Submitted to the Annals of Applied Statistics

IN-GAME WIN PROBABILITIES FOR THE NATIONAL RUGBY LEAGUE

BY TIANYU GUAN[1,*], ROBERT NGUYEN[2], JIGUO CAO[1,†] AND TIM SWARTZ[1,‡]

[1]*Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6* [*]*tianyug@sfu.ca;* [†]*jiguo_cao@sfu.ca;* [‡]*tim@stat.sfu.ca*

[2]*Department of Statistics, School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia robert.nguyen@unsw.edu.au*

This paper develops new methods for providing instantaneous in-game win probabilities for the National Rugby League. Besides the score differential, betting odds and real-time features extracted from the match event data are also used as inputs to inform the in-game win probabilities. Rugby matches evolve continuously in time and the circumstances change over the duration of the match. Therefore, the match data are considered as functional data, and the in-game win probability is a function of the time of the match. We express the in-game win probability using a conditional probability formulation, the components of which are evaluated from the perspective of functional data analysis. Specifically, we model the score differential process and functional feature extracted from the match event data as sums of mean functions and noises. The mean functions are approximated by B-spline basis expansions with functional parameters. Since each match is conditional on a unique kickoff win probability of the home team obtained from the betting odds (i.e. the functional data are not independent and identically distributed), we propose a weighted least squares method to estimate the functional parameters by borrowing the information from matches with similar kickoff win probabilities. The variance and covariance elements are obtained by the maximum likelihood estimation method. The proposed method is applicable to other sports when suitable match event data are available.

1. Introduction. In recent years, analytics have made a profound impact on sports where great investments have been made in the "big" professional sports of basketball (the National Basketball Association), football (the National Football league), soccer (major European leagues), hockey (the National Hockey League), and baseball (Major League Baseball). Many teams now have their own analytics staff where decisions are scrutinized across many areas of the sporting operation including strategy, drafting, salaries, player evaluation, and marketing. For a survey of some of the work that has been done in sports analytics, see Albert, Glickman, Swartz, and Koning (2017).

Whereas the National Rugby League (NRL) may be considered a big sport (it has the greatest television viewership of any sport in Australia), the NRL is underrepresented in the sports analytics literature. For example, in a search of the archives of the *Journal of Quantitative Analysis in Sports* (founded in 2005), the authors were unable to find a single article devoted to the rugby league. Similarly, in a search of *Australian & New Zealand Journal of Statistics* we were only able to find a single article devoted to the rugby league (Lee (1999)). However, there have been many papers written on the rugby league from the sports

[*] T. Guan is a PhD candidate, and J. Cao and T. Swartz are Professors, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. R. Nguyen is a Ph.D. candidate, Department of Statistics, School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia. Cao and Swartz have been partially supported by the Natural Sciences and Engineering Research Council of Canada. This work was initiated while Nguyen visited the Department of Statistics and Actuarial Science at Simon Fraser University.

Keywords and phrases: In-game win probability, event data, NRL, functional data analysis, model validation

science perspective and a small sample of these include Glassbrook, Doyle, Alderson, and Fuller (2019), Booth and Orr (2017), Windt, Gabbett, Ferris, and Khan (2017), Seitz, Rivière, de Villarreal, and Haff (2014), King, Jenkins, and Gabbett (2009), and Gabbett (2005).

In an attempt to grow the game, the NRL is adding an analytics focus to the sport (see www.nrl.com.stats). In particular, to provide additional excitement to the television viewing experience, the NRL would like to include in-game win probabilities. The idea is that such a graphic may be presented in a small corner of the screen, and be continually updated as the game circumstances change. The graphic would be appealing to the NRL fan base and also to punters. The continual update precludes highly computational techniques, and of course, the predictions of the in-game win probabilities ought to be accurate.

Prediction of the in-game win probabilities has been investigated in other major sports, such as basketball, football, and hockey. For example, in basketball, Stern (1994) developed a Brownian motion model to investigate the score differential process. Gable and Redner (2012) and Clauset, Kogan and Redner (2015) built computational random walk models for analyzing the scoring processes in basketball games. Štrumbelj and Vračar (2012) and Vračar, Štrumbelj and Kononenko (2016) used possession-based Markov models to simulate basketball matches. Data snapshots approaches were developed by Kayhan and Watkins (2018, 2019). Song, Gao and Shi (2020) obtained in-game predictions by fitting a gamma process based model for the total points process of basketball. A multiresolution stochastic process model was proposed by Cervone, D'Amour, Bornn, and Goldsberry (2016) to quantify the expected possession value, a concept that is similar to the in-game win probability. In football, Lock and Nettleton (2014) employed a random forest method to provide the in-game win probability of the National Football League (NFL), whereas Robberechts, Van Haaren and Davis (2019) introduced a Bayesian statistical model. For the National Hockey League (NHL), Buttrey, Washburn, and Price (2011) proposed to predict the scoring process by fitting a Poisson process and Pettigrew (2015) used the in-game win probabilities to assess the offensive productivity of the NHL players. However, the in-game win probability is a new concept in rugby. In this article, we focus on the NRL matches and propose methods to estimate the in-game win probabilities from the perspective of functional data analysis (FDA), which uses the information of the score differential, the features extracted from the in-game event data and a unique kickoff win probability of the home team for each match.

The NRL has provided us with four seasons of detailed event data (2016-2019) which we use to inform the in-game win probabilities. Our approach begins with a conditional probability formulation where our main interest concerns the evaluation of the in-game posterior win probability. Specifically, the in-game win probability is expressed by a conditional probability formulation with components of a unique kickoff win probability of the home team and conditional joint densities of the score differential and event feature that arises from the in-game event data conditional on the event that the home team wins or losses the match and the unique kickoff win probability of the home team. The challenge is the development of an accurate model for which the posterior probability can be evaluated in real time. The accuracy provided by the model relies on the domain knowledge of the sport; hence we search for data and covariates that have high predictive capability.

A rugby league match is 80 minutes in duration and that circumstances change over the duration of the match. Therefore, we consider the match data as functional data and the in-game win probability is a function of the time of the match. The distributions that are specified in our model are determined via FDA. FDA is a relatively new branch of statistics where regression methods are extended to the study of curves or functions. There is an extensive literature on FDA. The most popular techniques in FDA include various smoothing methods (e.g. Ramsay and Silverman (2005), Chapter 3, de Boor (2001) and Wand and Jones (1995)), functional principal component analysis (e.g., Besse and Ramsay (1986), Bosq (2000), Cardot (2000),

and Yao, Müller and Wang (2005a)), functional linear regression model (Hastie and Mallows (1993), Hall and Horowitz (2007), Cardot, Ferraty and Sarda (2003), Yuan and Cai (2010), and Yao, Müller and Wang (2005b)), and clustering and classification of functional data (e.g., James and Sugar (2003), Jacques and Preda (2014), Leng and Müller (2006), and Delaigle and Hall (2012)). For a broad theoretical, methodological and practical introduction to functional data analysis, interested readers are referred to the monographs by Ramsay and Silverman (2005), Ferraty and Vieu (2006), Ramsay, Hooker and Graves (2009), Horváth and Kokoszka (2012), Hsing and Eubank (2015), and Kokoszka and Reimherr (2017), and the review papers by Morris (2015) and Wang, Chiou and Müller (2016) and references therein. FDA also has many applications in other areas. For instance, Ainsworth, Routledge and Cao (2011) applied FDA for ecosystem research, in which they studied the relationship between river flow and salmon abundance. Luo, Cao, Gallagher, and Wiles (2013) estimated the intensity of ward admissions and investigated its effect on emergency department access in public hospitals by using FDA methods. However, little has been done on applying FDA in sports, except for Chen and Fan (2018) who investigated the score differential process in basketball by employing FDA.

In this paper, we apply FDA to evaluate the components in the conditional probability formulation that we use to express the in-game win probabilities. We model the functional feature extracted from the match event data and score differential processes as sums of smooth mean functions which are approximated by non parametric smoothing techniques and noises from Brownian motions. The mean functions are approximated by B-spline basis expansions with functional parameters. In FDA, a typical application involves the analysis of a sample of realizations from independent and identically distributed (iid) functions (e.g. Ramsay and Silverman (2005), Chapter 3.2.4, Cai and Hall (2006) and Hall and Horowitz (2007)). A novelty in our work is that the matches are not iid, because each match is conditional on a unique kickoff win probability of the home team. Therefore, we propose a weighted least squares method to estimate the functional parameters, which borrows the information from matches with similar kickoff win probabilities. The variance and covariance elements are obtained by the maximum likelihood estimation method. A key feature of our work is that the general approach for estimating in-game win probabilities may be used in any sport that has event data. Event data consists of a chronological record of well-defined events that occur during a match which are relevant to the match and are recorded with a time stamp. The necessary modifications to alternative sports would involve the determination of the relevant event data which is predictive and sport specific.

In Section 2, we begin with a discussion of the data that is at our disposal. We then outline a model from which we obtain the in-game posterior win probability. The model consists of distributions that are specified via FDA methods. The FDA methodology is explained in detail. In Section 3, we consider the utilization of the event data to provide good predictions. There are many potential insights from a game that are relevant. We use the domain knowledge from the rugby league for the specification. We then demonstrate that our estimated in-game win probabilities change during a match in expected ways. In Section 4, we demonstrate that our estimated win probabilities are reliable. We conclude with a short discussion in Section 5.

2. Model development.

2.1. *Available Data.* The NRL consists of 16 teams and each team plays 24 games during the regular season. The NRL has graciously given us access to event data for the resultant 769 regular season matches that have taken place during the four seasons 2016-2019. Event data are detailed match data that go well beyond box score data. With event data, every time an

event occurs during a match (e.g. field goal, try, tackle, etc.), characteristics of the event are recorded (e.g. location on the pitch, players involved, time of the match, etc.). In the NRL, 2.1 events are recorded on average per second, and there are up to 410 characteristics that can be recorded as an event. The events and characteristics are obtained through cameras and optical recognition software that carry out the data collection process in real time. Our dataset is a huge matrix with rows corresponding to events and columns corresponding to characteristics. Our dataset has 8,144,905 events (rows) obtained over the four seasons.

An important component of our work which is developed in Section 3 is the determination of relevant event data to inform the in-game win probabilities. We propose several choices of the event data that assist us to inform the in-game win probabilities. Specifically, we choose the event feature, missed tackle differential, to illustrate the proposed method. In our development and without loss of generality, the in-game win probabilities and data will refer to the home team.

For the time being, for a particular match, we will refer to $X(t)$ as a random functional feature that arises from the event data relative to the home team defined on a time interval $[0, 80]$ minutes. Note that $X(t)$ may be multivariate. For example, it is obvious that the average field position by the home team is a measure of dominance and it may be a good predictor of the home team's chance of winning the match.

Another important predictor of the in-game win probability of the home team is the current score differential. We will refer to $D(t)$ as the number of points by which the home team is defeating the road team at time t. Note that $D(t) < 0$ indicates that the road team is winning by $|D(t)|$ points at time t.

Finally, another important predictor of the in-game win probability of the home team is a measure of its strength relative to the road team. This is not something immediately available from the event data, and therefore we sourced an additional dataset. The website http://www.aussportsbetting.com/data/historical-nrl-results-and-odds-data/ gives closing betting odds of NRL matches immediately prior to kickoff. A nice feature of the betting odds is that they take into account everything that is relevant to a match including home team advantage, injuries, travel, etc. Betting odds are also known to be efficient; otherwise, sportsbooks would not exist. Therefore, we can rely on the betting odds as providing reliable information concerning the win-probability of the home team at the time of kickoff.

Betting odds arise in various formats, and we will refer to odds provided in the European format. Odds o_h on the home team indicate that a winning bet of \$1 on the home team will result in a payout of $\$o_h$. Clearly, $o_h \geq 1$. Similarly, odds o_r on the road team indicate that a winning bet of \$1 on the road team will result in a payout of $\$o_r$. We ignore the rare event that a match can end in a draw as this does not affect the subsequent calculations. Draws occur roughly 4.94% of the time in the NRL. Now, some simple probability calculations involving expectations yield that the probability of the home team winning is $p_h = 1/o_h$ and the probability of the road team winning is $p_r = 1/o_r$. However, these calculations do not take into account the vigorish (i.e. the expected profit) by the sportsbook, and therefore $p_h + p_r > 1$. We therefore remove the vigorish and set the kickoff probability that the home team wins the match as $p_0 = p_h/(p_h + p_r)$.

Therefore, to review, the inputs to our model which we use to estimate in-game win probabilities for the home team are given by:

$$
\begin{aligned}
&X(t) \equiv \text{functional feature extracted from the event data relative to the home team at time } t \\
&D(t) \equiv \text{score differential in favour of the home team at time } t \\
&p_0 \equiv \text{kickoff probability of the home team winning based on sportsbook odds}
\end{aligned}
\tag{1}
$$

2.2. *Model overview.* In this subsection, we present a model based on the inputs given by (1). We let W denote the event that the home team wins the match and let $\overline{W}$ denote the event that the home team does not win the match, and it is the posterior probability of W which is our quantity of interest. We therefore obtain the expression

$$\begin{aligned}\text{Prob}(W \mid X(t)=x(t), D(t)=d(t), p_0) &= \frac{f(x(t),d(t)|W,p_0)\ \text{Prob}(W|p_0)}{f(x(t),d(t)|W,p_0)\ \text{Prob}(W|p_0) + f(x(t),d(t)|\overline{W},p_0)\ \text{Prob}(\overline{W}|p_0)} \\ &= \frac{f(x(t),d(t)|W,p_0)\ p_0}{f(x(t),d(t)|W,p_0)\ p_0 + f(x(t),d(t)|\overline{W},p_0)\ (1-p_0)}\ , \end{aligned} \tag{2}$$

where $f(x(t), d(t) \mid W, p_0)$ represents the conditional joint density of $X(t)$ and $D(t)$ given W and p_0 , and $f(x(t), d(t) \mid \overline{W}, p_0)$ represents the conditional joint density of $X(t)$ and $D(t)$ given $\overline{W}$ and p_0. We observe that (2) is a simple expression for the purposes of calculation. However, for the application to television broadcasts, we emphasize that it is necessary that the component distributions in (2) need to be evaluated instantaneously.

2.3. *Estimation of model components using FDA.* This is the most technical portion of the paper where an atypical FDA structure is introduced and estimation techniques are developed to determine the conditional joint densities $f(x(t), d(t) \mid W, p_0)$ and $f(X(t), D(t) \mid \overline{W}, p_0)$ in (2). We illustrate the methodology with univariate $X(t)$ although the methods can be extended to multivariate $X(t)$. This subsection may be skimmed while still retaining the overall intent of the paper.

We begin by focusing on the $f(X(t), D(t) \mid W, p_0)$ term where $f(X(t), D(t) \mid \overline{W}, p_0)$ is handled in a similar fashion. Given W and p_0, we assume that

$$\begin{aligned} X(t) &= \mu_X(t, W, p_0) + \epsilon_X(t, W), \\ D(t) &= \mu_D(t, W, p_0) + \epsilon_D(t, W), \end{aligned}$$

where $\mu_X(t, W, p_0)$ is the expected value of the functional feature $X(t)$ given the home team winning and having a kickoff win probability of p_0. Similarly, $\mu_D(t, W, p_0)$ is the expected value of the score differential $D(t)$ given the home team winning and having a kickoff win probability of p_0. $\{\epsilon_X(t, W)\}_t$ and $\{\epsilon_D(t, W)\}_t$ are error processes.

For ease of notation, we use $\epsilon_X(t)$ and $\epsilon_D(t)$ to represent $\epsilon_X(t, W)$ and $\epsilon_D(t, W)$ respectively. In Section 3, we consider various choices for $X(t)$ that affect $\text{Var}(\epsilon_X(t))$ and the resultant estimation procedure. Suppose for now that $\epsilon_X(t)$ is a random variable that consists of independent incremental contributions up to time t. Therefore, we assume that $\epsilon_X(t)$ has mean 0 and variance $t\sigma_X^2$. However, we note that the following theory may be modified to accommodate other variance assumptions such as a constant variance. For $\epsilon_D(t)$, we also assume that it is based on a white noise process where we recognize that the score differential consists of incremental contributions during the match up to time t. Therefore, assuming that these contributions are independent and identically distributed, it is appropriate that $\epsilon_D(t)$ have mean 0 and variance $t\sigma_D^2$. These assumptions are equivalent to assuming that $\epsilon_X(t)$ and $\epsilon_D(t)$ are Brownian motion processes. We justify the normality assumptions in the supplementary document. The correlation between $X(t)$ and $D(t)$ is assumed to be invariant of t and let $\rho = \text{Corr}(X(t), D(t))$. Then, at time t, the noises are distributed as

$$\begin{pmatrix} \epsilon_X(t) \\ \epsilon_D(t) \end{pmatrix} \sim \text{Normal}\left(\mathbf{0}, t\mathbf{K}\right), \tag{3}$$

where $\mathbf{0} = (0, 0)^T$ and

$$\mathbf{K} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_D \\ \rho\sigma_X\sigma_D & \sigma_D^2 \end{pmatrix}.$$

For different time points t and t', $\text{Cov}(\epsilon_X(t), \epsilon_X(t')) = \min\{t, t'\}\sigma_X^2$, $\text{Cov}(\epsilon_D(t), \epsilon_D(t')) = \min\{t, t'\}\sigma_D^2$, and $\text{Cov}(\epsilon_X(t), \epsilon_D(t')) = \min\{t, t'\}\rho\sigma_X\sigma_D$.

We further assume that $\mu_X(t, W, p_0)$ and $\mu_D(t, W, p_0)$ are continuous smooth functions and we approximate these functions by linear combinations of basis functions as follows

$$\begin{aligned} \mu_X(t, W, p_0) &= \sum_{k=1}^{K} a_k(W, p_0) b_k(t), \\ \mu_D(t, W, p_0) &= \sum_{k=1}^{K} c_k(W, p_0) b_k(t), \end{aligned} \tag{4}$$

where b_k are predetermined basis functions. Up until this point, except for the variance assumptions associated with the noise terms, this is a standard setup in FDA applications (see, Chapter 3 in Ramsay and Silverman (2005), for example).

With our initial concentration on the specification of $f(x(t), d(t) \mid W, p_0)$, we restrict our data to matches where the home team has won (i.e. W is observed). Assume that we have functional data $\{(X_i(t_{ij}), D_i(t_{ij})) : i = 1, \ldots, N; j = 1, \ldots, n_i\}$. We also have the kickoff win probability p_{i0} associated with the ith match.

An aspect of our problem that makes it different from a typical FDA application is that the functional data are not iid. Specifically, the functional distribution of the ith match is conditional on p_{i0} (the kickoff win probability of the home team in the ith match). Suppose that $t_{i1} < t_{i2} < \cdots < t_{in_i}$ and let

$$\boldsymbol{\Sigma}_{i0} = \begin{bmatrix} t_{i1} & t_{i1} & t_{i1} & \ldots & t_{i1} \\ t_{i1} & t_{i2} & t_{i2} & \ldots & t_{i2} \\ t_{i1} & t_{i2} & t_{i3} & \ldots & t_{i3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ t_{i1} & t_{i2} & t_{i3} & \ldots & t_{in_i} \end{bmatrix}.$$

Therefore, to address the estimation of the a's and c's in (4), we minimize the functions

$$\begin{aligned} H_a(\mathbf{a}) &= \sum_{i=1}^{N} \exp\left\{\tfrac{-(p_0 - p_{i0})^2}{\gamma}\right\} (\mathbf{X}_i - \mathbf{B}_i\mathbf{a})^T \mathbf{G}_i^{-1} (\mathbf{X}_i - \mathbf{B}_i\mathbf{a}) \\ H_c(\mathbf{c}) &= \sum_{i=1}^{N} \exp\left\{\tfrac{-(p_0 - p_{i0})^2}{\gamma}\right\} (\mathbf{D}_i - \mathbf{B}_i\mathbf{c})^T \mathbf{H}_i^{-1} (\mathbf{D}_i - \mathbf{B}_i\mathbf{c}) \end{aligned} \tag{5}$$

where $\mathbf{a} = (a_1(W, p_0), \ldots, a_K(W, p_0))^T$, $\mathbf{c} = (c_1(W, p_0), \ldots, c_K(W, p_0))^T$, $\mathbf{X}_i$ is an $n_i \times 1$ vector with the jth element $X_i(t_{ij})$, $\mathbf{D}_i$ is an $n_i \times 1$ vector with the jth element $D_i(t_{ij})$, $\mathbf{B}_i$ is the $n_i \times K$ matrix with the (j, k)th element $b_k(t_{ij})$, and $\mathbf{G}_i = \mathbf{H}_i = \boldsymbol{\Sigma}_{i0}$. In (5), $\gamma > 0$ is a tuning parameter. The term $\exp\left\{-(p_0 - p_{i0})^2/\gamma\right\}$ assigns more weight to matches that have similar kickoff win probabilities to the generic value p_0.

The proposed estimation procedure is based on the minimization of the functions H_a and H_c. What makes the equations in (5) unusual is that $\text{E}(X_i(t_{ij})|W, p_{i0})$ and $\text{E}(D_i(t_{ij})|W, p_{i0})$ do not equal the specified expressions $\sum_{k=1}^{K} a_k(W, p_0) b_k(t_{ij})$ and $\sum_{k=1}^{K} c_k(W, p_0) b_k(t_{ij})$. Equality would only exist if the X_i and D_i were observed under the generic value p_0, where again, we emphasize that the functional data from different matches don't have the same conditional distribution because each match is conditional on a unique kickoff win probability p_{i0}. This provides the motivation for the exponential terms; we assign more weight to observations for which the generic p_0 is closer to the observed p_{i0}.

With a little bit of work, it can be shown that for fixed γ, the minimization of H_a and H_c yields the analytic expressions

$$\begin{aligned}\hat{\mathbf{a}} &= \left(\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{G}_i^{-1} \mathbf{B}_i\right)^{-1} \left(\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{G}_i^{-1} \mathbf{X}_i\right), \\ \hat{\mathbf{c}} &= \left(\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{H}_i^{-1} \mathbf{B}_i\right)^{-1} \left(\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{H}_i^{-1} \mathbf{D}_i\right),\end{aligned} \tag{6}$$

where $v_i = v_i(p_0, \gamma) = \exp\left\{-(p_0 - p_{i0})^2/\gamma\right\}$.

With estimated $\hat{a}$'s and $\hat{c}$'s, we now turn to more traditional estimation procedures. Let $\hat{a}_{ik} = \hat{a}_k(W, p_{i0})$ and $\hat{c}_{ik} = \hat{c}_k(W, p_{i0})$. Based on the data and modelling assumptions, the resulting likelihood is given by

$$L(\sigma_X, \sigma_D, \rho \big| W, p_0) = \prod_{i=1}^{N} (2\pi)^{-n_i} |\mathbf{\Sigma}_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{w}_i - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{w}_i - \boldsymbol{\mu}_i)\},$$

where

$$\mathbf{w}_i = \begin{pmatrix} \mathbf{X}_i \\ \mathbf{D}_i \end{pmatrix}, \ \boldsymbol{\mu}_i = \begin{pmatrix} \boldsymbol{\mu}_{iX} \\ \boldsymbol{\mu}_{iD} \end{pmatrix}$$

with $\boldsymbol{\mu}_{iX} = (\sum_{k=1}^{K} \hat{a}_{ik} b_k(t_{i1}), \ldots, \sum_{k=1}^{K} \hat{a}_{ik} b_k(t_{in_i}))^T$, $\boldsymbol{\mu}_{iD} = (\sum_{k=1}^{K} \hat{c}_{ik} b_k(t_{i1}), \ldots, \sum_{k=1}^{K} \hat{c}_{ik} b_k(t_{in_i}))^T$, and $\mathbf{\Sigma}_i$ is the Kronecker product of $\mathbf{K}$ and $\mathbf{\Sigma}_{i0}$, denoted by $\mathbf{K} \otimes \mathbf{\Sigma}_{i0}$. The parameters σ_X, σ_D, and ρ appear in matrix $\mathbf{K}$. The likelihood can then be maximized to provide estimates

$$\begin{aligned}\hat{\sigma}_X^2 &= D_{xx} \ / \ v_0, \\ \hat{\sigma}_D^2 &= D_{dd} \ / \ v_0, \\ \hat{\rho} &= D_{xd} \ / \ \sqrt{D_{xx} D_{dd}}\,,\end{aligned} \tag{7}$$

where

$$\begin{aligned} v_0 &= \textstyle\sum_{i=1}^{N} n_i, \\ D_{xx} &= \textstyle\sum_{i=1}^{N} \left(\mathbf{X}_i - \mathbf{B}_i \hat{\mathbf{a}}_i\right)^T \mathbf{G}_i^{-1} \left(\mathbf{X}_i - \mathbf{B}_i \hat{\mathbf{a}}_i\right), \\ D_{xd} &= \textstyle\sum_{i=1}^{N} \left(\mathbf{X}_i - \mathbf{B}_i \hat{\mathbf{a}}_i\right)^T \mathbf{\Sigma}_{i0}^{-1} \left(\mathbf{D}_i - \mathbf{B}_i \hat{\mathbf{c}}_i\right), \\ D_{dd} &= \textstyle\sum_{i=1}^{N} \left(\mathbf{D}_i - \mathbf{B}_i \hat{\mathbf{c}}_i\right)^T \mathbf{H}_i^{-1} \left(\mathbf{D}_i - \mathbf{B}_i \hat{\mathbf{c}}_i\right), \end{aligned} \tag{8}$$

with $\hat{\mathbf{a}}_i = (\hat{a}_{i1}, \ldots, \hat{a}_{iK})^T$, $\hat{\mathbf{c}}_i = (\hat{c}_{i1}, \ldots, \hat{c}_{iK})^T$. Finally, the parameter γ is tuned as described in Section 3.

Putting this all together, suppose that there is a new match l with a kickoff win probability p_{l0}, and we observe event data $x_l(t)$ and score differential $d_l(t)$ at time t. Then

$$\begin{aligned}\hat{f}(x_l(t), d_l(t) | W, p_{l0}) = \frac{1}{2\pi \, t\hat{\sigma}_X \hat{\sigma}_D \sqrt{1-\hat{\rho}^2}} \exp\Bigg\{ -\frac{1}{2(1-\hat{\rho}^2)} \Bigg[\frac{\left(x_l(t) - \sum_{k=1}^{K} \hat{a}_k(W, p_{l0}) b_k(t)\right)^2}{t\hat{\sigma}_X^2} - \\ \frac{2\hat{\rho} \left(x_l(t) - \sum_{k=1}^{K} \hat{a}_k(W, p_{l0}) b_k(t)\right) \left(d_l(t) - \sum_{k=1}^{K} \hat{c}_k(W, p_{l0}) b_k(t)\right)}{t\hat{\sigma}_X \hat{\sigma}_D} \\ + \frac{\left(d_l(t) - \sum_{k=1}^{K} \hat{c}_k(W, p_{l0}) b_k(t)\right)^2}{t\hat{\sigma}_D^2} \Bigg] \Bigg\}.\end{aligned}$$

Similarly, we can obtain $\hat{f}(x_l(t), d_l(t)|\overline{W}, p_{l0})$ by using the data of matches where the home team has lost (i.e. $\overline{W}$ is observed). Then by (2), we can simply estimate the posterior in-game win probability at time t for match l.

2.4. *In-Game win probabilities for the second half of the match.* The method proposed in Section 2.3, which we call the split FDA method, is used to estimate the in-game win probabilities in the first half of the match. By splitting the matches based on whether the home team wins or loses the match and using additional information provided by the functional feature X extracted from the event data, the split FDA method is expected to provide good predictions of the in-game win probabilities in the first half of the match. However, the method may not provide reasonable estimates toward the end of the match. The estimated in-game win probability at time $t = 80$ using the split FDA method may not be exactly 1 when the home team has won (W) or 0 when the home team has lost ($\overline{W}$). Moreover, the score differential $D(t)$ given that the home team won or lost is not normally distributed when the match approaches the end. For example, Figure 1 displays the histograms of the score differentials at time $t = 75$ for the matches that the home teams won (left) and lost (right) in the 2016 - 2018 seasons. We can observe that the distributions for $D(75)$ given either W or $\overline{W}$ are skewed.

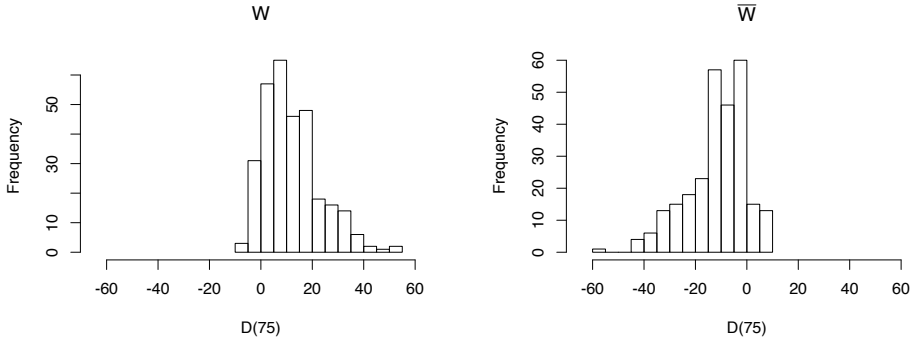

Fig 1: Left panel: histogram of the score differentials at $t = 75$ for the 311 matches in the training data (2016 - 2018 seasons) that the home teams won. Right panel: histogram of the score differentials at $t = 75$ for the 241 matches in the training data (2016 - 2018 seasons) that the home teams lost.

To overcome the above problems, we consider a method which does not split the matches into matches that home teams win and matches that home teams lose. We call this method the joint FDA method. The joint FDA method uses only the score differentials and kickoff win probabilities as inputs since the score differential is the most dominant factor that impacts the in-game win probabilities towards the end of a match. We assume that the score differential process follows a Brownian motion model with independent increments, which implies that $D(80) - D(t)$ is independent of $D(t)$. Therefore, we have

$$\begin{aligned} \text{Prob}(W \mid D(t) = d(t), p_0) &= \text{Prob}(D(80) > 0 \mid D(t) = d(t), p_0) \\ &= \text{Prob}(D(80) - D(t) > -d(t)|p_0). \end{aligned} \tag{9}$$

We assume that given p_0

$$D(t) = \mu_D(t, p_0) + \epsilon_{D\text{joint}}(t),$$

where $\mu_D(t, p_0)$ is the expected value of the score differential $D(t)$ with a kickoff win probability p_0 and $\epsilon_{D\text{joint}}(t)$ is the noise with mean 0 and variance $t\sigma^2_{D\text{joint}}$. We approximate $\mu_D(t, p_0)$ by $\sum_{k=1}^{K} \tilde{c}_k(p_0)b_k(t)$. We assume that $\epsilon_{D\text{joint}}(t)$ can be modeled as a Brownian motion process. Since the Brownian motion model assumes independent increments, $\text{Cov}(\epsilon_{D\text{joint}}(t), \epsilon_{D\text{joint}}(t')) = \min\{t, t'\}\sigma^2_{D\text{joint}}$ for different time points t and t'. Therefore,

$$D(80) - D(t) \sim \text{Normal}\left(\mu_{\Delta D}(t, p_0)), (80 - t)\sigma^2_{D\text{joint}}\right)$$

where $\mu_{\Delta D}(t, p_0) = \sum_{k=1}^{K} \tilde{c}_k(p_0)(b_k(80) - b_k(t))$. We estimate $\tilde{c}_k$'s and obtain $\hat{\sigma}_{D\text{joint}}$ by the relevant parts of (6) - (8) using all matches instead of restricting to the matches that the home teams win. Then for a new match l with a kickoff win probability p_{l0} and observed score differential $d_l(t)$ at time t, (9) yields

$$\widehat{\text{Prob}}(W \mid D_l(t) = d_l(t), p_{l0}) = \Phi\left(\frac{d_l(t) + \hat{\mu}_{\Delta D}(t, p_{l0})}{\sqrt{80 - t}\,\hat{\sigma}_{D\text{joint}}}\right),$$

where $\hat{\mu}_{\Delta D}(t, p_{l0}) = \sum_{k=1}^{K} \hat{\tilde{c}}_k(p_0)(b_k(80) - b_k(t))$ and Φ represents the cumulative distribution function of the standardized normal distribution. Compared to the split FDA method, the joint FDA method is more sensitive to the scoring events (see Section 4 for more details).

Now let $p_{\text{split}}(t)$ and $p_{\text{joint}}(t)$ denote the in-game win probabilities at time t obtained by the split FDA method and joint FDA method respectively. Let $w(t) = \frac{80-t}{40}$, then we use the weighted average $p(t) = w(t)p_{\text{split}}(t) + (1 - w(t))p_{\text{joint}}(t)$ to estimate the in-game win probability at time t in the second half of the match when $40 < t \leq 80$.

3. Results. We begin by considering appropriate choices for the functional match event feature $X(t)$. When a game is being viewed, there are often indications that one of the teams is gaining an upper hand in the match. The variable $X(t)$ is chosen to quantitatively reflect this sort of dominance as a predictor of winning the match. In Table 1, we propose several choices that are intended to reflect dominance by the home team. All of the variables presented in Table 1 are recorded with respect to the home team.

TABLE 1

Potential choices of event data where all variables are measured with respect to the home team and larger values denote increasing superiority.

Event feature	Description
$X_1(t)$	tackle differential up to time t
$X_2(t)$	tackle differential during the most recent 10 minutes at time t
$X_3(t)$	missed tackle differential up to time t
$X_4(t)$	missed tackle differential during the most recent 10 minutes at time t

For clarity, a missed tackle is one where a player on the team of interest may have been tackled, but the tackle was unsuccessful. Therefore, the missed tackle differential with respect to the home team is favourable to the home team if the variable is positive. Now, we are not suggesting that the variables proposed in Table 1 are the best choices. For example, Parmar et al. (2017) investigated key performance indicators in the professional rugby league. However, the variables in Table 1 are easy to calculate based on live match data. We imagine that experts with detailed domain knowledge of the rugby league may be able to propose improved variables from the point of view of prediction. However, to illustrate the proposed methods, we will hereafter use the variable $X_3(t)$ in Table 1 as the event data of interest. For ease of notation, we denote $X_3(t)$ as $X(t)$. We also emphasize that the choice of the event data impacts the modelling distribution (3) and the estimation equations given by (5)-(8).

The basis functions $b_k(t)$ introduced in (4) are cubic B-splines. For details on B-spline approximation, see de Boor (2001). Specifically, we choose 9 equally spaced knots over the interval $[0, 80]$ minutes and this results in $K = 11$ cubic B-spline basis functions as depicted in Figure 2. This selection of knots and splines leads to flexible shapes that can be used to express $\mu_X(t, W, p_0)$ and $\mu_D(t, W, p_0)$ in (4) and $\mu_D(t, p_0)$ in Section 2.4.

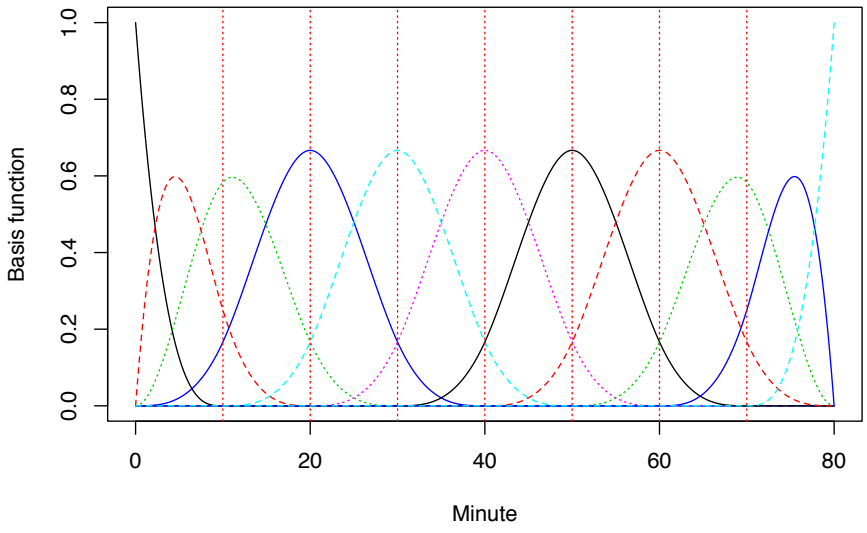

Fig 2: Cubic B-spline basis functions defined on 9 equally spaced knots over the interval $[0, 80]$ minutes.

Before proceeding to estimation, it is good to have a sense of the data. In Table 2 and 3, we provide descriptive statistics of data collected from 731 NRL regular-season matches from $2016 - 2019$. We observe that there is indeed a home-field advantage as the average score differential in favour of the home team is 1.8 points. We also observe that the average missed tackle differential is positive which is also evidence of the home team advantage. The score differential curves and the missed tackle differential curves for the 731 matches are plotted in Figures 3 and 4, respectively. On average, it seems that both the differential and missed tackle differential are linear with respect to the time of the match. This is consistent with a process whereby the better team separates itself from the weaker team in a consistent manner over the course of a match.

TABLE 2
Descriptive statistics of the scores corresponding to all 731 *matches from the four regular seasons* $(2016 - 2019)$ *of the NRL.*

Variable	Min Value	Max Value	Average	Std Dev
Home Team Score	0	64	21.1	10.8
Road Team Score	0	62	19.2	10.1
Score Differential wrt Home Team	-62	58	1.8	16.9

TABLE 3
Descriptive statistics of the missed tackles (at the end of the match) corresponding to all 731 *matches from the four regular seasons (*$2016-2019$*) of the NRL.*

Variable	Min Value	Max Value	Average	Std Dev
Home Team Missed Tackle	8	48	23.5	6.7
Road Team Missed Tackle	8	44	22.5	6.2
Missed Tackle Differential wrt Home Team	-31	34	1.0	9.6

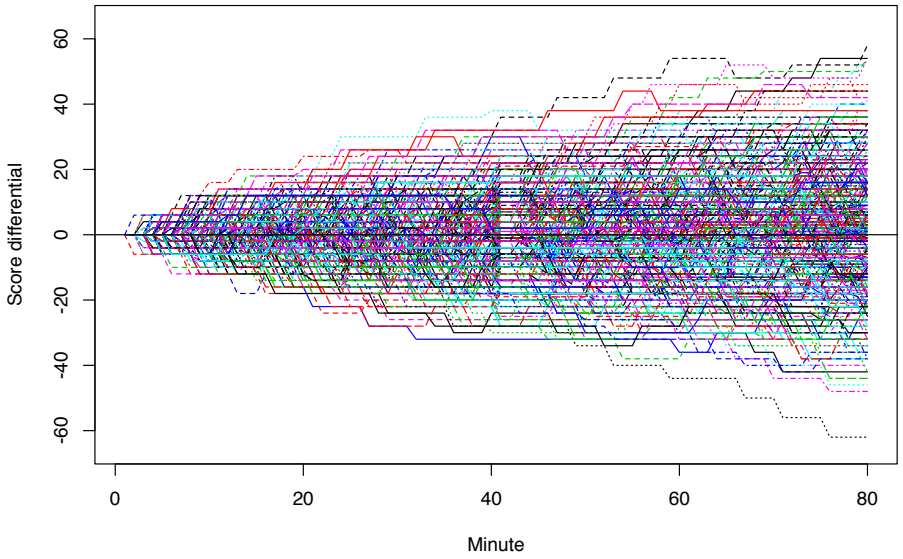

Fig 3: The score differential curves for all the 731 matches from the four regular seasons $(2016-2019)$ of the NRL.

Having specified the basis functions, the procedure in Section 2.3 requires the estimation of the parameters σ_X, σ_D, and ρ as specified in the multivariate normal distribution (3). We first restrict estimation to data where the home team has won (i.e. W) and we note that in the training data set (matches in $2016-2018$ seasons) there are 311 matches that fit this criterion. Based on the specification of the tuning parameter $\gamma = 0.01$, the chosen basis functions and the determination of the a_k and c_k terms, we obtain

$$\hat{\sigma}_X = 1.33,$$
$$\hat{\sigma}_D = 2.06,$$
$$\hat{\rho} = 0.21\ .$$

These estimates appear to be sensible in terms of the descriptive statistics provided in Table 2 and 3. In particular, we note a positive correlation $\hat{\rho}$ which suggests that $X(t)$ and $D(t)$ tend to work in tandem.

Using the training data $(2016-2018$ seasons) where the home team has not won (i.e. $\overline{W}$), there are 241 matches, and we similarly obtain

$$\hat{\sigma}_X = 1.31,$$
$$\hat{\sigma}_D = 2.00,$$
$$\hat{\rho} = 0.22\ .$$

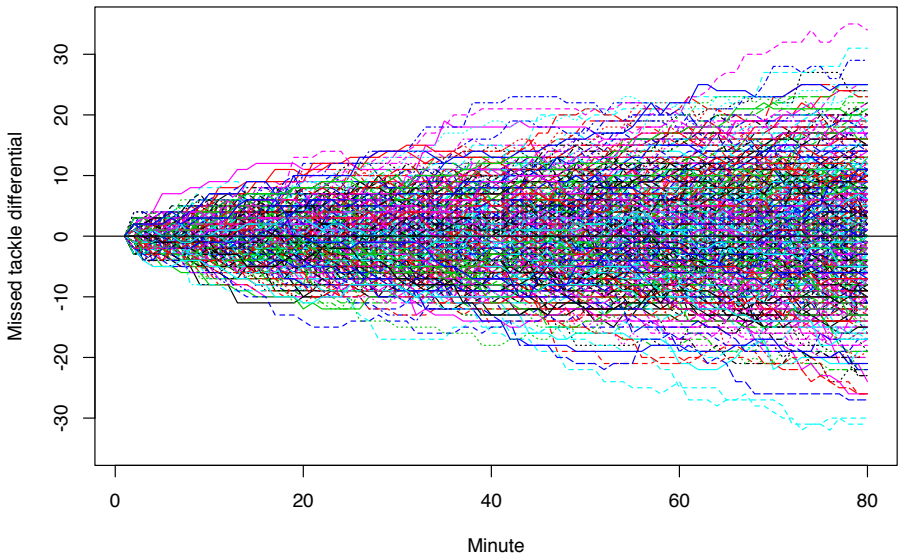

Fig 4: The missed tackle differential curves for all the 731 matches from the four regular seasons ($2016-2019$) of the NRL.

When we use all the 552 matches (i.e. W or $\overline{W}$) in the training data set ($2016-2018$ seasons), we obtain

$$\hat{\sigma}_{D\text{joint}} = 1.63.$$

Our estimation procedure involves a tuning parameter γ. We select the tuning parameter by fivefold cross-validation. Specifically, we randomly split the matches in $2016-2018$ seasons into five groups. For each unique group, we take it as a holdout test data set, On the remaining groups, for a particular γ, we fit the model parameters $(a, c, \sigma_X, \sigma_D, \rho)$ using the split FDA method and $(\tilde{c}, \sigma_{D\text{joint}})$ using the joint FDA method. We then apply the split FDA method and joint FDA method to estimate the home team win probability at time t on the hold out data set. If, for a given match at time t, the estimated in-game win probability is larger (smaller) than 0.5 and the home team eventually wins (loses) the match, then the prediction is considered to be correct. We repeat this procedure over all matches in the test set and all times to give the overall correct prediction rate. For both the split FDA and joint FDA methods, the choice $\gamma = 0.01$ yields the highest average overall correct prediction rate over all the five groups. In Figure 5, we show the estimated mean functions of the X and D processes with $\gamma = 0.01$ and various kickoff probabilities of the home team winning p_0. The top and middle panels present the estimated mean functions of X and D using the split FDA method. We observe that the plots exhibit the expected behaviors. For example, in matches where the home team wins, mean differentials in both X and D increase as the game progresses. When a curve is wiggly, we attribute this to lack of data. For example, in the top right plot where $p_0 = 0.8$, there are not many matches where the home team is heavily favored and they lose. The bottom panel of Figure 5 shows that for the joint FDA method, in general, the mean score differentials increase when the kickoff win probability is larger than 0.5 (i.e. $p_0 = 0.6$ and 0.8) and decrease when $p_0 = 0.2$ and 0.4.

4. Model validation. Obviously, there is a random component to sport and this is part of its appeal. If matches were perfectly predictable, then there would be no point in holding

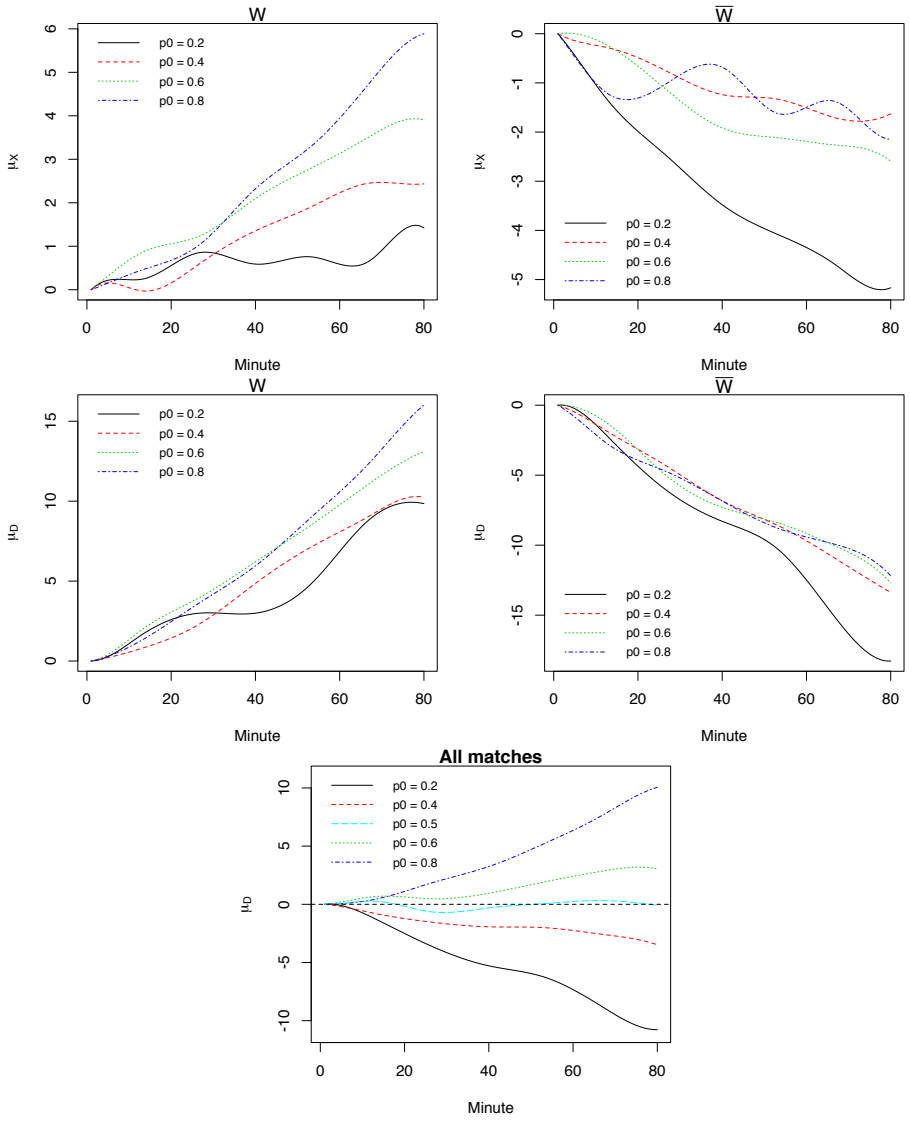

Fig 5: Top panel: estimated $\hat{\mu}_X(t, W, p_0)$ and $\hat{\mu}_X(t, \overline{W}, p_0)$ using the split FDA method. Middle panel: estimated $\hat{\mu}_D(t, W, p_0)$ and $\hat{\mu}_D(t, \overline{W}, p_0)$ using the split FDA method. Bottom panel: estimated $\hat{\mu}_D(t, p_0)$ using the joint FDA method.

sporting competitions. Therefore, our investigation in this section involves an assessment of whether our predictions are reasonable - they cannot and should not be perfect predictions.

We should not use the same data to both fit models and carry out the model assessment. We therefore fit our model using the first three seasons $2016 - 2018$ of the event data and use the fitted model to predict the match outcomes in the 2019 season for which there are 179 matches. We then compare the actual 2019 match outcomes with the predicted outcomes.

In Figure 6, we investigate the predictive capability of the split FDA method (dashed curve), joint FDA method (dotted curve), and the proposed weighted method (solid curve). We consider the estimated probability that the home team wins at times $t = 1, \ldots, 75$ for the 2019 data. It is sensible to only consider predictions up to the 75th minute as many sportsbooks terminate in-match betting towards the end of matches. A reason for this is that possession of the ball near the end of a close match is critical and becomes more important

than both X and D in the determination of fair betting odds. Punters could exploit this situation. If an estimated probability exceeds 0.5, then this indicates a prediction in favour of the home team. At time t, we compare the 2019 match predictions with the actual match results and obtain the correct prediction rate. As one would expect, Figure 6 demonstrates that the correct prediction rates obtained by all methods improve as matches progress in time. This figure shows that the split FDA method provides higher correct prediction rates than the joint FDA method for the first 40 minutes of the game, whereas the joint FDA method performs better in the second half especially when the game approaches the end. We observe that the methods yield good results exceeding 80% accuracy by the 55th minute.

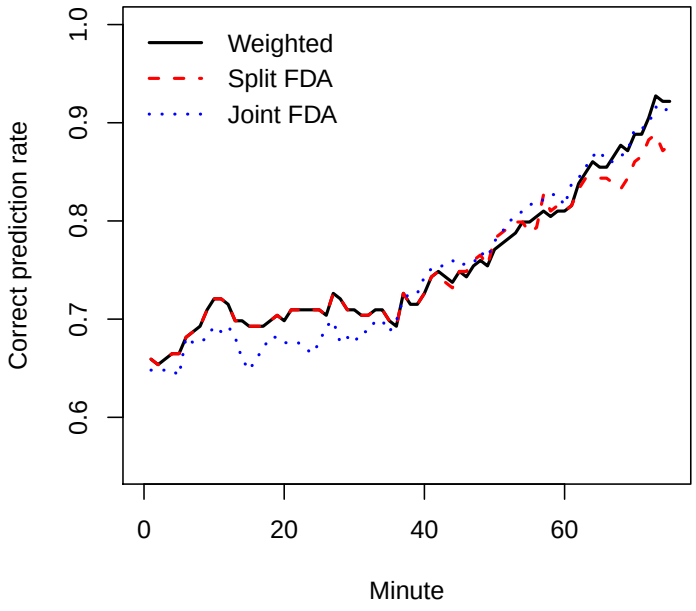

Fig 6: The correct prediction rates for the 2019 NRL season obtained by the split FDA method (dashed curve), the joint FDA method (dotted curve), and the proposed weighted method (solid curve). Note that the proposed weighted method has the same result as the split FDA method in the first 40 minutes of the match.

To investigate whether our estimated in-game win probabilities are reliable, we randomly select four matches from the 2019 season where the home teams won. In Figure 7, the solid curves are the predicted in-game win probabilities by the proposed method annotated with scoring events (dashed vertical lines). Recall that the proposed method applies the split FDA method in the first half of the match, whereas in the second half of the match, a weighted average of the estimates obtained by the split FDA method and the joint FDA method is used. In comparison, we also include the estimated in-game win probabilities by the split FDA method and joint FDA method. We can see that the joint FDA method is more sensitive to scoring events. For example, the top right plot shows a match that the home team won. When the home team scored at the 7th minute, the in-game probability obtained by the joint FDA

method increases dramatically, whereas the proposed method is less sensitive to the scoring event. Sensibly, we observe that the predicted win probabilities are impacted by scoring (discontinuous jumps).

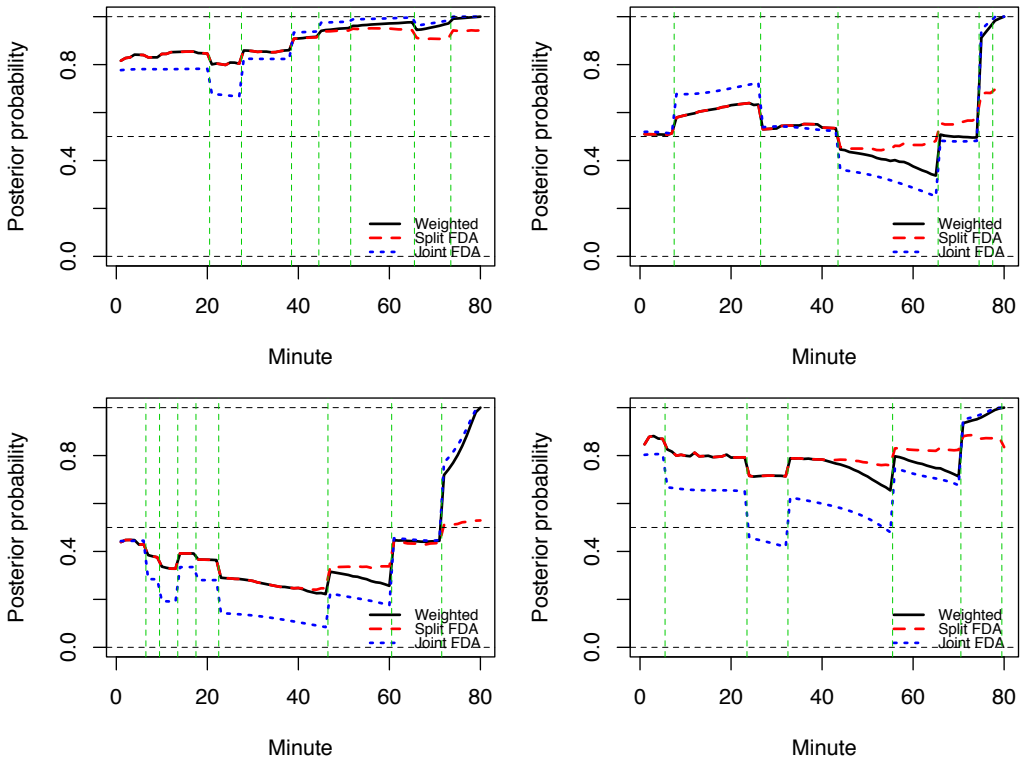

Fig 7: Predicted instantaneous in-game win probabilities by the proposed method (solid curves), the split FDA method (long-dashed curves), and the joint FDA method (short-dashed curves) for four randomly selected matches from the 2019 season where the home team won. The dashed horizontal lines indicate the values of 0, 0.5, and 1. The dashed vertical lines indicate the times when the score changed.

We also compare the proposed weighted method to the Brownian motion model that was proposed by Stern (1994) to study the scoring process of basketball. Stern (1994) estimated the drift and variance parameters of the Brownian motion model by treating the game outcome as a binary response and maximizing the profit regression likelihood. We use "BMM" to represent the Brownian motion model proposed by Stern (1994). One limitation of the BMM method is that the probability that a home team wins the match when it leads d points at time t is assumed to be the same for any basketball game. Figure 8 (a) compares the correct prediction rates obtained by the proposed method (solid curve) and the BMM approach (dashed curve). It is observed that our proposed method outperforms the BMM approach in the first half of the match. In Figure 8 (b), we present the predicted in-game probabilities by the proposed method and the BMM approach. The results indicate that the BMM approach is very sensitive to the scoring events. Because the BMM method does not use the information of the betting odds, the predicted in-game probabilities at time 0 are 0.6 for all matches in 2019.

To see how $X(t)$ impacts the estimation procedure, we consider two scenarios. In Scenario I, the split FDA method predicts the in-game win probabilities using both the event data X

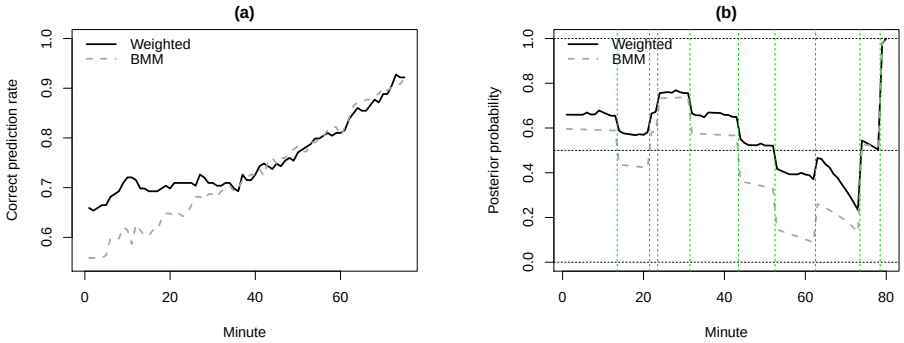

Fig 8: (a) The correct prediction rates for the 2019 NRL season obtained by the proposed weighted method (solid curve) and the BMM approach (dashed curve). (b) Predicted instantaneous in-game win probabilities by the proposed weighted method (solid curves) and the BMM approach (dashed curves) for one randomly selected match from the 2019 season. The dashed horizontal lines indicate the values of 0, 0.5, and 1. The dashed vertical lines indicate the times when the score changed.

and the score differential D, whereas in Scenario II, the split FDA method predicts the in-game win probabilities using only the score differential D. For both scenarios, the joint FDA method uses only the score differentials. We select a match played on 6th April 2019 between the Melbourne Storm (home) and the Canterbury-Bankstown Bulldogs. The half time score is 6 (Storm) - 12 (Bulldogs) and the full time score is 18 (Storm) - 16 (Bulldogs). More details about the match can be found at https://www.nrl.com/draw/nrl-premiership/2019/round-4/storm-v-bulldogs/.

In Figure 9, we present the predicted instantaneous in-game win probabilities for the match under Scenario I and Scenario II together with the score differentials and missed tackle differentials. The solid curve in Figure 9 represents the predictions obtained using Scenario I, and the dotted curve represents the predictions based on Scenario II. The kickoff win probability $p_0 = 0.85$ indicates that the Storm was heavily favored. We can see from Figure 9 that the road team scored on the 6th minute of the match, and after that, the predicted in-game win probabilities based on D only (Scenario II) decreased to below 0.8. In contrast, the missed tackle differentials keep positive for most of the time in the first half of the match. This indicates that even though the Storm were trailing, there was reason to be hopeful that they would turn the match around. We observe that the predicted in-game win probabilities based on Scenario I are greater than those based on Scenario II for the entire game except for the short time interval between the 24th and 32nd minute. Clearly, the example demonstrates the added value in the event data $X(t)$ through the superiority of Scenario I over Scenario II.

5. Discussion. We have developed a model that provides instantaneous in-game win probabilities for the National Rugby League. The model has distributional components that are informed by FDA techniques.

There are various future research directions associated with our work. First, the approach is general and is applicable to other sports whenever suitable event data are available. Second, there are obvious gambling questions that may be explored with respect to our predictions. Finally, the choice of the functional event feature $X(t)$ impacts our estimation procedure, and we have focused on the missed tackle differential. We believe that experts with detailed domain knowledge of the rugby league may be able to propose better predictive choices for $X(t)$. Although we illustrate the use of univariate $X(t)$, our methods can be extended to multivariate settings.

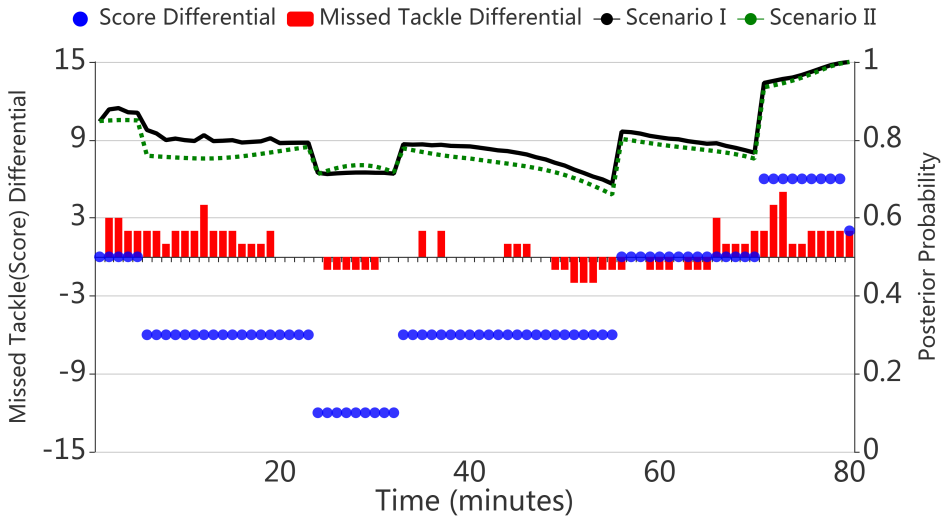

Fig 9: Predicted instantaneous in-game win probabilities for the match Storm versus Bulldogs on 6th April 2019 by Scenario I (———) and Scenario II (·····). The dots indicate the score differentials of the match. The bars indicate the missed tackle differentials of the match.

Supplementary Document. The supplementary document contains the numerical results for justification of the normality assumption.

Acknowledgments. The authors would like to thank the Editor, the Associate Editor and two anonymous referees for their constructive comments, which are very helpful to improve the quality of this paper. We are particularly grateful to the NRL for providing the data.

REFERENCES

ALBERT, J.A., GLICKMAN, M.E., SWARTZ, T.B. AND KONING, R.H., EDITORS (2017). *Handbook of Statistical Methods and Analyses in Sports*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

AINSWORTH, L.M., ROUTLEDGE R., AND CAO J (2011). Functional Data Analysis in Ecosystem Research: the Decline of Oweekeno Lake Sockeye Salmon and Wannock River Flow. *Journal of Agricultural, Biological, and Environmental Statistics* **16** 282–300.

BESSE, P. AND RAMSAY, J.O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51** 285–311.

BOOTH, M. AND ORR, R. (2017). Time-loss injuries in sub-elite and emerging rugby league players. *Journal of Sports Science and Medicine* **16** 295–301.

BOSQ, D. (2000). *Linear processes in function spaces : theory and applications*. Springer, New York.

BUTTREY, S.E., WASHBURN, A.R. AND PRICE, W.L. (2011). Estimating NHL scoring rates. *Journal of Quantitative Analysis in Sports* **7** 1–18.

CAI, T. T. AND HALL, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34** 2159–2179.

CARDOT, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics* **12** 503–538.

CARDOT, H., FERRATY, F. AND SARDA, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13** 571–591.

CERVONE, D., D'AMOUR, A., BORNN, L. AND GOLDSBERRY, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of American Statistical Association* **111** 585–599.

CHEN, T. AND FAN, Q (2018). A functional data approach to model score difference process in professional basketball games. *Journal of Applied Statistics* **45** 112–127.

CLAUSET, A., KOGAN, M. AND REDNER, S. (2015). Safe leads and lead changes in competitive team sports. *Physical Review E* **91** 062815.

DE BOOR, C. (2001). *A practical Guide to Splines*. Springer-Verlag, New York.

DELAIGLE, A. AND HALL, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B* **74** 267–286.

FERRATY, F. AND VIEU, P. (2006). *Nonparametric Functional Data Analysis*. Springer, New York.

GABBETT, T.J. (2005). Science of rugby league football: A review. *Journal of Sports Sciences* **23** 961–976.

GABLE, A. AND REDNER, S. (2012). Random walk picture of basketball scoring. *Journal of Quantitative Analysis in Sports* **8** 1–20.

GLASSBROOK, D.J, DOYLE, T.L.A, ALDERSON, J.A. AND FULLER, J.T. (2019). The demands of professional rugby league match-play: a meta-analysis. *Sports Medicine - Open* **5** Article number: 24.

HALL, P. AND HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics* **35** 70–91.

HASTIE, T. AND MALLOWS, C. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 140–143.

HORVÁTH, L. AND KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. New York: Springer.

HSING, T. AND EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Chichester: Wiley.

JACQUES, J. AND PREDA, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis* **71** 92–106.

JAMES, G.M. AND SUGAR, C.A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98** 397–408.

KAYHAN, V. O. AND WATKINS, A. (2018). A data snapshot approach for making real-time predictions in basketball. *Big Data* **6** 96–112.

KAYHAN, V. O. AND WATKINS, A. (2019). Predicting the point spread in professional basketball in real time: a data snapshot approach. *Journal of Business Analytics* **2** 63–73.

KING, T., JENKINS, D. AND GABBETT, T. (2009). A time-motion analysis of professional rugby league match-play. *Journal of Sports Sciences* **27** 213–219.

KOKOSZKA, P. AND REIMHERR, M. (2017). *Introduction to Functional Data Analysis*. Boca Raton: Chapman and HallCRC.

LEE, A. (1999). Applications: Modelling rugby league data via bivariate negative binomial regression. *Australian & New Zealand Journal of Statistics* **14** 141–152.

LENG, X. AND MÜLLER, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** 68–76.

LOCK, D. AND NETTLETON, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports* **10** 197–205.

LUO, W., CAO, J., GALLAGHER, M. AND WILES, J. (2013). Estimating the intensity of ward admission and its effect on emergency department access block. *Statistics in Medicine* **32** 2681–2694.

MORRIS, J.S. (2015). Functional regression. *Annual Review of Statistics and Its Application* **2** 321–359.

PARMAR, N, JAMES, N., HUGHES, M., JONES, H. AND HEARNE, G. (2017). Team performance indicators that predict match outcome and points difference in professional rugby league. *International Journal of Performance Analysis in Sport* **17** 1044–1056.

PETTIGREW, S. (2015). Assessing the offensive productivity of NHL players using in-game win probabilities. *In Proceedings of the 9th MIT Sloan Sports Analytics Conference*.

RAMSAY, J.O., HOOKER, G. AND GRAVES, S. (2009). *Functional Data Analysis with R and Matlab*. Springer, New York.

RAMSAY, J.O. AND SILVERMAN, B.W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York.

ROBBERECHTS, P., VAN HAAREN, J. AND DAVIS, J. (2019). Who will win it? An in-game win probability model for football. *In Proceedings of the 6th Workshop on Machine Learning and Data Mining for Sports Analytics, 20 September 2019, page 13.* Würzburg, Germany.

SEITZ, L.B., RIVIÈRE, M., DE VILLARREAL, E.S. AND HAFF, G.G. (2014). The athletic performance of elite rugby league players is improved after an 8-week small-sided game training intervention. *Journal of Strength and Conditioning Research* **28** 971–975.

SONG, K., GAO, Y. AND SHI, J. (2020). Making real-time predictions for NBA basketball games by combining the historical data and bookmaker's betting line. *Physica A* **547** 124411.

STERN, H.S. (1994). A Brownian motion model for the progress of sports scores. *Journal of the American Statistical Association* **89** 1128–1134.

ŠTRUMBELJ, E. AND VRAČAR, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting* **28** 532–542.

VRAČAR, P., ŠTRUMBELJ, E. AND KONONENKO, I. (2016). Modeling basketball play-by-play data. *Expert Systems with Applications* **44** 58–66.

WANG, J.-L, CHIOU, J.-M. AND MÜLLER, H.-G. (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application* **3** 257–295.

WAND, M.P. AND JONES, M.P. (1995). *Kernel smoothing*. Chapman and Hall/CRC.

WINDT, J., GABBETT, T.J., FERRIS, D. AND KHAN, K.M. (2017). Training load–injury paradox: is greater preseason participation associated with lower in-season injury risk in elite rugby league players? *British Journal of Sports Medicine* **51** 645–650.

YAO, F., MÜLLER, H.-G. AND WANG, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100** 577–590.

YAO, F., MÜLLER, H.-G. AND WANG, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33** 2873–2903.

YUAN, M. AND CAI, T.T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics* **38** 3412–3444.