Physica A 547 (2020) 124411

Contents lists available at ScienceDirect

Physica A

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

Making real-time predictions for NBA basketball games by combining the historical data and bookmaker's betting line

Kai Song [a,b], Yiran Gao [c], Jian Shi [a,b,*]

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*
[b] *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China*
[c] *Beijing StatusWin Lottery Operations Technology Ltd., Beijing, China*

ARTICLE INFO

Article history:
Received 16 October 2019
Received in revised form 4 February 2020
Available online 28 February 2020

Keywords:
Sports science
NBA
In-play predictions
Gamma process
The total points process

ABSTRACT

The paper presents a gamma process based model for the total points processes of NBA basketball matches. This model obtains a useful formula for the in-play prediction. What is more, we employ the bookmaker's betting line to adjust the original gamma process model. The out-of-sample forecasting performances are evaluated, and more profoundly, this model can produce a positive return on the over–under betting market. Besides, our model has an application in monitoring the betting market, which may be useful to bettors.

1. Introduction

Basketball is one of the most popular sports in the world and the National Basketball Association (NBA) is the premier professional basketball league in the United States. Many issues related to NBA basketball matches have been studied, for example, player assessment and ranking [1], the teams' persistent behavior [2,3], the importance of the franchise player and the hot-hand effect [4] and outcomes prediction [5,6]. Forecasting the outcomes of future matches has been being investigated extensively by researchers and different statistical models have been proposed. Nevertheless, a majority of papers deal with pre-match forecasting [7–11]. In this paper, we aim to develop a model which can provide in-play predictions when a match is underway.

In-play models forecast the final outcomes conditional on the current information when a match is in progress. Play-by-play data record the in-match scoring events in detail. Therefore, there are several papers that focus on modeling the progression of a match via play-by play data [12–15]. Owing to the fast-paced character of basketball contest, scoring events happen frequently. Consequently, it is difficult to access and process the play-by-play data for models to make predictions in real time, which limits their applications. In general, scores are recorded at several discrete moments without the in-match events. Chen and Fan [16] investigated the score difference process via functional data analysis based on the data generated every thirty seconds. However, forecasting the outcomes of future matches is not addressed in this paper. Kayhan and Watkins [17] developed a data snapshot method to produce real-time winning probabilities for NBA games, which is essentially based on the empirical frequency. Stern [18] modeled the score difference process using the Brownian motion model and developed a simple formula for the in-play win probability. Since then, Glasson [19] and Ryall [20] made improvements to the Stern's model using the bookmaker's spread betting lines and Elo ratings respectively. Following a similar research thinking, we model the total points process as a gamma process in this paper and then obtain a useful formula for in-play predictions. After that, we improve the original gamma process model by means of the bookmaker's betting line.

[*] Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China.
E-mail addresses: kaisong@amss.ac.cn (K. Song), yr.gao@8win.com (Y. Gao), jshi@iss.ac.cn (J. Shi).

https://doi.org/10.1016/j.physa.2020.124411
0378-4371/

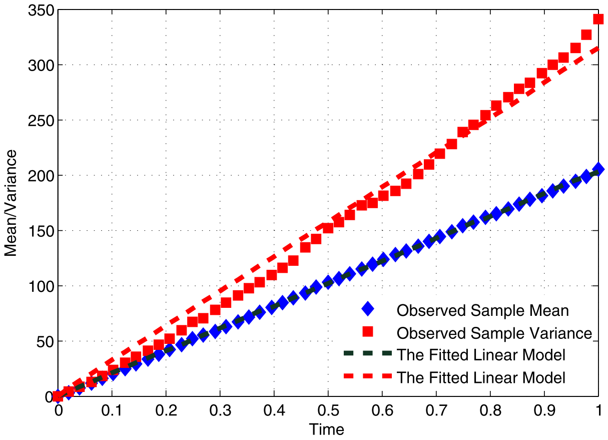


Fig. 1. The sample mean and variance of total points graphed over time.

The gamma process is widely used in degradation modeling, where deterioration is assumed to accumulate gradually over time in a sequence of tiny increments, see [21] for more details. The gamma process $\{X(t), t \geq 0\}$ is defined as the random process satisfying: (1) $X(t) = 0$ with probability one; (2) $X(s) - X(t) \sim Gamma(\lambda, \eta(s) - \eta(t))$ for all $s > t \geq 0$, where $Gamma(\cdot, \cdot)$ is the gamma distribution and $\eta(t)$ is a non-decreasing and right-continuous function with $\eta(0) \equiv 0$; (3) $X(t)$ has independent increments. In this case, $X(t)$ follows $Gamma(\lambda, \eta(t))$ with scale parameter λ and shape function $\eta(t)$ and its probability density function is written as

$$f(x; \lambda, \eta(t)) = \frac{\lambda^{\eta(t)}}{\Gamma(\eta(t))} e^{-\lambda x} x^{\eta(t)-1} I(x \geq 0), \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function. Furthermore, the expectation and variance of $X(t)$ are

$$E[X(t)] = \frac{\eta(t)}{\lambda} \text{ and } Var[X(t)] = \frac{\eta(t)}{\lambda^2}. \tag{2}$$

Note that the total points process is nonnegative, non-decreasing and nearly continuous and the total points increase over time due to a series of scoring events, therefore, the gamma process is considered to be a good option to depict the total points process.

To our knowledge, there is no paper addressing the problem of modeling the total points process. In this paper, we develop a gamma process based model for the total points process. The rest of this paper is structured as follows. In Section 2, we describe the data used here and present an initial study of the total points process. In Section 3, a gamma process based model is proposed, and thus obtaining a formula for the in-play prediction. In Section 4, we conduct an empirical study to evaluate the proposed model and then illustrate its application. Finally, Section 5 concludes.

2. Data

NBA regular season data from 2015–2016 to 2017–2018 are collected, where the total points processes are recorded every minute. The overtimes, which are added to a match ended in a tie so that one team can win the match, are counted as part of the 48th minute. Besides, the in-play betting odds on the over–under betting market from the bookmaker Bet365 are also available. All these data are downloaded from www.nowgoal.com/nba.htm website. More explicitly, our data are formed in the following way: for a match k at time t, we define

$$Y_k(t) = HomeTeamScore_k(t) + AwayTeamScore_k(t), \tag{3}$$

where the 48-minute regulation period is transformed to the unit interval, i.e., $t = 0, 1/48, \ldots, 1$.

Before establishing our specific model, we consider the features of the data by examining the 2015–2016 season's data for illustration. The sample mean and variance of total points are plotted against the time in Fig. 1, as well as the corresponding fitted linear models, where we can see that the trends are nearly linear. Therefore, we would be inclined to think that the total points process can be well described by a gamma process with a linear shape function. We fit a gamma process model to the total points process and then plot the sample quantiles against the quantiles of the fitted gamma process at the end of each quarter respectively, which is shown in Fig. 2. To our satisfaction, the points lie near the y = x line roughly, which demonstrates that modeling the total points process using the gamma process with a linear shape function is feasible. The results at the end of the fourth quarter show some divergence, of which one possible reason is that the matches with overtimes have an influence on the distribution of the final total points.

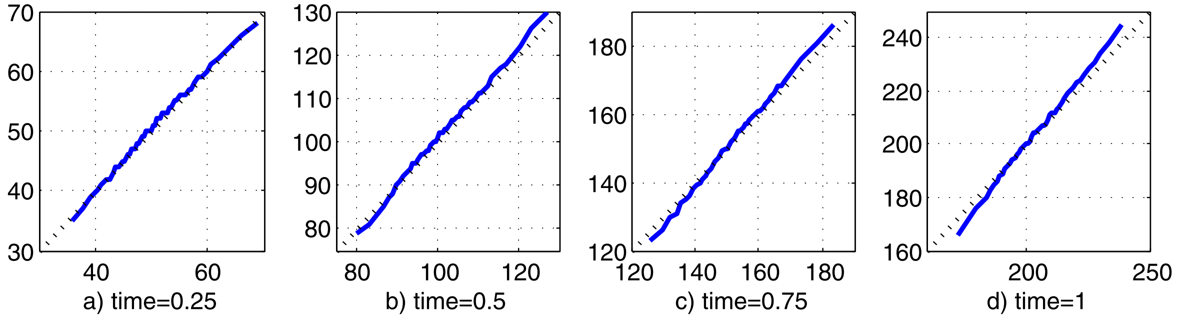


Fig. 2. The sample quantiles plotted against the quantiles of the fitted gamma process. The dashed diagonal is the y = x line.

3. Methodology

3.1. The gamma process model

Let $Y(t)$ denote the total points scored by the home and away teams at time t, where t refers to the fraction of a match having been completed. Then, the gamma process model can be written as:

$$Y(t) \sim Gamma(\lambda, \eta t) \tag{4}$$

and $Y(s) - Y(t)$, $s > t$, is independent of $Y(t)$ with

$$Y(s) - Y(t) \sim Gamma(\lambda, \eta(s-t)). \tag{5}$$

Under the gamma process model, the probability that the final total points are greater than a threshold τ [i.e., $Y(1) > \tau$] given the current total points at time t [i.e., Y(t)=h] is

$$\begin{aligned} P\{Y(1) > \tau \mid Y(t) = h\} &= P\{Y(1) - Y(t) > \tau - h\} \\ &= \int_{\tau-h}^{+\infty} \frac{\lambda^{\eta(1-t)}}{\Gamma(\eta(1-t))} e^{-\lambda x} x^{\eta(1-t)-1} dx \\ &= \int_{\lambda(\tau-h)}^{+\infty} \frac{1}{\Gamma(\eta(1-t))} e^{-x} x^{\eta(1-t)-1} dx \\ &\triangleq Gammainc(\lambda(\tau-h), \eta(1-t)), \end{aligned} \tag{6}$$

where $\Gamma(\cdot)$ and $Gammainc(\cdot,\cdot)$ denote the gamma function and the upper incomplete gamma function respectively. The model parameters are estimated using the method of moments and the details are given in the appendix.

3.2. The adjusted gamma process model

Although the gamma process model provides a simple and useful formula for the conditional probability in (6), it does not take account of differences between the matches, especially the difference in ability between the teams. As we know, the betting lines provided by the bookmaker imply prior beliefs of difference between the matches. Then, we extend the gamma process model to incorporate this information.

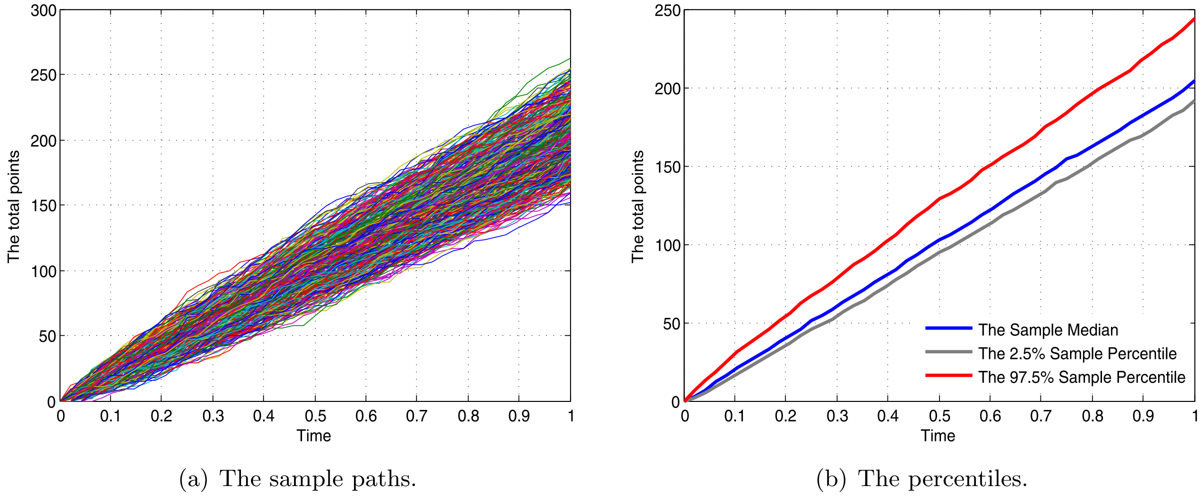
More specifically, for a particular match k, let b_{k0} denote the pre-match total points betting line. b_{k0} is an initial estimate of the final total points according to the bookmaker. Therefore, we assume that the expectation of the final total points equals b_{k0}, i.e,

$$E[Y_k(1)] = \frac{\eta}{\lambda} = b_{k0}. \tag{7}$$

Once the model parameters are estimated, the scale parameter $\hat{\lambda}$ is fixed. We then obtain a specific estimation $\hat{\eta}_k = \hat{\lambda} \cdot b_{k0}$ for the match k. In this case, the matches are considered differently, i.e., $Y_k(t) \sim Gamma(\lambda, \eta_k t)$. For illustration, we plot the total points processes of different matches of 2015–2016 season against the time in Fig. 3. We can see that the paths vary from match to match, which demonstrates that the adjusted model does make sense. Besides, we also plot the sample median and percentiles in Fig. 3 to have a better intuition what scores are more likely.

4. Empirical study

For a regular season, we fit the model according to the first half of matches, and then make out-of-sample predictions for the remaining matches. To evaluate the in-play forecasting performance, the predictions are made every minute. We repeat this procedure for each regular season separately.



(a) The sample paths.

(b) The percentiles.

Fig. 3. Graphic display of the total points process.

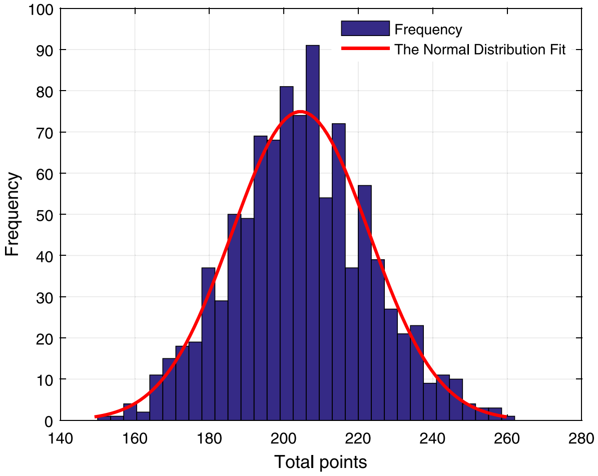


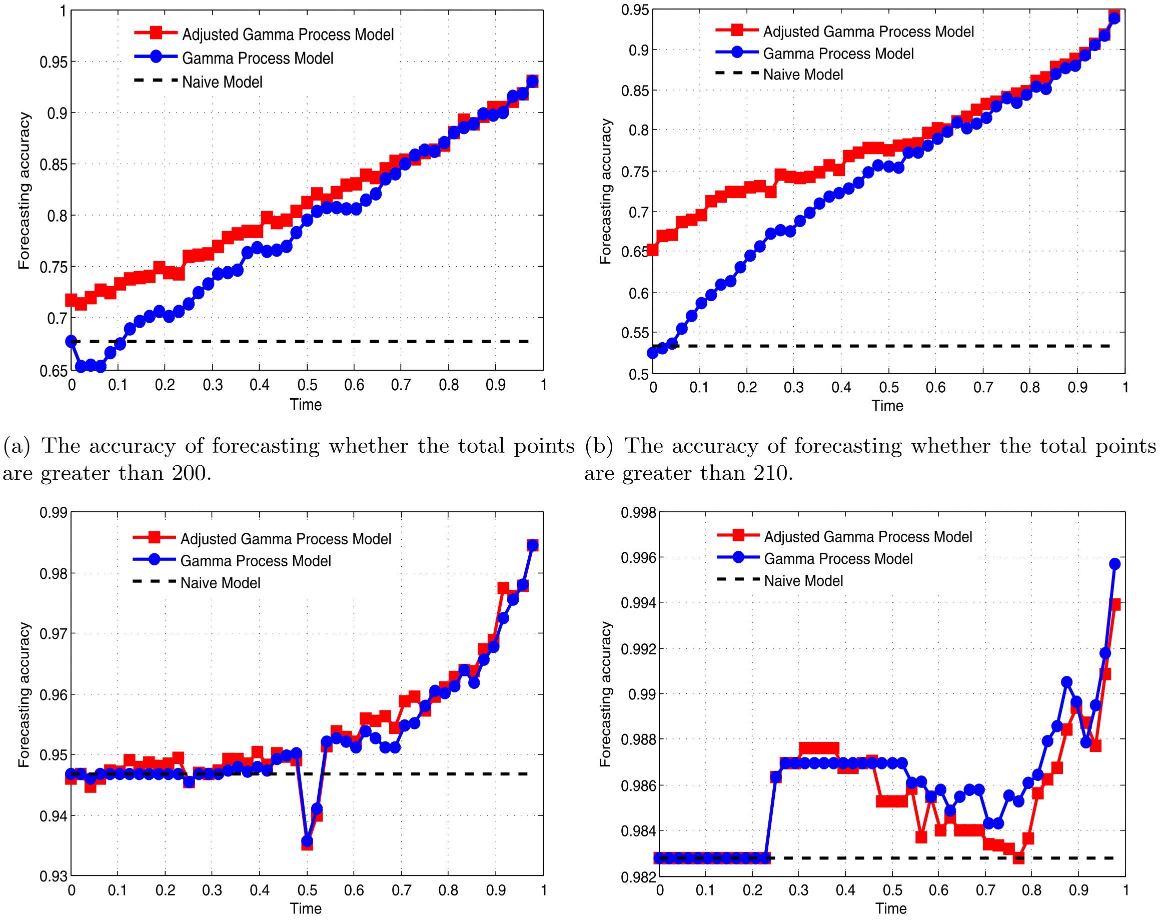
Fig. 4. Histogram of the final total points.

4.1. Forecasting accuracy performance

In this section, we evaluate our model regarding the forecasting accuracy which is the percentage of game results being predicted correctly. Here, a naive model based on the empirical frequency is used for comparison. Intuitively, we present a histogram of the final total points as well as the normal distribution fit in Fig. 4, taking the 2015–2016 season as an example, which shows that the normal distribution could provide an acceptable model fit visually. More specifically, let $p_{k,\tau}$ be the empirical frequency of games which ended with more than τ points prior to the match k. If $p_{k,\tau}$ is larger that 0.5, we predict that more points would be scored in the match k, and denote $\delta_{k,\tau} = 1$. Otherwise, we predict that less points would be scored in the match k, and denote $\delta_{k,\tau} = 0$. Then, the naive forecasting accuracy, with n games in the test set, can be calculated as follows:

$$\frac{1}{n}\sum_{k=1}^{n} I\{\delta_{k,\tau} = w_{k,\tau}\}, \tag{8}$$

where $w_{k,\tau}$ is the actual outcome of the match k (1 if the total points does exceed τ, otherwise it equals 0) and $I\{\cdot\}$ is the indicator function. For illustration, the regular game results we are mainly interested in are: whether the final total points are greater than 200 and 210. Besides, predicting the relatively improbable results, for example, whether the total points exceeding 240 or being smaller than 170, are also investigated. Nevertheless, it is worth noting that predicting the regular games precisely plays an more important role in obtaining positive returns in the betting market. Fig. 5 shows the forecasting results averaged across all three seasons. Form Figs. 5(a) and 5(b), we can see that the trends of both models



(a) The accuracy of forecasting whether the total points are greater than 200.

(b) The accuracy of forecasting whether the total points are greater than 210.

(c) The accuracy of forecasting whether the total points are greater than 240.

(d) The accuracy of forecasting whether the total points are smaller than 170.

Fig. 5. Out-of-sample forecasting performance evaluation.

are increasing, which may be explained by the fact that we can observe more information as time goes on, resulting in predicting the final outcomes more accurately. Besides, the gamma process model adjusted by the bookmaker's betting line has a superior performance. On the other hand, for the relatively improbable events, the models tend to predict correctly, and thus have higher forecasting accuracy. Furthermore, the models have no much difference regarding the forecasting accuracy. What is more, the gamma process model and the adjusted gamma process model outperform the naive model based on the empirical frequency overall.

4.2. Betting performance

Here, we combine the adjusted gamma process model with a betting strategy to bet with the market to assess our proposed model. To begin with, we introduce the over–under betting market briefly. The bookmaker publishes the odds for the over, the odds for the under and the total points betting line which are updated as the match goes on. Then, bettors can wager throughout the match as long as the odds and the total points betting line are available at that time. One can bet on more or less than the published total points betting line. If a bettor bets on the over, he or she wins if the final total points are greater than the published betting line. If a bettor bets on the under, he or she wins if the final total points are less than the published betting line. In this case, the successful bettor can obtain a net profit which equals the corresponding odds minus one for one unit bet. If the final total points equal the published betting line, the bet is canceled and the wager is refunded [22,23]. One basic betting strategy is based on the expected value of one unit bet, which is given by

$$EV(A) = P(A) * Odds(A) - 1, \tag{9}$$

Table 1
Summary of betting results.

Season	Total bet	Net profit	Rate of return (%)
2015–2016	9429	984.64	10.44
2016–2017	11297	970.44	8.59
2017–2018	12160	1629.44	13.40
Average	32886	3584.52	10.90

Table 2
Summary of the descriptive statistics analysis.

Variable	Mean	Standard deviation	Skewness	Kurtosis
Home teams' scores	103.5737	11.5867	0.0726	3.0297
Away teams' scores	100.9202	11.1673	0.1415	2.8771
Winning teams' scores	107.6222	10.3169	0.1550	3.0140
Losing team's scores	96.8717	9.9096	0.0462	2.8645

where the event A is the over or the under; $P(A)$ is our estimated probability; $Odds(A)$ is the bookmaker's odds [24,25]. To reduce risk, we only bet when $EV(A)$ is greater than zero. We make bets every minute and then calculate the average rate of return. The matches whose in-play odds are not available at the corresponding time are not considered in the current bet. Table 1 displays the betting results, including the total bet, net profit, and rate of return. We can obtain positive returns on the over–under betting market, which is a testament to the effectiveness of our proposed model.

4.3. Application in monitoring the betting market

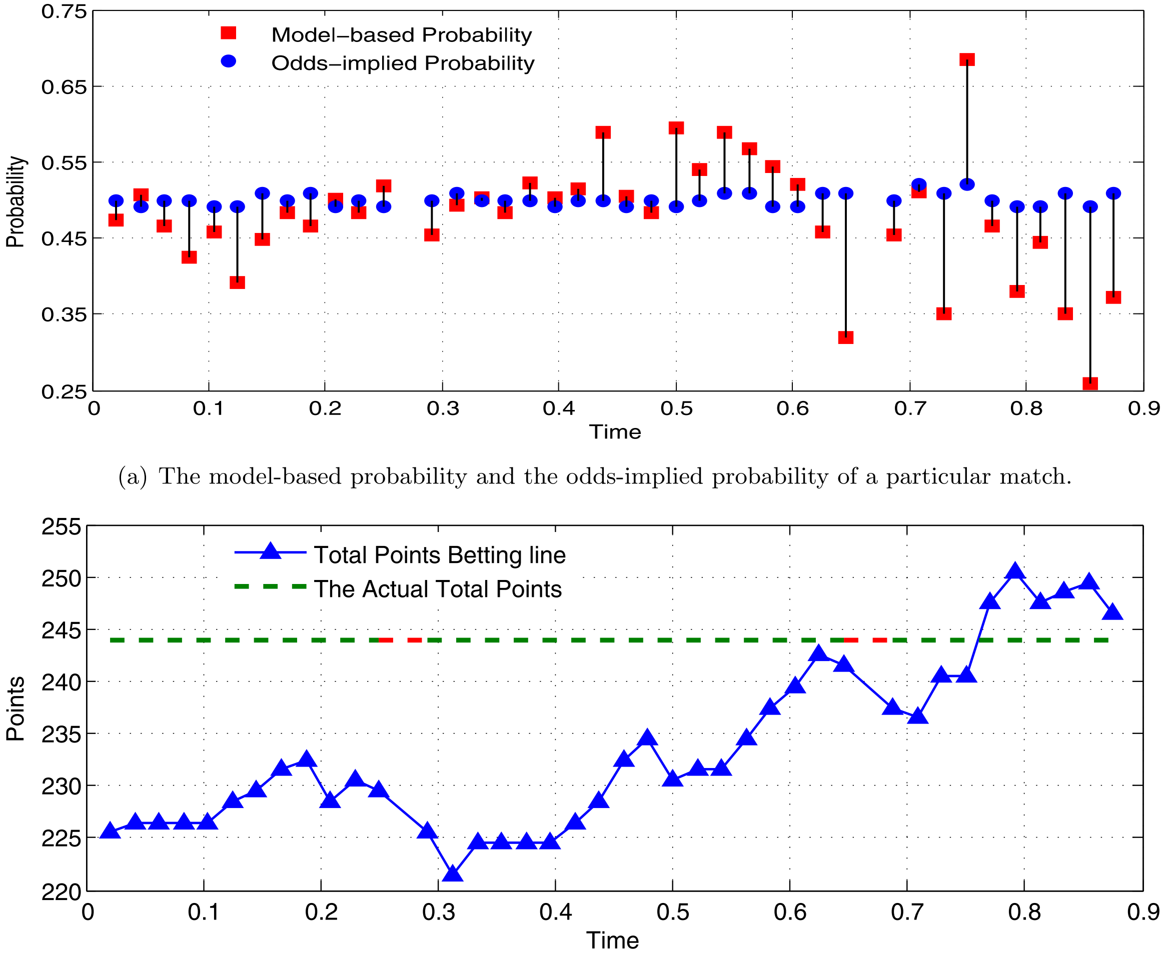
Bettors are very interested in the abnormal betting activity, mainly reflected by the volume of betting, because insiders have information unknown to others. When there is abnormal behavior in the market, the bookmaker begins to modify the odds to avoid the exposure risk [26]. However, the statistical model does not react to this information. Therefore, we compare the model-based probability with the odds-implied probability in real time and suspicions arise where the two kinds of probabilities diverge. See the appendix for how to calculate the odds-implied probability. Fig. 6 shows the model-based probability and the odds-implied probability of a particular match, as well as the in-play betting lines. Take the 41th minute as an example, the betting line is greater than the actual total points, therefore the model-based probability that the total points being above the betting line is small, which is a reasonable result. However, the odds-implied probability deviates from this plausible result. Thus, we have grounds to suspect that there is abnormal betting activity. It is worth noting that the bookmaker sometimes has addition information such as the player injury, which also prompts the bookmaker to adjust the odds. In this case, deviation of the model-based probability from that implied by the odds should be assessed further.

5. Conclusion and discussion

This paper treats the problem of modeling the total points processes of NBA basketball matches. After an initial analysis of the observed data, we employ the gamma process with a linear shape function to describe the total points process. At the same time, a useful formula for in-play predictions is developed under the gamma process model. To develop a match-specific model, the original gamma process model is adjusted using the bookmaker's betting line. And the empirical study demonstrates that this modification does make sense. What is more, our model can have a positive return on the over–under betting market. In addition, the model can also be used to monitor the abnormal betting behavior. Lastly, we note that this model can be used to model other high-scoring sports, for example Australian Rules football.

The main modeled quantity of this paper is the total points process. While there is some information loss, modeling the total points is also interesting and worthwhile, and it has a direct application in the over–under betting market. What is more, we can make use of the information from the over–under betting market to improve the forecasting accuracy in this framework. As a side note, it is found that the gamma process is also applicable to the separate home and away scoring processes. In addition, using the NBA data, some potentially interesting questions can be investigated, for example, the differences in distributions of the home/away teams' scores and in particular of the winning/losing teams' scores. Here, we present some results of descriptive statistics in Table 2. It shows that the home team outscores the away team by approximately 2.65 points, while the winning team has a big advantage of 10.75 points over the losing team. The performance of losing teams is relatively stable, having the lowest standard deviation. All the scores could be depicted by the normal distribution according to the values of skewness and kurtosis. These results might reflect potential influences of the situation on teams' behaviors. Further work is outside the scope of the current paper, but is under progress and we expect to present some useful findings in the future.

It is also noteworthy that there are several papers [27,28] which use the random process, like the Poisson process, to investigate the scoring in basketball. However, they treat the one-point shots, two-point shots and three-point shots



(a) The model-based probability and the odds-implied probability of a particular match.

(b) The corresponding in-play betting lines and the final total points.

Fig. 6. A real example of monitoring the betting market.

without distinction, and any kind of point scored by either team is considered an event. The elapsed time between events are the main focus in these papers, and the Weibull distribution, the Lognormal distribution and the power-law distribution are more often used to model its distribution. Essentially, the number of events over time is a counting process. However, our paper aims at the total points process, and it is not a counting process because the kind of point scored (1, 2, or 3) should be considered. The total points scored in a time interval involve both the number of events and the kind of point scored.

We take account of differences between the matches by using the betting line to adjust the shape parameter. The variability between teams also influence a match's outcome, thus, how to incorporate it into the model is notable. Maybe the random effect model is a viable option.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank the editor and anonymous referees for their constructive comments and suggestions which have considerably improved this paper. We also thank the Beijing StatusWin Lottery Operations Technology Ltd for providing the data.

Appendix A. The model parameters estimation

Let $y_k(t)$ be the observed total points process for the match k. Note that

$$E[Y_k(\frac{j}{48}) - Y_k(\frac{j-1}{48})] = \frac{\eta}{48\lambda},\ Var[Y_k(\frac{j}{48}) - Y_k(\frac{j-1}{48})] = \frac{\eta}{48\lambda^2}. \tag{A.1}$$

Then, we estimate the parameters as follows:

$$\hat{\lambda} = \frac{\bar{y}}{V},\ \hat{\eta} = 48 \cdot \hat{\lambda} \cdot \bar{y}, \tag{A.2}$$

where $\bar{y} = \frac{1}{48N}\sum_{k=1}^{N}\sum_{j=1}^{48}[y_k(\frac{j}{48}) - y_k(\frac{j-1}{48})]$ and $V = \frac{1}{48N}\sum_{k=1}^{N}\sum_{j=1}^{48}[y_k(\frac{j}{48}) - y_k(\frac{j-1}{48}) - \bar{y}]^2$; N is the number of matches in the training set.

Appendix B. The odds-implied probability

The implied probability can be estimated by

$$P_{over} = \frac{1/O}{1/O + 1/U},\ P_{under} = \frac{1/U}{1/O + 1/U} \tag{B.1}$$

where O and U are the betting odds for the over and the under respectively [29].

References

[1] W.W. Cooper, J.L. Ruiz, I. Sirvent, Selecting non-zero weights to evaluate effectiveness of basketball players with DEA, European J. Oper. Res. 195 (2) (2009) 563–574.
[2] P. Ferreira, What detrended fluctuation analysis can tell us about NBA results, Physica A 500 (2018) 92–96.
[3] A. Kononovicius, Illusion of persistence in NBA 1995–2018 regular season data, Physica A 520 (2019) 250–256.
[4] M. Oldham, A.T. Crooks, Drafting agent-based modeling into basketball analytics, in: Spring Simulation Conference, SpringSim, 2019, pp. 1–12.
[5] K. Song, Q. Zou, J. Shi, Modelling the scores and performance statistics of NBA basketball games, Comm. Statist. Simulation Comput. (2018) http://dx.doi.org/10.1080/03610918.2018.1520878.
[6] W. Cai, D. Yu, Z. Wu, X. Du, T. Zhou, A hybrid ensemble learning framework for basketball outcomes prediction, Physica A 528 (2019) 121461.
[7] R.T. Stefani, Football and basketball predictions using least squares, IEEE Trans. Syst. Man Cybern. 7 (2) (1977) 117–121.
[8] R.T. Stefani, Improved least squares football, basketball, and soccer predictions, IEEE Trans. Syst. Man Cybern. 10 (2) (1980) 116–123.
[9] T. Baghal, et al., Are the "four factors" indicators of one factor? An application of structural equation modeling methodology to NBA data in prediction of winning percentage, J. Quant. Anal. Sports 8 (1) (2012) 1–17.
[10] E.S. Jones, Predicting outcomes of NBA basketball games (Ph.D. thesis), North Dakota State University, 2016.
[11] M. Hans, Modeling and forecasting the outcomes of NBA basketball games, J. Quant. Anal. Sports 12 (1) (2016) 31–41.
[12] E. Štrumbelj, P. Vračar, Simulating a basketball match with a homogeneous Markov model and forecasting the outcome, Int. J. Forecast. 28 (2) (2012) 532–542.
[13] A. Gabel, S. Redner, Random walk picture of basketball scoring, J. Quant. Anal. Sports 8 (1) (2012) 1–20.
[14] P. Vračar, E. Štrumbelj, I. Kononenko, Modeling basketball play-by-play data, Expert Syst. Appl. 44 (C) (2016) 58–66.
[15] S. Merritt, A. Clauset, Scoring dynamics across professional team sports: tempo, balance and predictability, EPJ Data Sci. 3 (1) (2014) 1–21.
[16] T. Chen, Q. Fan, A functional data approach to model score difference process in professional basketball games, J. Appl. Stat. 45 (1) (2018) 112–127.
[17] O.V. Kayhan, W. Alison, A data snapshot approach for making real-time predictions in basketball, Big Data 6 (2) (2018) 96–112.
[18] H.S. Stern, A brownian motion model for the progress of sports scores, J. Amer. Statist. Assoc. 89 (427) (1994) 1128–1134.
[19] S. Glasson, A Brownian motion model for the progress of Australian Rules football scores, in: Eighth Australasian Conference on Mathematics and Computers in Sport, 2006, pp. 216–225.
[20] R. Ryall, Predicting outcomes in Australian rules football (Ph.D. thesis), Royal Melbourne Institute of Technology University, 2011.
[21] J.M.V. Noortwijk, A survey of the application of gamma processes in maintenance, Reliab. Eng. Syst. Saf. 94 (1) (2009) 2–21.
[22] BeaBet, 2019, https://www.beabetterbettor.com/nba/odds/, (Accessed 24 December 2019).
[23] TheLines, 2019, https://www.thelines.com/betting/in-play/, (Accessed 24 December 2019).
[24] G. Boshnakov, T. Kharrat, I.G. Mchale, A bivariate Weibull count model for forecasting association football scores, Int. J. Forecast. 33 (2) (2017) 458–466.
[25] R. Ryall, A. Bedford, An optimized ratings-based model for forecasting Australian rules football, Int. J. Forecast. 26 (3) (2010) 511–517.
[26] D. Forrest, I.G. McHale, Using statistics to detect match fixing in sport, IMA J. Manag. Math. 30 (4) (2019) 431–449.
[27] J.M. Martín-González, Y.D.S. Guerra, J.M. García-Manso, E. Arriaza, T. Valverde-Estévez, The Poisson model limits in NBA basketball: Complexity in team sports, Physica A 464 (2016) 182–190.
[28] Y.D.S. Guerra, J.M. Martín-González, S.S. Montesdeoca, D.R. Ruiz, N.A. López, J.M. García-Manso, Basketball scoring in NBA games: An example of complexity, J. Syst. Sci. Complexity 26 (1) (2013) 94–103.
[29] E. Štrumbelj, On determining probability forecasts from betting odds, Int. J. Forecast. 30 (4) (2014) 934–943.