(N^\star(1-t) = x - \ell|\lambda_t^A, \lambda_t^B)\end{aligned} \tag{8}$$

where $\lambda_t^A, \lambda_t^B, \ell$ are given by market expectations at time t.

Two conditional probabilities of interest are he chances that the home team A wins,

$$\begin{aligned}\mathbb{P}(N(1) > 0|\lambda_t^A, \lambda_t^B, N(t) = \ell) \\ = \mathbb{P}(\ell + N^\star(1-t) > 0|\lambda_t^A, \lambda_t^B) \\ = \mathbb{P}(\text{Skellam}(\lambda_t^A, \lambda_t^B) > -\ell|\lambda_t^A, \lambda_t^B) \\ = \sum_{x > -\ell} e^{-(\lambda_t^A + \lambda_t^B)} \left(\frac{\lambda_t^A}{\lambda_t^B}\right)^{x/2} I_{|x|}(2\sqrt{\lambda_t^A \lambda_t^B}).\end{aligned} \tag{9}$$

and the conditional probability of a draw at time t is

$$\begin{aligned}\mathbb{P}(N(1) = 0|\lambda_t^A, \lambda_t^B, N(t) = \ell) \\ = \mathbb{P}(\ell + N^\star(1-t) = 0|\lambda_t^A, \lambda_t^B) \\ = \mathbb{P}(\text{Skellam}(\lambda_t^A, \lambda_t^B) = -\ell|\lambda_t^A, \lambda_t^B) \\ = e^{-(\lambda_t^A + \lambda_t^B)} \left(\frac{\lambda_t^A}{\lambda_t^B}\right)^{-\ell/2} I_{|\ell|}(2\sqrt{\lambda_t^A \lambda_t^B}).\end{aligned} \tag{10}$$

The conditional probability at time t of home team A losing is $1 - \mathbb{P}(N(1) > 0|\lambda_t^A, \lambda_t^B, N(t) = \ell)$. We now turn to the calibration of our model from given market odds.

2.2 Market calibration

Our information set at time t, denoted by $\mathcal{I}_t$, includes the current lead $N(t) = \ell$ and the market odds for $\{\text{Win, Lose, Draw, Score}\}_t$, where $\text{Score}_t = \{(i-j) : i, j = 0, 1, 2, \ldots\}$. These market odds can be used to calibrate a Skellam distribution which has only two parameters λ_t^A and λ_t^B. The best fitting Skellam model with parameters $\{\hat{\lambda}_t^A, \hat{\lambda}_t^B\}$ will then provide a better estimate of the market's information concerning the outcome of the game than any individual market (such as win odds) as they are subject to a "vig" and liquidity. Suppose that the fractional odds for all possible final score outcomes are given by a bookmaker. In this case, the bookmaker pays out three times the amount staked by the bettor if the outcome is indeed 2 - 1. Fractional odds are used in the UK, while money-line odds are favored by American bookmakers with 2 : 1 ("two-to-one") implying that the bettor stands to make a \$200 profit on a \$100 stake. The market implied probability makes the expected winning amount of a bet equal to 0. In this case, the implied probability $p = 1/(1+3) = 1/4$ and the expected winning amount is $\mu = -1^\star(1-1/4)+3^\star(1/4) = 0$. We denote this odds as $odds(2, 1) = 3$. To convert all the available odds to implied probabilities, we use the identity

$$\mathbb{P}(N_A(1) = i, N_B(1) = j) = \frac{1}{1 + odds(i, j)}.$$

The market odds matrix, O, with elements $o_{ij} = odds(i-1, j-1)$, $i, j = 1, 2, 3\ldots$ provides all possible combinations of final scores. Odds on extreme outcomes are not offered by the bookmakers. Since the probabilities are tiny, we set them equal to 0. The sum of the possible probabilities is still larger than 1 (see Dixon and Coles (1997) and Polson and Stern (2015)). This "excess" probability corresponds to a quantity known as the "market vig." For example, if the sum of all the implied probabilities is 1.1, then the expected profit of the bookmaker is 10%. To account for this phenomenon, we scale the probabilities to sum to 1 before estimation.

To estimate the expected scoring rates, λ_t^A and λ_t^B, for the sub-game $N^\star(1-t)$, the odds from a bookmaker should be adjusted by $N_A(t)$ and $N_B(t)$. For example, if $N_A(0.5) = 1$, $N_B(0.5) = 0$ and $odds(2, 1) = 3$ at half time, these observations actually says that the odds for the second half score being 1-1 is 3 (the outcomes for the whole game and the first half are 2-1 and 1-0 respectively, thus the outcome for the second half is 1-1). The adjusted $odds^\star$ for $N^\star(1-t)$ is calculated using the original odds as well as the current scores and given by

$$odds^\star(x, y) = odds(x + N_A(t), y + N_B(t)). \tag{11}$$

At time t $(0 \le t \le 1)$, we calculate the implied conditional probabilities of score differences using odds information

$$\begin{aligned}\mathbb{P}(N(1) = k|N(t) = \ell) = \mathbb{P}(N^\star(1-t) = k - \ell) \\ = \frac{1}{c} \sum_{i-j=k-\ell} \frac{1}{1 + odds^\star(i, j)}\end{aligned} \tag{12}$$

where $c = \sum_{i,j} \frac{1}{1+odds^\star(i,j)}$ is a scale factor, $\ell = N_A(t) - N_B(t)$, $i, j \ge 0$ and $k = 0, \pm 1, \pm 2\ldots$

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

Moments of the Poisson distribution make it straightforward to derive the moments of a Skellam random variable with parameters λ^A and λ^B. The unconditional mean and variance are given by

$$E[N(1)] = E[W_A(1)] - E[W_B(1)] = \lambda^A - \lambda^B,$$

$$V[N(1)] = V[W_A(1)] + V[W_B(1)] = \lambda^A + \lambda^B.$$

Therefore, the conditional moments are given by

$$\begin{cases} E[N(1)|N(t) = \ell] = \ell + (\lambda_t^A - \lambda_t^B), \\ V[N(1)|N(t) = \ell] = \lambda_t^A + \lambda_t^B. \end{cases} \tag{13}$$

We also need to ensure that $\hat{E}[N(1)|N(t) = \ell] - \ell \leq \hat{V}[N(1)|N(t) = \ell]$. A method of moments estimate of λ's is given by the solution to

$$\begin{cases} \hat{E}[N(1)|N(t) = \ell] = \ell + (\lambda_t^A - \lambda_t^B), \\ \hat{V}[N(1)|N(t) = \ell] = \lambda_t^A + \lambda_t^B, \end{cases} \tag{14}$$

where $\hat{E}$ and $\hat{V}$ are the expectation and variance calculated using market implied conditional probabilities, could be negative. To address this issue, we define the residuals

$$\begin{cases} D_E = \hat{E}[N(1)|N(t) = \ell] - [\ell + (\lambda_t^A - \lambda_t^B)], \\ D_V = \hat{V}[N(1)|N(t) = \ell] - (\lambda_t^A + \lambda_t^B). \end{cases} \tag{15}$$

We then calibrate parameters by adding the constraints $\lambda_t^A \geq 0$ and $\lambda_t^B \geq 0$ and solving the following equivalent constrained optimization problem.

$$\left(\hat{\lambda}_t^A, \hat{\lambda}_t^B\right) = \underset{\lambda_t^A, \lambda_t^B}{\arg\min} \left\{D_E^2 + D_V^2\right\} \quad \text{subject to } \lambda_t^A \geq 0, \lambda_t^B \geq 0 \tag{16}$$

Figure 2 illustrates a simulation evolution of an EPL game between Everton and West Ham (March 5th, 2016) with their estimated parameters. It provides a discretized version of Figure 1 in Polson and Stern (2015). The outcome probability of first half and updated second half are given in the left two panels. The top right panel illustrates a simulation-based approach to visualizing how the model works in the dynamic evolution of score difference. In the bottom left panel, from half-time onwards, we also simulate a set of possible Monte Carlo paths to the end of the game. This illustrates the discrete nature of our Skellam process and how the scores evolve.

2.3 Model diagnostics

To assess the performance our score-difference Skellam model calibration for the market odds, we have collected data from ladbrokes.com on the correct score odds of 18 EPL games (from October 15th to October 22nd, 2016) and plot the calibration result in Figure 3. The Q-Q plot of log(*odds*) is also shown. In average, there are 13 different outcomes per game, i.e., $N(1) = -6, -5, \ldots 0, \ldots, 5, 6$. In total 238 different outcomes are used. We compare our Skellam implied probabilities with the market implied probabilities for every outcome of the 18 games. If the model calibration is sufficient, all the data points should lie on the diagonal line.

Figure 3 left panel demonstrates that our Skellam model is calibrated by the market odds sufficiently well, except for the underestimated draw probabilities. Karlis and Ntzoufras (2009) describe this underestimation phenomenon in a Poisson-based model for the number of goals scored. Following their approach, we apply a zero-inflated version of Skellam distribution to improve the fit on draw probabilities, namely

$$\begin{cases} \tilde{P}(N(1) = 0) = p + (1 - p)P(N(1) = 0) \\ \tilde{P}(N(1) = x) = (1 - p)P(N(1) = x) \qquad \text{if } x \neq 0. \end{cases} \tag{17}$$

Here $0 < p < 1$ is an inflation factor and $\tilde{P}$ denotes the inflated probabilities. We also consider another type of inflation here

$$\begin{cases} \tilde{P}(N(1) = 0) = (1 + \theta)P(N(1) = 0) \\ \tilde{P}(N(1) = x) = (1 - \gamma)P(N(1) = x) \qquad \text{if } x \neq 0 \end{cases} \tag{18}$$

where θ is the inflation factor and $P(N(1) = 0) = \gamma/(\gamma+\theta)$.

Both types of inflation factors have the corresponding interpretation regarding the bookmakers' way of setting odds. With the first type of factor, the bookmakers generate two different set of probabilities, one specifically for the draw probability (namely the inflation factor p) and the other for all the outcomes using the Skellam model. The "market vig" for all the outcomes is a constant. With the second type, the bookmakers use the Skellam model to generate the probabilities for all the outcomes. Then they apply a larger "market vig" for draws than others. Yates (1982) also point out the "collapsing" tendency in forecasting behavior, whereby the bookmakers are inclined to report forecasts of 50% when they feel they know little about the event. In Figure 3 right panel, we see that the Skellam implied log(*odds*) has a heavier right tail than the market implied log(*odds*). This effect results from the overestimation of extreme outcomes, which in turn is due to market microstructure effect due to the market "vig".

To assess the out-of-sample predictive ability of the Skellam model, we analyze the market (win, lose, draw) odds for 1520 EPL games (from 2012 to 2016, 380 games

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

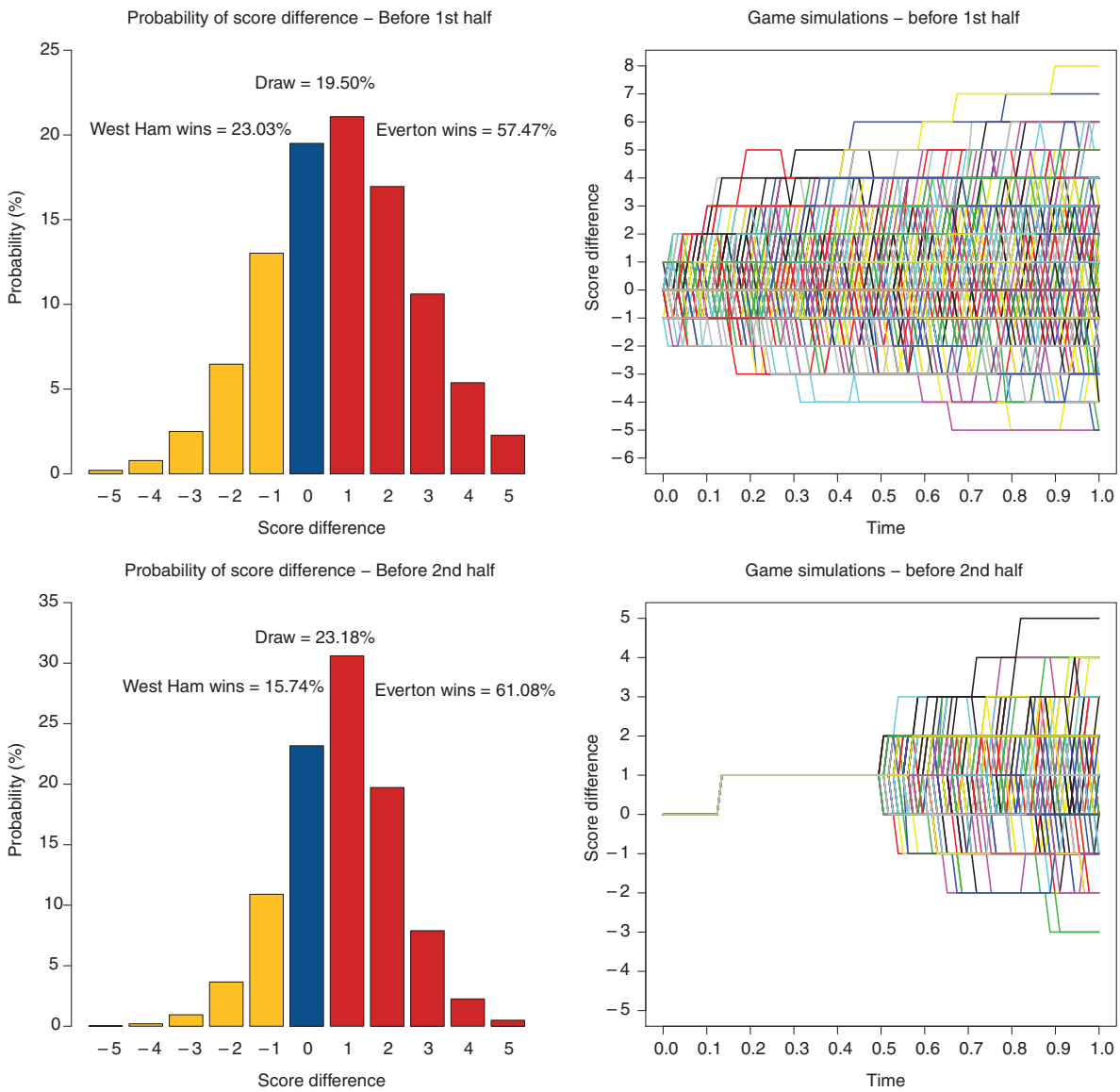

Figure 2: The Skellam process model for winning margin and game simulations. The top left panel shows the outcome distribution using odds data before the match starts. Each bar represents the probability of a distinct final score difference, with its color corresponding to the result of win/lose/draw. Score differences larger than 5 or smaller than −5 are not shown. The top right panel shows a set of simulated Skellam process paths for the game outcome. The bottom row has the two figures updated using odds data available at half-time.

per season). However, the sample covariance of the end of game scores, $N_A(1)$ and $N_B(1)$, is close to 0. If we assume parameters stay the same, then the estimates are $\hat{\lambda}^A = 1.5$ and $\hat{\lambda}^B = 1.2$. Since the probabilities of win, lose and draw sum to 1, we only plot the market implied probabilities of win and draw. In Figure 4 left panel, the draw probability is nearly a non-linear function of the win probability. To illustrate our model, we set the value of $\lambda^A\lambda^B = 1.5 \times 1.2 = 1.8$ and plot the curve of Skellam implied probabilities (red line). We further provide the inflated Skellam probabilities (blue line for the first type and green line for the second type). As expected, the non-inflated Skellam model (red line) underestimates the draw probabilities while the second type inflated Skellam model (green line) produces the better fit. We also group games by the market implied winning probability of home teams $P(N(1) > 0)$: (0.05, 0.1], (0.1, 0.15], $\cdots$, (0.8,0.85]. We calculate the frequency of home team winning for each group. In Figure 4 right panel, the barplot of frequencies (x-axis is regarding scaled odds) shows that the market is efficient, i.e., the frequency is close to the corresponding market implied probability and our Skellam model is calibrated to the market outcome for this dataset.

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

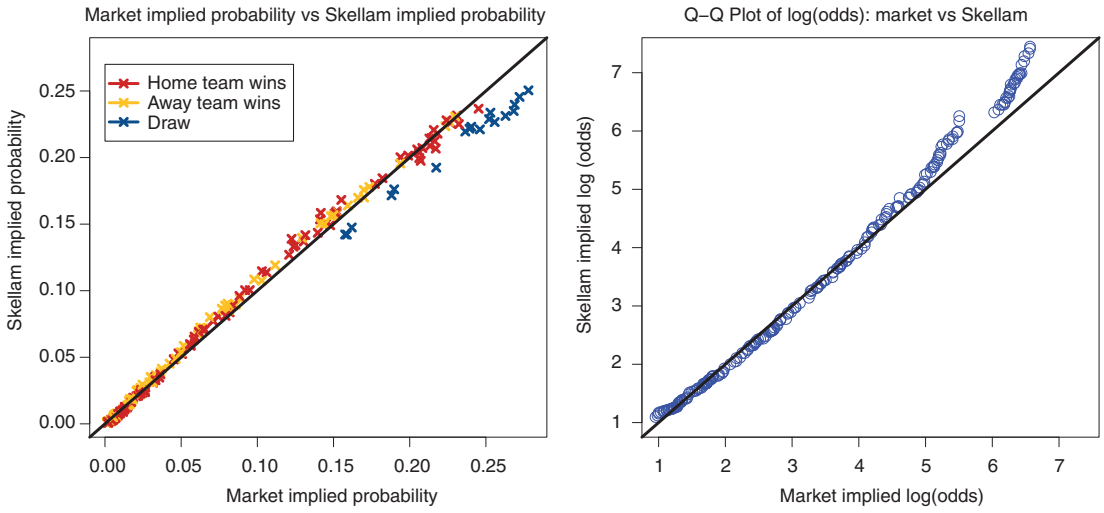

Figure 3: Left: Market implied probabilities for the score differences versus Skellam implied probabilities. Every data point represents a particular score difference; Right: Market log(odds) quantiles versus Skellam implied log(odds) quantiles. Market odds (from ladbrokes.com) of 18 games in EPL 2016–2017 are used (in average 13 score differences per game). The total number of outcomes is 238.

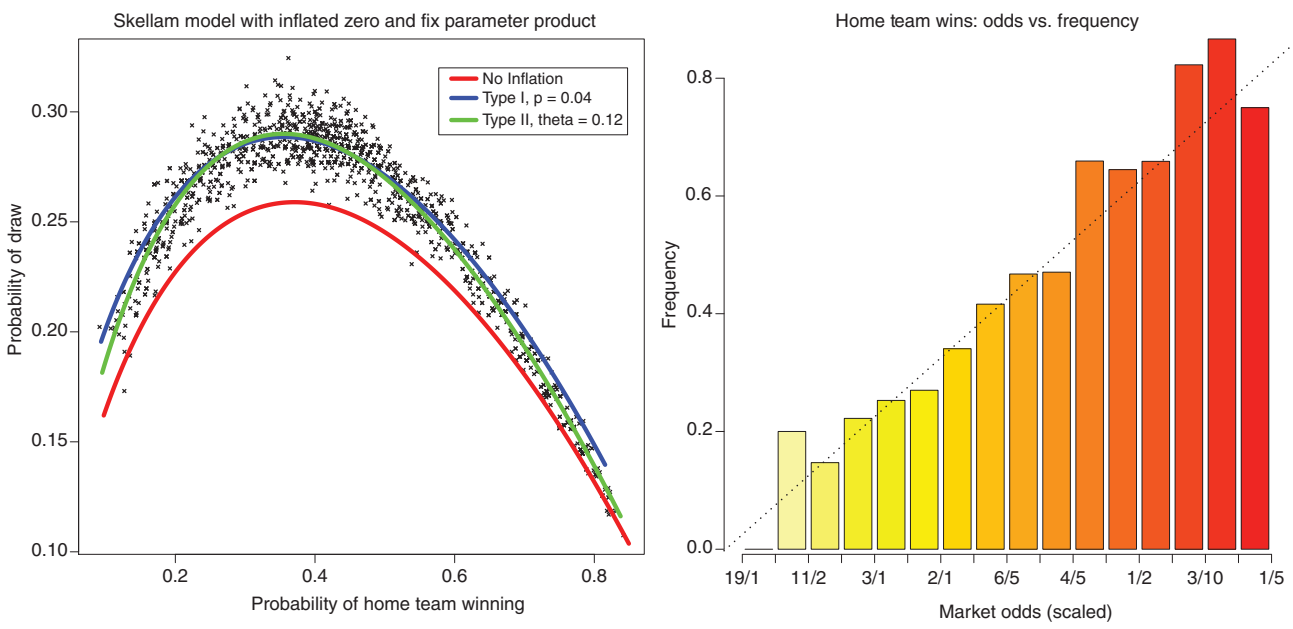

Figure 4: Left: Market implied probabilities of win and draw. The fitted curves are Skellam implied probabilities with fixed $\lambda^A\lambda^B = 1.8$. Right: Market odds and result frequency of home team winning. 1520 EPL games from 2012 to 2016 are used. The dashed line represents: Frequency = Market Implied Probability.

2.4 Time-varying extension

One extension that is clearly warranted is allowing for time-varying $\{\lambda_t^A, \lambda_t^B\}$ where the Skellam model is recalibrated dynamically through updated market odds during the game. We use the current $\{\lambda_t^A, \lambda_t^B\}$ to project possible results of the match in our Skellam model. Here $\{\lambda_t^A, \lambda_t^B\}$ reveal the market expectation of scoring difference for both teams from time t to the end of the game as the game progresses. Similar to the martingale approach of Polson and Stern (2015), $\{\lambda_t^A, \lambda_t^B\}$ reveal the best prediction of the game result. From another point of view, this approach is the same as assuming homogeneous rates for the rest of the game.

An alternative approach to time-varying $\{\lambda_t^A, \lambda_t^B\}$ is to use a Skellam regression with conditioning information such as possession percentages, shots (on goal), corner kicks, yellow cards, red cards, etc. We would expect jumps

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

in the $\{\lambda_t^A, \lambda_t^B\}$ during the game when some important events happen. A typical structure takes the form

$$\begin{cases} \log(\lambda_t^A) = & \alpha_A + \beta_A X_{A,t-1} \\ \log(\lambda_t^B) = & \alpha_B + \beta_B X_{B,t-1}, \end{cases} \tag{19}$$

estimated using standard log-linear regression.

Our approach relies on the betting market being efficient so that the updating odds should contain all information of game statistics. Using log differences as the dependent variable is another alternative with a state space evolution. Koopman et al. (2014) adopt stochastically time-varying densities in modeling the Skellam process. Barndorff-Nielsen, Pollard and Shephard (2012) is another example of the Skellam process with different integer valued extensions in the context of high-frequency financial data. Further analysis is required, and this produces a promising area for future research.

3 Example: Everton vs. West Ham (3/5/2016)

We collect the real-time online betting odds data from ladbrokes.com for an EPL game between Everton and West Ham on March 5th, 2016. By collecting real-time online betting data for every 10-minute interval, we can show the evolution of betting market prediction on the final result. We do not account for the overtime for both 1st half and 2nd half of the match and focus on a 90-minute game.

3.1 Implied Skellam probabilities

Table 1 shows the raw data of odds right the game. We need to transform odds data into probabilities. For example, for the outcome 0-0, 11/1 is equivalent to a probability of 1/12.

Table 1: Original odds data from Ladbrokes before the game started.

Everton\ West Ham	0	1	2	3	4	5
0	11/1	12/1	28/1	66/1	200/1	450/1
1	13/2	6/1	14/1	40/1	100/1	350/1
2	7/1	7/1	14/1	40/1	125/1	225/1
3	11/1	11/1	20/1	50/1	125/1	275/1
4	22/1	22/1	40/1	100/1	250/1	500/1
5	50/0	50/1	90/1	150/1	400/1	
6	100/1	100/1	200/1	250/1		
7	250/1	275/1	375/1			
8	325/1	475/1				

Then we can calculate the marginal probability of every score difference from -4 to 5. We neglect those extreme scores with small probabilities and rescale the sum of event probabilities to one.

In Figure 5, the probabilities estimated by the model are compared with the market implied probabilities. As we see, during the course of the game, the Skellam assumption suffices to approximate market expectation of score difference distribution. This set of plots is evidence of goodness-of-fit the Skellam model.

Table 2 shows the model implied probability for the outcome of score differences before the game, compared with the market implied probability. As we see, the Skellam model appears to have longer tails. Different from independent Poisson modeling in Dixon and Coles (1997), our model is more flexible with the correlation between two teams. However, the trade-off of flexibility is that we only know the probability of score difference instead of the exact scores.

Finally, we can plot these probability paths in Figure 6 to examine the behavior of the two teams and represent the market predictions on the final result. Notably, we see the probability change of win/draw/loss for important events during the game: goals scoring and a red card penalty. In such a dramatic game, the winning probability of Everton gets raised to 90% before the first goal of West Ham in 78th minutes. The first two goals scored by West Ham in the space of 3 minutes completely reverses the probability of winning. The probability of draw gets raised to 90% until we see the last-gasp goal of West Ham that decides the game.

3.2 How the market forecast adapts

A natural question arises to how does the market odds (win, lose, draw and actual score) adjust as the game evolves. This is similar to option pricing where Black-Scholes model uses its implied volatility to show how market participants' beliefs change. Our Skellam model mimics its way and shows how the market forecast adapts to changing situations during the game. See Merton (1976) for references of jump models.

Our work builds on Polson and Stern (2015) who define the implied volatility of a NFL game. For an EPL game, we simply define the implied volatility as $\sigma_{IV,t} = \sqrt{\lambda_t^A + \lambda_t^B}$. As the market provides real-time information about λ_t^A and λ_t^B, we can dynamically estimate $\sigma_{IV,t}$ as the game proceeds. Any goal scored is a discrete Poisson shock to the expected score difference (Skellam process) between

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

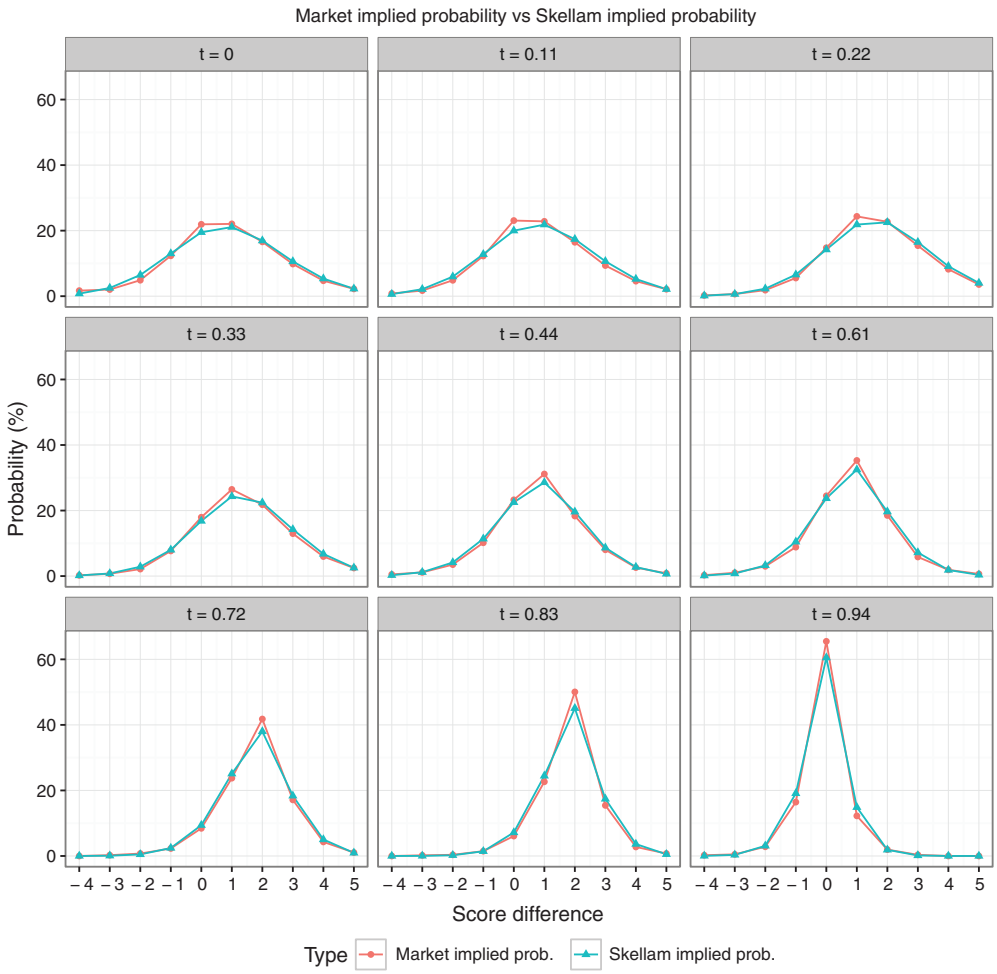

Figure 5: Market implied probabilities versus the probabilities estimated by the model at different time points, using the parameters given in Table 3 .

Table 2: Market implied probabilities for the score differences versus Skellam implied probabilities at different time points. The estimated parameters $\hat{\lambda}^A = 2.33$, $\hat{\lambda}^B = 1.44$.

Score difference	−4	−3	−2	−1	0	1	2	3	4	5
Market prob. (%)	1.70	2.03	4.88	12.33	21.93	22.06	16.58	9.82	4.72	2.23
Skellam prob.(%)	0.78	2.50	6.47	13.02	19.50	21.08	16.96	10.61	5.37	2.27

the teams, and our odds implied volatility measure will be updated.

Figure 6 plots the path of implied volatility throughout the course of the game. Instead of a downward sloping line, we see changes in the implied volatility as critical moments occur in the game. The implied volatility path provides a visualization of the conditional variation of the market prediction for the score difference. For example, when Everton lost a player by a red card penalty at the 34th minute, our estimates $\hat{\lambda}_t^A$ and $\hat{\lambda}_t^B$ change accordingly. There is a jump in implied volatility and our model captures the market expectation adjustment about the game prediction. The change in $\hat{\lambda}_A$ and $\hat{\lambda}_B$ are consistent with the findings of Vecer et al. (2009) where the scoring intensity of the penalized team drops while the scoring intensity of the opposing team increases. When a goal is scored in the 13th minute, we see the increase of $\hat{\lambda}_t^B$ and the market expects that the underdog team is pressing to come back into the game, an effect that has been well-documented in the literature. Another important effect that we observe at the end of the game is that as goals are scored (in the 78th and 81st minutes), the market expectation is that the implied volatility increases again as one might expect.

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

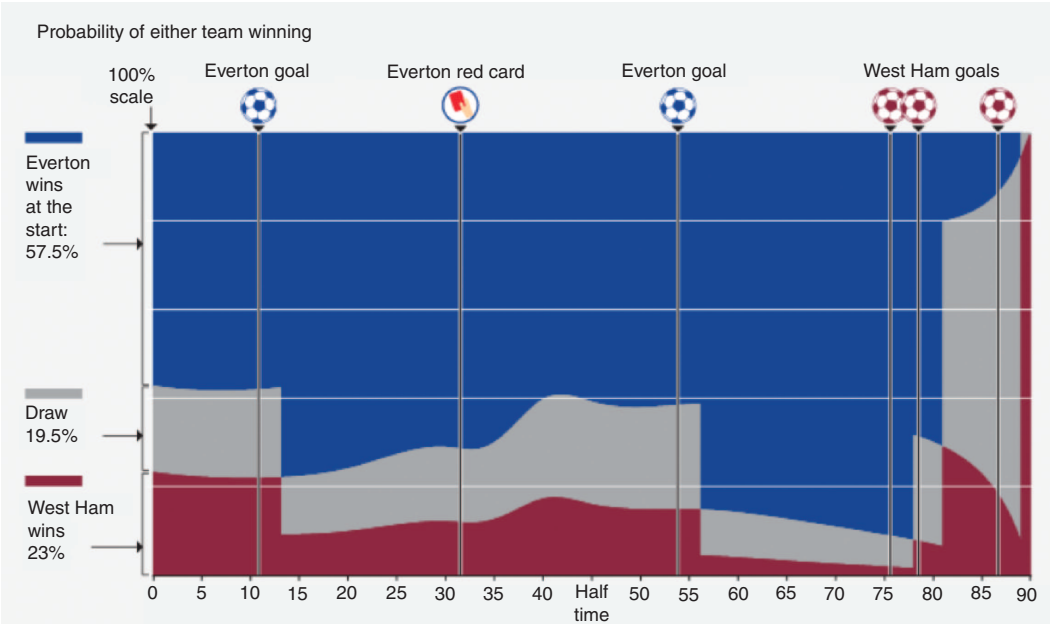

Figure 6: The betting market data for Everton and West Ham is from ladbrokes.com. Market implied probabilities (expressed as percentages) for three different results (Everton wins, West Ham wins and draw) are marked by three distinct colors, which vary dynamically as the game proceeds. The solid black line shows the evolution of the implied volatility (defined in Section 3.2). The dashed line shows significant events in the game, such as goals and red cards. Five goals in this game are 13' Everton, 56' Everton, 78' West Ham, 81' West Ham and 90' West Ham.

Table 3: The calibrated $\{\hat{\lambda}_t^A, \hat{\lambda}_t^B\}$ divided by $(1-t)$ and the implied volatility during the game. $\{\lambda_t^A, \lambda_t^B\}$ are expected goals scored for rest of the game. The less the remaining time, the less likely to score goals. Thus $\{\hat{\lambda}_t^A, \hat{\lambda}_t^B\}$ decrease as t increases to 1. Diving them by $(1-t)$ produces an updated version of $\hat{\lambda}_0$'s for the whole game, which are in general time-varying (but not decreasing necessarily).

t	0	0.11	0.22	0.33	0.44	0.50	0.61	0.72	0.83	0.94	1
$\hat{\lambda}_t^A/(1-t)$	2.33	2.51	2.53	2.46	1.89	1.85	2.12	2.12	2.61	4.61	0
$\hat{\lambda}_t^B/(1-t)$	1.44	1.47	1.59	1.85	2.17	2.17	2.56	2.90	3.67	5.92	0
$(\hat{\lambda}_t^A+\hat{\lambda}_t^B)/(1-t)$	3.78	3.98	4.12	4.31	4.06	4.02	4.68	5.03	6.28	10.52	0
$\sigma_{IV,t}$	1.94	1.88	1.79	1.70	1.50	1.42	1.35	1.18	1.02	0.76	0

Figure 7 compares the updating implied volatility of the game with implied volatilities of fixed $(\lambda^A + \lambda^B)$. At the beginning of the game, the red line (updating implied volatility) is under the "$(\lambda^A + \lambda^B = 4)$"-blue line; while at the end of the game, it's above the "$(\lambda^A + \lambda^B = 8)$"-blue line. As we expect, the value of $(\hat{\lambda}_t^A + \hat{\lambda}_t^B)/(1-t)$ in Table 3 increases throughout the game, implying that the game became more and more intense and the market continuously updates its belief in the odds.

4 Discussion

The goal of our analysis is to provide a probabilistic methodology for calibrating real-time market odds for the evolution of the score difference for a soccer game. Rather than directly using game information, we use the current odds market to calibrate a Skellam model to provide a forecast of the final result. To our knowledge, our study is the first to offer an interpretation of the betting market and to show how it reveals the market expectation of the game result through an implied volatility. One area of future research is studying the index betting. For example, a soccer game includes total goals scored in match and margin of superiority (see Jackson (1994)). The latter is the score difference in our model, and so the Skellam process directly applies.

Our Skellam model is also valid for low-scoring sports such as baseball, hockey or American football with a discrete series of scoring events. For NFL score prediction, Baker and McHale (2013) propose a point process model

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

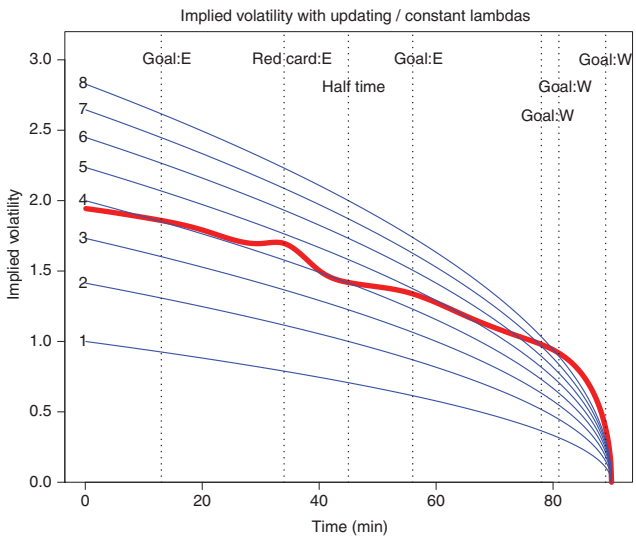

Figure 7: Red line: the path of implied volatility throughout the game, i.e., $\sigma_t^{red} = \sqrt{\hat{\lambda}_t^A + \hat{\lambda}_t^B}$. Blue lines: the path of implied volatility with constant $\lambda^A + \lambda^B$, i.e., $\sigma_t^{blue} = \sqrt{(\lambda^A + \lambda^B) * (1-t)}$. Here $(\lambda^A + \lambda^B) = 1, 2, \ldots, 8$.

that performs as well as the betting market. On the one hand, our model has the advantage of implicitly considering the correlation between goals scored by both teams but on the other hand, ignores the sum of goals scored. For high-scoring sports, such as basketball, the Brownian motion adopted by Stern (1994) is more applicable. Rosenfeld (2012) provides an extension of the model that addresses concerns of non-normality and uses a logistic distribution to estimate the relative contribution of the lead and the remaining advantage. Another avenue for future research, is to extend the Skellam model to allow for the dependent jumpiness of scores which is somewhere in between these two extremes (see Glickman and Stern (1998), Polson and Stern (2015) and Rosenfeld (2012) for further examples.)

Our model allows the researcher to test the inefficiency of EPL sports betting from a statistical arbitrage viewpoint. More importantly, we provide a probabilistic approach for calibrating dynamic market-based information. Camerer (1989) shows that the market odds are not well-calibrated and that an ultimate underdog during a long losing streak is underpriced on the market. Golec and Tamarkin (1991) test the NFL and college betting markets and find bets on underdogs or home teams win more often than bets on favorites or visiting teams. Gray and Gray (1997) examine the in-sample and out-of-sample performance of different NFL betting strategies by the probit model. They find the strategy of betting on home team underdogs averages returns of over 4 percent, over commissions. In summary, a Skellam process appears to fit the dynamics of EPL soccer betting very well and produces a natural lens to view these market efficiency questions.

Acknowledgment: We would like to thank the two referees and the associate editor for their valuable comments on the content of our manuscript and their suggestions for improving the document.

References

Alzaid, A. A. and M. A. Omair. 2010. "On the Poisson Difference Distribution Inference and Applications." *Bulletin of the Malaysian Mathematical Sciences Society* 8(33):17–45.

Avery, C. and J. Chevalier. 1999. "Identifying Investor Sentiment from Price Paths: The Case of Football Betting." *The Journal of Business* 72(4):493–521.

Baker, R. D. and I. G. McHale. 2013. "Forecasting Exact Scores in National Football League Games." *International Journal of Forecasting* 29(1):122–130.

Barndorff-Nielsen, O. E., D. G. Pollard and N. Shephard. 2012. "Integer-valued Levy Processes and Low Latency Financial Econometrics." *Quantitative Finance* 12(4, SI):587–605.

Barndorff-Nielsen, O. E. and N. Shephard. 2012. "Basics of Levy Processes." Technical report, Economics Group, Nuffield College, University of Oxford.

Camerer, C. F. 1989. "Does the Basketball Market Believe in the 'Hot Hand'?" *American Economic Review* 22(4):76–76.

Dixon, M. J. and S. G. Coles. 1997, January. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 46(2):265–280.

Dixon, M. J. and P. F. Pope. 2004. "The Value of Statistical Forecasts in the UK Association Football Betting Market." *International Journal of Forecasting* 20(4):697–711.

Fitt, A. D. 2009. "Markowitz Portfolio Theory for Soccer Spread Betting." *IMA Journal of Management Mathematics* 20(2):167–184.

Fitt, A. D., C. J. Howls and M. Kabelka. 2005. "Valuation of Soccer Spread Bets." *Journal of the Operational Research Society* 57(8):975–985.

Glickman, M. E. and H. S. Stern. 1998. "A State-Space Model for National Football League Scores." *Journal of the American Statistical Association* 93(441):25–35.

Golec, J. and M. Tamarkin. 1991. "The Degree of Inefficiency in the Football Betting Market: Statistical Tests." *Journal of Financial Economics* 30(2):311–323.

Gray, P. K. and S. F. Gray. 1997. "Testing Market Efficiency: Evidence from the NFL Sports Betting Market." *The Journal of Finance* 52(4):1725–1737.

Jackson, D. A. 1994. "Index Betting on Sports." *The Statistician* 43(2):309.

Karlis, D. and I. Ntzoufras. 2003. "Analysis of Sports Data by Using Bivariate Poisson Models." *Journal of the Royal Statistical Society: Series D (The Statistician)* 52(3):381–393.

Karlis, D. and I. Ntzoufras. 2009. "Bayesian Modelling of Football Outcomes: Using the Skellam's Distribution

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM

for the Goal Difference." *IMA Journal of Management Mathematics* 20(2):133–145.

Koopman, S. J., R. Lit and A. Lucas. 2014. "The Dynamic Skellam Model with Applications." *Tinbergen Institute Discussion Paper 14-032/IV/DSF73*.

Lee, A. J. 1997. "Modeling Scores in the Premier League: Is Manchester United Really the Best?" *Chance* 10(1):15–19.

Levitt, S. D. 2004. "Why are Gambling Markets Organized So Differently from Financial Markets?" *Economic Journal* 114(3):223–246.

Maher, M. J. 1982. "Modelling Association Football Scores." *Statistica Neerlandica* 36(3):109–118.

Merton, R. C. 1976. "Option Pricing When Underlying Stock Returns are Discontinuous." *Journal of Financial Economics* 3(1–2):125–144.

Polson, N. G. and H. S. Stern. 2015. "The Implied Volatility of a Sports Game." *Journal of Quantitative Analysis in Sports* 11(2):145–153.

Rosenfeld, J. W. 2012. "An In-game Win Probability Model of the NBA." Thesis, Harvard University.

Sellers, K. F. 2012. "A Distribution Describing Differences in Count Data Containing Common Dispersion Levels." *Advances and Applications in Statistical Sciences* 7(3):35–46.

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.

Stern, H. S. 1994. "A Brownian Motion Model for the Progress of Sports Scores." *Journal of the American Statistical Association* 89(427):1128–1134.

Vecer, J., F. Kopriva and T. Ichiba. 2009. "Estimating the Effect of the Red Card in Soccer: When to Commit an Offense in Exchange for Preventing a Goal Opportunity." *Journal of Quantitative Analysis in Sports* 5(1):1–20.

Yates, J. F. 1982. "External Correspondence: Decompositions of the Mean Probability Score." *Organizational Behavior and Human Performance* 30(1):132–156.

Brought to you by | University of Chicago
Authenticated
Download Date | 10/25/17 7:46 PM