Flexible marked spatio-temporal point processes with applications to event sequences from association football

Santhosh Narayanan[1,4], Ioannis Kosmidis[1,4], and Petros Dellaportas[2,3,4]

[1]University of Warwick, Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
[2]Department of Statistical Science, University College London, Gower St. London WC1E 6BT, UK
[3]Department of Statistics, Athens University of Economics and Business, 76 Patission Str., Athens 10434, Greece
[4]The Alan Turing Institute

March 9, 2021

Abstract

We develop a new family of marked point processes by focusing the characteristic properties of marked Hawkes processes exclusively to the space of marks, providing the freedom to specify a different model for the occurrence times. This is possible through a decomposition of the joint distribution of marks and times that allows to separately specify the conditional distribution of marks given the filtration of the process and the current time. We develop a Bayesian framework for the inference and prediction from this family of marked point processes that can naturally accommodate process and point-specific covariate information to drive cross-excitations, offering wide flexibility and applicability in the modelling of real-world processes. The framework is used here for the modelling of in-game event sequences from association football, resulting not only in inferences about previously unquantified characteristics of the game dynamics and extraction of event-specific team abilities, but also in predictions for the occurrence of events of interest, such as goals, corners or fouls, in a specified interval of time.

Keywords: *Bayesian inference*; *Hamiltonian Monte Carlo*; *team abilities*; *branching structure*

1 Introduction

Football is one of the most popular team sports and is an example of an invasive sport, where two opposing teams compete for the possession of the ball with the dual objective of attacking to score a goal and defending against attacks from the opposition. Most analyses in football are typically done manually by studying video footage or using simple frequency analysis of match events. Hence, there is huge scope to improve the efficiency of the data-analytic methods as well as the quality of performance evaluation. However, the analysis of football data is mathematically challenging due to the continuous interaction between players within and across the two teams. As an introduction, we describe the event data from football and survey the existing work in this area before arguing how marked point processes are well-suited to developing a modelling foundation to achieve our goal of describing the game dynamics.

1.1 Football event data

Over the last decade, the availability of spatio-temporal data from team sports has inspired research into the application of statistical methods for team and player performance evaluation. A comprehensive survey of the recent research efforts into the spatio-temporal analysis of team sports is provided in Gudmundsson and Horton (2017). There are two primary types of spatio-temporal data collected from team sports. Movement data consists of samples of timestamped locations in the plane tracking the movement of all players and the ball during the game. Player movement is captured using fixed cameras in optical tracking systems, that process the images to obtain the trajectories. Event data streams, on the other hand, record the sequence of events that occur during the game and are collected manually by trained analysts who watch video feeds of the games through a special annotation software. As our work is motivated by the availability of event data from football, we focus on reviewing research that use event data streams. Event data are less dense than movement data, but richer in the sense that they contain more information about what is happening in the game. Events broadly fall into two categories; player events such as passes and shots; and stoppage events such as fouls, end of game etc. Every event is annotated with a timestamp, its location, its type (pass, foul, etc.), the players involved, and team information.

A popular research topic based on event data is the network analysis of player interaction. Models for player interaction can quantify a team's playing style as well as the importance of an individual player within the team. Players are identified as nodes of a network and are connected using directed edges whose weights are proportional to the number of successful passes between the two players. Passing networks were first applied to team sports in Passos et al. (2011) to study a team's collective behaviour in water polo. Grund (2012) studied the degree centrality of passing networks in football, which quantifies the importance of nodes in the network based on the number of edges. They showed that teams that rely heavily on key players performed relatively worse. Duch et al. (2010) used flow centrality to assess player performance by capturing the fraction of times that a player intervenes in those paths that result in a shot on goal. They also take into account defensive behaviour by letting each player start a number of paths proportional to the number of balls they recover. Clemente et al. (2015) studied the density and heterogeneity of passing networks and showed that high heterogeneity leads to formation of sub-communities, meaning there is a low level of cooperation between the players of a team. Pena and Touchette (2012) looked at other centrality measures such as closeness and eigenvector centrality as well as clustering in football passing networks.

Another use of event data is in the identification of frequently occurring sequences of passes between a small group of players within the same team. In Borrie et al. (2002), passes are identified by the zones in the pitch they start and end in and frequently occurring sequences are detected by also taking into account the time intervals between passes. Wang et al. (2015) proposed an unsupervised approach to automatically detect tactical patterns in football. They present the Team Tactic Topic model based on Latent Dirichlet Allocation to identify tactics from pass sequences. Interesting visualisations are provided for the most successful tactics as well as how a team's tactical patterns evolve over a season. Van Haaren et al. (2016) also look at automatic discovery of patterns in attacking strategy. They use a data-driven approach to determine a number of spatial features about the areas occupied during a continuous possession phase of a team. The features are then used to cluster similar phases together to identify frequently occurring event sequences within the cluster. Decroos et al. (2017) partition the game using overlapping intervals to create subsequences of events to use as a feature to predict a goal event in the near future. They compute similarity between subsequences using Dynamic Time Warping, a distance measure for time-dependent sequences.

Extracting game states from event sequences to quantify the value of player actions or to make predictions of the game outcome is another interesting area of research. Routley and

Schulte (2015) used Markov decision processes for valuing player actions in Ice Hockey. Game states are derived from contextual features like game score and time remaining along with the recent history of events. The associated reward for an action in the Markov decision process gives the value of the player action. A similar approach based on game states is taken in Decroos et al. (2018) to value player actions in football. They train a classification model to calculate the probability a game state will lead to a goal in the near future, where each game state is described using over 150 features. The value of a player action is then calculated by the shift in the predicted goal probability before and after the action. Other approaches for predicting goal probabilities based on a current game state are by Mackay (2017) and Robberechts et al. (2019). Approaches based on game states involve significant effort into feature engineering and with the use of learning algorithms like gradient boosting that limit parameter interpretations, the methods provide, typically, little insight into the dynamics of the game.

The major focus of existing methods in team sport analysis appear to be tailored towards individual player performance evaluation or identifying specific patterns in team play. Most approaches take the route of summarising the event data into compact representations like networks and game states. In this paper, we take a more holistic approach to study football as a dynamic system and model the entire sequence of events within a game. Such a model, that captures all event interactions, is attractive for predicting the occurrence of the rare goal scored events, that determine the outcome of the game.

1.2 Point processes

Phenomena that are observed as a sequence of events happening over time can be represented using point processes. While point processes can describe a random collection of points in any general space, we limit ourselves to the case in which the points denote events that occur along a time axis. Such point processes, having a natural order in which the points occur, are suitable for a wide range of real-world applications and are well studied in probability theory (see, for example, Daley and Vere-Jones, 2003).

Processes in which points are identified only by the occurrence times are referred to as univariate point processes. Multivariate point processes, on the other hand, are those in which two or more types of points are observed and are specified by associating a random variable, say m, to each point in a univariate point process, where the realised value of m gives the point type. If m is allowed to be a general random variable, not restricted to be a category giving the point type, then we refer to m as a mark and the process as a marked point process. An example application of a marked point process with continuous marks is in seismology, where the magnitude of an earthquake is recorded in addition to the time of occurrence. In this paper, we model event sequences observed in football using marked point processes with discrete marks used to denote the event type.

When event sequence data are analysed using point process models, an important distinction is between empirical models and mechanistic models as noted by Diggle (2013). Empirical models have the solitary aim of describing the patterns in the observed data, while mechanistic models go beyond that and attempt to capture the underlying process that generated the data. Mechanistic models for marked point processes are typically specified using a joint conditional intensity for the occurrence times and the marks and in general are not flexible enough to be applied to complex real-world phenomena. The joint modelling of the components of the process can also be challenging and it is common to make strong restrictive assumptions like separability (González et al., 2016) to simplify the model. In this paper, we present a flexible mechanistic modelling framework for marked point processes that are suitable for a wide range of applications without the need for assumptions like separability.

We produce a family of marked point processes that generalises the classical Hawkes process, a mathematical model for self-exciting processes proposed in Hawkes (1971) that can be used to

model a sequence of arrivals of some type over time, for example, earthquakes in Ogata (1998). Each arrival excites the process in the sense that the chance of a subsequent arrival increases for a period of time after the initial arrival and the excitation from previous arrivals add up. Marked Hawkes processes are typically specified using a joint conditional intensity function for the occurrence times and the marks (see, for example, Rasmussen, 2013, expression 2.2). The marked Hawkes process model captures the magnitudes of all cross-excitations between the various event types as well as the rate at which these excitations decay over time. Excitation leads to clustering of events in time as the process is driven by an intensity that increases with every arrival for a short period of time. However, in applications like the event sequences observed in football, the events tend not to cluster in time and the marked Hawkes process model is not suitable. The joint modelling of the times and the marks has to be decoupled to restrict the excitation property of the process exclusively to the dimension of the marks.

We take advantage of the decomposition of a multivariate distribution function that motivated the partial likelihood in Cox (1975). The joint conditional distribution for a marked point process can be factorised into a probability density function for the next event time conditioned on the past occurrences times and marks, and a probability mass function for the event mark conditioned on the time of occurrence and the filtration of the process. Therefore, an alternative approach to specify a marked point process model is to specify the conditional distribution functions for the times and the marks separately. We derive the conditional distribution function for the marks from a marked Hawkes process, which gives us then the freedom to specify the conditional distribution for the times separately. In this way, we are able construct marked point process models that retain the characteristic properties, of Hawkes processes, such as excitation in the model for the marks while avoiding the strong clustering of event times.

We develop a framework for Bayesian inference of such flexible marked point processes, which is realised through the Stan (Stan Development Team, 2020) software for statistical modelling. Stan implements a variant of the Hamiltonian Monte Carlo algorithm, originally proposed by Duane et al. (1987), to generate samples from the posterior distribution of the parameters. The Bayesian models we consider are compared using the out-of-sample log predictive density.

We define marked point processes for the modelling of touch-ball events in football, which along with time and event type information also carry location information. As we illustrate, the family of marked point processes can be readily enriched to handle all times, event types and locations. We are also able to incorporate team information in a direct way that captures the relative abilities of the teams for each event type. We develop a method based on association rules to reduce the increased model complexity introduced by model extensions. The rule-based approach identifies significant event interactions within sequences by placing thresholds on particular measures of significance. We then evaluate the accuracy of the excitation based models by comparing against two baseline models and confirm the superior performance of the models with excitation effects.

We provide a detailed parameter description showing how the model parameters can be used to gain valuable insights into football. The excitation component of the proposed model captures both the magnitudes and the durations of all pairwise event interactions across different locations. From the conversion rate parameters, we are able to confirm the well-known home advantage effect, and quantify the relative performance of each team when playing games at their home venue compared to away. We also discuss how the team ability parameters can be used to obtain rankings for the teams by event type, that can be used as predictors of team performance. The team ability parameters also capture some interesting differences in the playing styles of the teams, that are not immediately apparent. In this way, the model along with its parameters can be used to develop a deeper understanding of the game-play by the coaching staff and inform strategic decision making. The proposed model can also be used to simulate the sequence of events in a game to obtain real-time predictions of event probabilities.

4

Table 1: A snapshot of the data: a sequence of events and their attributes.

| second	minute	team ID	player ID	type	outcome	x	y	end x	end y
0	0	1	68312	Pass	Successful	49.1	51.0	52.5	44.8
2	0	1	14036	Pass	Successful	52.2	44.5	36.7	60.6
3	0	1	79050	Pass	Successful	36.7	60.6	24.9	39.1
5	0	1	14107	Pass	Unsuccessful	25.0	37.9	97.0	22.9
11	0	2	73379	Tackle	Successful	1.9	73.7		
15	0	2	73379	Pass	Successful	5.5	65.3	20.9	21.5
17	0	2	6292	Pass	Successful	20.9	21.5	29.0	38.5
19	0	2	26820	Foul	Successful	25.8	37.4	25.8	37.4
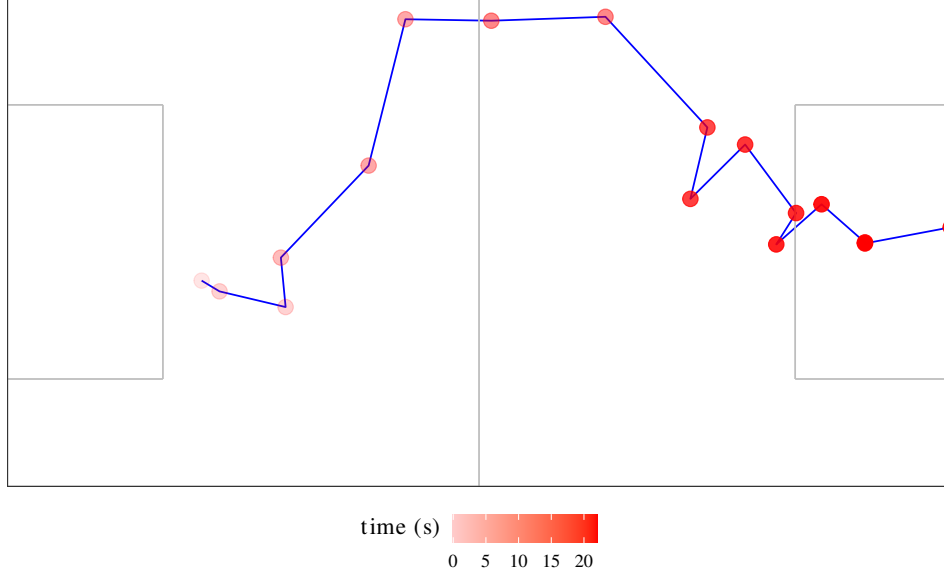
Figure 1: The sequence of events that led to the goal scored by Jack Wilshere for Arsenal against Norwich City (voted the best goal of the 2013/14 season).

The simulator results in predictions that can enhance, among others, the viewing experience of televised games. Finally, like Hawkes Processes, the proposed model also allows the recovery of the hidden branching structure of the process that quantifies the relative contributions of the background and previous occurrences to the triggering of a new event.

The developments in this paper can be readily applied to many other team sports like rugby, hockey, basketball etc. As none of the methods have been tailored specifically to football or even sports for that matter, they can also be applied to a wide range of applications that generate event data streams.

2 Data

2.1 Description and descriptives

The data that motivated this work was provided by Stratagem Technologies Ltd, and it consists of all touch-ball events from all English Premier League games in the 2013/14 season. A touch-ball event is an event where a player has acted on the ball by touching it with some part of their body. In total, the data consists of over half a million touch-ball events recorded over the season along with other attributes. A snapshot of the data is provided in Table 1. The league is contested by a total of 20 teams and follows a round-robin tournament scheduling, where each
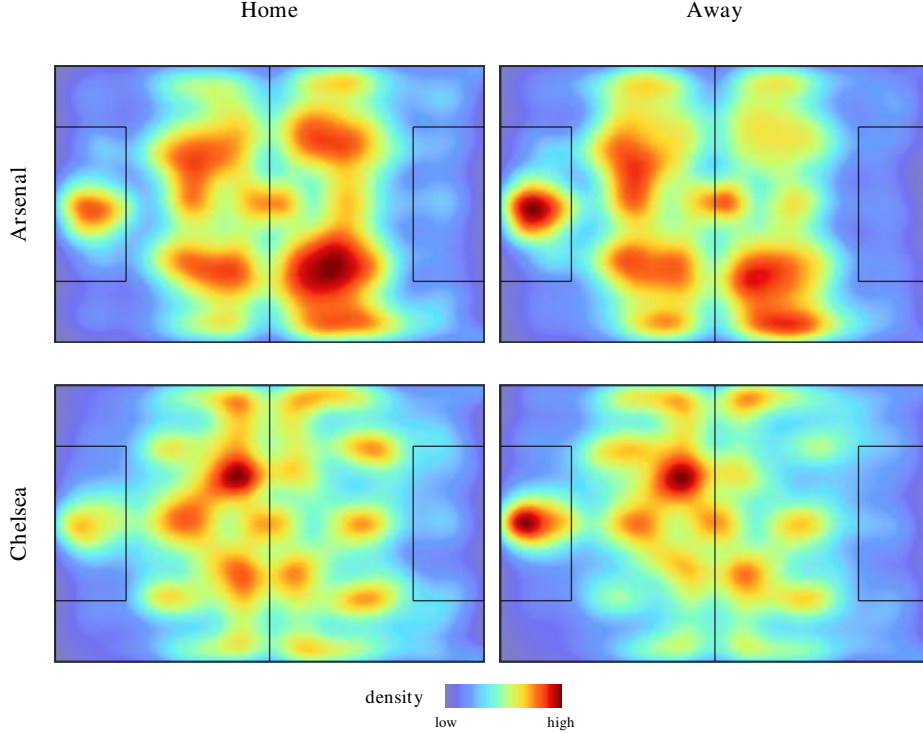
Home   Away

Figure 2: Heat maps showing the density of ball-touches for Arsenal and Chelsea in their home and away games in the 2013/14 season. In all heat maps the team is attacking to the right.

team plays every other team at their home and away venues, which results in a total of 380 games over the season.

Each game comprises two halves that are separated by an interruption of approximately 15 minutes. In what follows, we refer to each uninterrupted game half as a game period. For each touch-ball event, we have records of the event type, time-stamp, (x, y) co-ordinates of its location in the playing field, team and player ids, game period, event outcome (successful/unsuccessful) and the end (x, y) co-ordinates if the touch-ball event is a Pass. Table 2 gives the frequency of each of the 22 distinct touch-ball event types recorded in the data.

Table 2: Frequencies of the 22 distinct types of touch-ball events in the data.

event type	frequency	event type	frequency
Pass	376924	SavedShot	4971
BallRecovery	36908	Save	4910
Clearance	25462	CornerAwarded	4100
Tackle	14581	MissedShots	4076
TakeOn	13607	OffsidePass	1582
BallTouch	13517	Claim	1181
Aerial	12871	Goal	1052
Interception	10422	Punch	380
Dispossessed	8897	ShotOnPost	187
Foul	8238	Smother	122
KeeperPickup	5208	CrossNotClaimed	81

Figure 1 shows the trajectory of the ball during an attacking move that led to a goal in the 18th minute of the game between Arsenal and Norwich City on October 19, 2013. The goal was scored by Jack Wilshere for Arsenal and was voted as the best goal in the English Premier League for the 2013/14 season.

Latent game characteristics, such as the home advantage, and differences in playing styles and formations between the teams are also reflected in the touch-ball events. For example, Figure 2 compares the concentration of ball touches for Arsenal and Chelsea between their home and away games in the 2013/14 season. The playing field is plotted so that the team is always attacking to the right. It is clear that when the teams play at home the density of events is higher towards the opponent's goal. In fact, the point process modelling framework developed

Table 3: Composite event types along with their labels and observed frequencies in the data.

| m	mark label	count	m	mark label	count
1	Home_Win	10864	16	Away_Win	10829
2	Home_Dribble	3432	17	Away_Dribble	3123
3	Home_Pass_S	152140	18	Away_Pass_S	140975
4	Home_Pass_U	42344	19	Away_Pass_U	41462
5	Home_Shot	5127	20	Away_Shot	4107
6	Home_Keeper	3273	21	Away_Keeper	3555
7	Home_Save	2208	22	Away_Save	2702
8	Home_Clear	11780	23	Away_Clear	14059
9	Home_Lose	16534	24	Away_Lose	16515
10	Home_Goal	597	25	Away_Goal	455
11	Home_Foul	4229	26	Away_Foul	4009
12	Home_Out_Throw	8982	27	Away_Out_Throw	8396
13	Home_Out_GK	3084	28	Away_Out_GK	3697
14	Home_Out_Corner	2321	29	Away_Out_Corner	1779
15	Home_Pass_O	814	30	Away_Pass_O	768
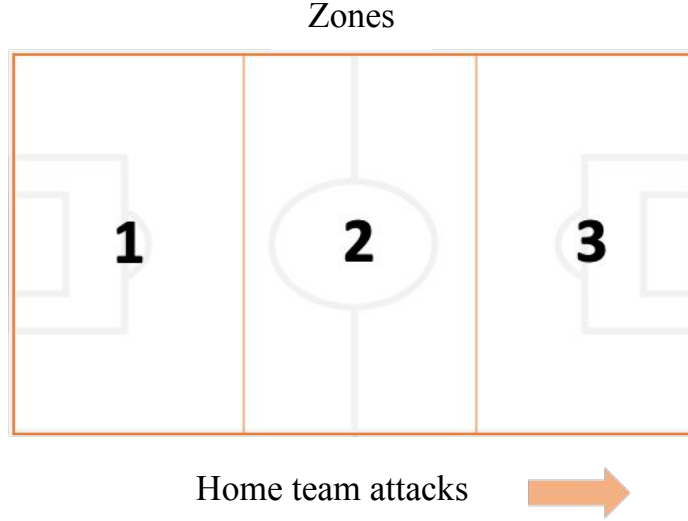
Zones

Figure 3: Mapping from event location in (x, y) coordinates to zones.

in this paper allows us to quantify home advantage by, for example, learning the relative ability of each team to retain possession when playing at home compared to away. (see, Section 6.6).

2.2 Data preparation

We combine the types and outcomes of touch-ball events into in-play and terminal composite events. The in-play composite events are Win, Dribble, Successful Pass (Pass_S), Unsuccessful Pass (Pass_U), Shot, Keeper, Save, Clear and Lose. Win denotes a player regaining possession of the ball from the opponent. Dribble is taking the ball forward with repeated slight touches. Passes are deemed to be successful when the ball is received by a teammate and unsuccessful otherwise. Shots include all attempts on the opponents' goal, including those missing the target. The Keeper event denotes the goal keeper taking possession of the ball into their hands by picking it up or claiming a cross. The Keeper event is unlike any other in-play event, as the goal keeper is allowed to hold the ball without being challenged for a period of time while waiting for opponents to clear the goal area. As a result, there is often a delay before the next event

Table 4: Snapshot of the final data prepared for modelling where each event, indexed by $i = 1, ..., n$, consists of the following components, the time of occurrence t_i, the zone z_i, and the mark m_i. The home and away team information for each game is assumed to be known and the first event (t_1, z_1, m_1) in each game period is considered to be deterministic and therefore, not modelled.

i	id	period	team_id	time (t_i)	zone (z_i)	mark (m_i)
1	101	1	1	0	2	18
2	101	1	1	1	2	19
3	101	1	2	3	1	8
4	101	1	1	6	3	16
5	101	1	1	8	3	18
6	101	1	1	15	2	18
7	101	1	1	16	1	19
8	101	1	2	19	1	12

even though the ball is technically in-play. Saves are events where the goal keeper prevents a shot from crossing the goal line. Clear events are those where a player moves the ball away from their goal area to safety while the Lose event is when a player loses possession of the ball. The terminal composite events are those which result in the ball going out-of-play and are Goal, Foul, Out_Throw, Out_GK, Out_Corner and Offside Pass (Pass_O). The terminal events interrupt the game resulting in a delay before play resumes. Each composite event is tracked for both the home and away teams. For this reason, we append "Home" or "Away" as a prefix to the event label to distinguish between the events of the two teams playing the game. This results in $M = 30$ distinct composite events, whose labels and observed frequencies are given in Table 3.

For each touch-ball event, the data also contains the associated (x, y) coordinates on the playing field. We partition the playing field into 3 zones of equal area. The zones and their corresponding labels are shown in Figure 3. Zone 1 is the region where the home team defends their goal, zone 3 is the region where the home team attacks, and zone 2 is the midfield region. It is natural to expect that the control a team has on the game depends on the zone the ball is at. For example, the home team is expected to retain possession of the ball more successfully in zone 1 as compared to zone 3.

Table 4 shows a snapshot of the data after its preparation, including a unique identifier for each game period in the data.

3 Marked point processes

3.1 Conditional intensity function

Sequences of events over time are conveniently represented as realisations of a point process. Oftentimes, the events can carry additional information, which are assumed to be realisations of random variables, referred to as marks. The collection of the times $\{t_i\}$ at which the events occur and the marks $\{m_i\}$ is a marked point process, whose ground process, is the processes for $\{t_i\}$ only.

A marked point process is typically specified through its joint conditional intensity function

$$\lambda^*(t, m) = \lambda_g^*(t) f^*(m \mid t), \tag{1}$$

where $\lambda_g^*(t)$ is the conditional intensity of the ground process and $f^*(m \mid t)$ is the conditional probability density or mass function of the mark m at time t. Both $\lambda_g^*(t)$ and $f^*(m \mid t)$ in (1)
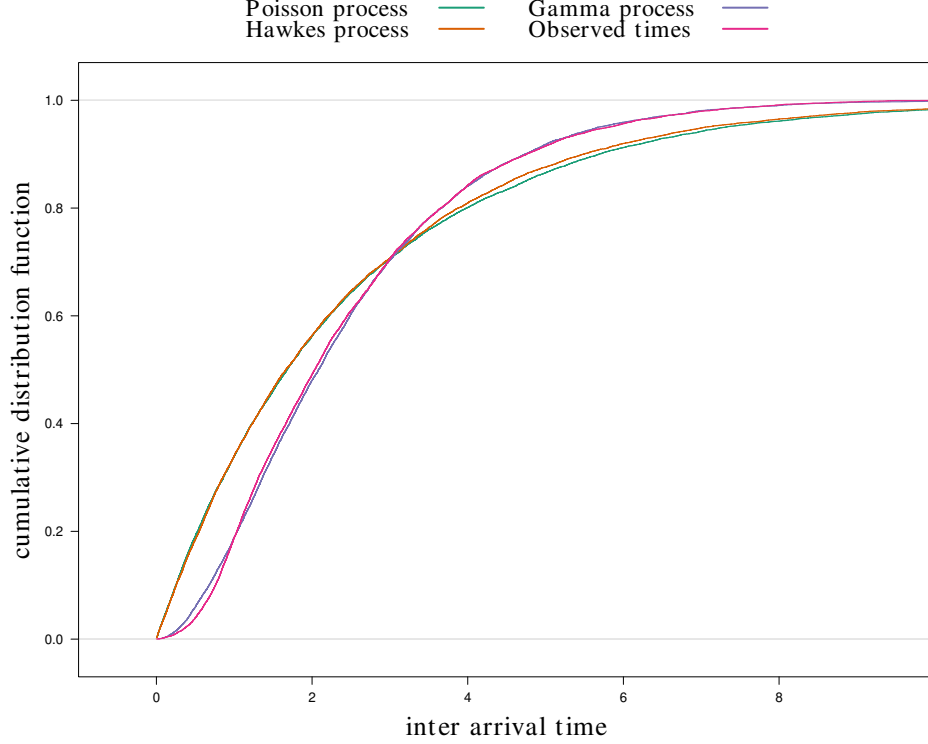
Figure 4: Comparing the cumulative distribution functions (CDFs) of the inter-arrival times of events simulated from a Poisson process (red), a Hawkes process (blue), a Gamma process (green) and observed event times in football (purple). Empirical CDFs were computed using 10,000 inter-arrival times in each case.

are understood as being conditional on $\mathcal{F}_{t^-}$, which is the filtration of the marked point process up to but not including t.

3.2 Marked Hawkes processes

Marked Hawkes processes are point processes whose defining characteristic is that they self-excite, meaning that each arrival increases the rate of future arrivals for a period of time. More formally, consider a realisation of a marked point process, consisting of event times $\{t_i\}$ with $t_i \in \mathbb{R}^+$ and $t_i > t_{i-1}$, and marks $m_i \in \{1, \ldots, M\}$ $(i = 1, \ldots, n)$, where M is the number of discrete marks. The marked Hawkes process is most intuitively specified using its mark dependent conditional intensity function $\lambda^*(t, m)$, which for an exponentially decaying intensity is (Rasmussen, 2013, expression 2.2),

$$\lambda^*(t, m) = \mu \delta_m + \sum_{t_j < t} \alpha \beta e^{-\beta(t - t_j)} \gamma_{m_j \to m} \,. \tag{2}$$

In (2), the parameter $\mu > 0$ is a constant background intensity and $\delta_m \in (0, 1)$ is the background mark probability for mark m with $\sum_{m=1}^{M} \delta_m = 1$. The parameter $\alpha \in (0, 1)$ is the excitation factor, $\beta > 0$ is the exponential decay rate and $\gamma_{m_j \to m} \in (0, 1)$ is the probability the excitation from an event of mark m_j triggers an event of mark m, with $\sum_{m=1}^{M} \gamma_{m_j \to m} = 1$ for any $m_j \in \{1, \ldots, M\}$.

9

3.3 Limitations of the marked Hawkes process model

The specification in (2) describes a marked Hawkes process that is linear in the sense that the excitations from different arrivals add up, not only increasing the probability of triggering an event of a particular type, but also concentrating the occurrence times for a certain period of time. For this reason, marked Hawkes processes have proven useful in a wide range of applications, where events tend to cluster in time, such as the modelling of earthquakes (Ogata, 1998), gang violence (Mohler et al., 2011) and financial market events (Bowsher, 2007).

However, in applications like the modelling of event sequences in football, each event triggers other events of a particular type with high probability, while it is not necessarily true that the occurrence times cluster.

Figure 4 shows the empirical distribution function of the first $10,000$ inter-arrival times from the first 7 games of the league season. We also plot the cumulative distribution functions of a homogeneous Poisson process and an unmarked Hawkes process, whose parameters are estimated from the aforementioned $10,000$ events using maximum likelihood. The fitted Hawkes process is far from the empirical distribution function, and almost identical to the fitted Poisson process confirming that the arrival times in football exhibit no clustering. This is evidence that Hawkes processes are not appropriate for modelling events such those observed in football, that tend not to cluster in time. Figure 4 also includes the cumulative distribution function of a fitted Gamma process, which, as is apparent provides an excellent fit to the observed inter-arrival times.

4 Specification of flexible marked point processes

4.1 Decoupling the modelling of times and marks

According to the decomposition of a multivariate distribution function in Cox (1975, expression 2) the likelihood of a marked point process can always be factorised as

$$\mathcal{L}(\mathcal{F}_{t_n} \mid \boldsymbol{\zeta}, \boldsymbol{\theta}) = \prod_{i=1}^{n} g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) \prod_{i=1}^{n} f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}), \tag{3}$$

where g and f are the conditional density and probability mass functions for the times and the marks, respectively, and $\boldsymbol{\zeta}, \boldsymbol{\theta}$ are unknown parameter vectors that may or may not share components. Therefore, an alternative approach to specify a marked point process is to specify the functions $g(\cdot \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta})$ and $f(\cdot \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta})$, separately, and combine them as in (3). The key insight in the current work is to derive the specification for the marks $f(\cdot \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta})$ from the joint conditional intensity function of a marked Hawkes process model, and then to specify a probability density function for the times $g(\cdot \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta})$ best suited to our application. In this way, we can restrict the characteristic excitation property of marked Hawkes processes exclusively to the modelling of the marks, and have the freedom to specify a different model for the occurrence times.

By the definition of the conditional intensity function for a marked point process in (1), $f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \lambda^*(t_i, m_i)/\sum_{m=1}^{M} \lambda^*(t_i, m)$. Plugging in $\lambda^*(t_i, m)$ from (2) in the latter expression, gives

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j < t_i} \alpha^* e^{-\beta(t_i - t_j)} \gamma_{m_j \to m_i}}{1 + \sum_{t_j < t_i} \alpha^* e^{-\beta(t_i - t_j)}}, \tag{4}$$

where $\alpha^* = \frac{\alpha \beta}{\mu}$. Expression (4) makes it immediately apparent, that the parameters μ and α of the marked Hawkes process as specified by (2) are not always identifiable for general specifications of $g(\cdot \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta})$ in (3). Apart from a mathematical fact, this is also rather

intuitive, because μ and α in (2) characterise the evolution of the Hawkes process in the time dimension and the sequence of marks is not sufficient to identify them.

The specification of the marked point process likelihood is complete once a probability density function $g(\cdot \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta})$ for the event times is specified.

We should highlight here that the marked point processes from the factorisation in (3) are generally different to the ones that result by assuming separability of the conditional intensity functions (see, for example, González et al., 2016, Section 6.5). A separable conditional intensity functions has the form

$$\lambda^*(t, m) = \lambda_g^*(t) f^*(m).$$

and implies that the conditional distribution of the mark does not depend on the occurrence time t. Separability is a convenient assumption because it allows for the sequence of marks to be modelled separately from the sequence of times. In contrast, the factorisation in (3) allows the conditional distribution of the mark to depend on the time of occurrence as well as the history, still allowing for estimating $\boldsymbol{\theta}$ separately from $\boldsymbol{\zeta}$, if $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ do not share components.

4.2 Parameter interpretation

In (4), the mark probability of each event in the sequence is determined by a combined additive effect from a background component and all previous occurrences. The first term δ_{m_i} in the numerator is the mark probability associated with the background component, while each term $\alpha^* \mathrm{e}^{-\beta(t_i - t_j)} \gamma_{m_j \to m_i}$ is the contribution from the excitation caused by a previous occurrence in the sequence.

The background mark probability $\delta_m \in (0, 1)$ is the probability an event has a mark m if the event is triggered solely by the background component. The excitation factor $\alpha^* \geq 0$ is a scaling factor applied to the contributions from the previous occurrences to the event mark probability. Large values of α^* indicate a stronger dependence of the process on its history, because the contributions from previous occurrences are weighted higher relative to the background component. The decay rate $\beta > 0$ is the exponential rate at which the excitations from previous occurrences decay over time. The parameter $\gamma_{m_j \to m_i} \in (0, 1)$ is the probability the excitation from an event of mark m_j triggers an event of mark m_i. In other words, $\gamma_{m_j \to m_i}$ can be viewed as the conversion rate for the transition from an event with mark m_j to an event with mark m_i.

In summary, as in marked Hawkes processes, the specification for the marks in (4) captures not only all cross-excitations between the various marks but also the rate at which these excitations decay over time.

4.3 Covariate-driven cross excitation

The conditional distribution of marks with probability mass function (4), allows to drive the cross-excitation of the marks using covariates. The conversion rates $\gamma_{m_j \to m}$ can be linked to a covariate vector $\boldsymbol{x} = (x_1, \dots, x_p)^\top$ observed at the current time through the baseline-category logit specification (see, for example, Agresti, 2007, Section 6.1)

$$\log\left(\frac{\gamma_{m_j \to m}}{\gamma_{m_j \to M}}\right) = \phi_{m_j \to m} + \boldsymbol{\omega}_m^\top \boldsymbol{x} \qquad (m = 1, \dots, M - 1), \tag{5}$$

where $\boldsymbol{\omega}_m$ is an unknown p-vectors of regression parameters. Then, keeping all covariates apart from x_t fixed, ω_{mt} is the log of the ratio of odds for category m versus the baseline category M at $x_t + 1$ to that at x_t $(t = 1, \dots, p)$. Also, by setting all covariates x_t equal to 0, $\phi_{m_j \to m_i}$ is the log of the ratio of odds for category m versus the baseline category M. The covariate vector $\boldsymbol{x}$ can include a combination of process-specific covariates that are time-invariant, and

event-specific covariates. For example, in Section 5.2, we use 5 to parameterise the marked point process in terms of the relative abilities of teams for each event type, and produce team rankings per event-type.

4.4 Spatio-temporal marked point processes

In applications like the modelling of in-game events in football, where there is a location associated with each event, we can readily extend the factorisation of the likelihood in (3) to include conditional densities for the event locations. We can write

$$\mathcal{L}(\mathcal{F}_{t_n} \mid \boldsymbol{\psi}) = \prod_{i=1}^{n} g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) \prod_{i=1}^{n} h(z_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\eta}) \prod_{i=1}^{n} f(m_i \mid t_i, z_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}), \qquad (6)$$

where $\{z_i\}$ is the collection of random variables corresponding to the spatial component of the process, which is characterised by the conditional probability mass or density functions $h(\cdot \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\eta})$ with $\boldsymbol{\eta}$ being a parameter vector that may or may not share parameters with $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}$, and $\boldsymbol{\psi} = (\boldsymbol{\zeta}^\top, \boldsymbol{\eta}^\top, \boldsymbol{\theta}^\top)^\top$. The filtration $\mathcal{F}_{t_i}$ now includes all times, marks and locations up to time t_i.

5 Bayesian modelling of in-game event sequences

5.1 Preamble

The framework for specifying flexible marked point processes of Section 4 is rather attractive for the modelling of in-game event sequences in football and other team sports. Firstly, cross-excitation of in-game events is a natural assumption because any event in an event sequence is likely to be triggered by one or more of the previous events. For example, following a corner kick, the next event is with high probability one among a shot on goal, a defensive clearance or a claim by the keeper. Such effects can be naturally and readily captured by the parameters of the conditional mark distribution in (4), which involves not only the magnitudes of all cross-excitations between the various event types but also the rate at which these excitations decay over time. Secondly, the preliminary analyses in Section 3.3 provides strong evidence that occurrence times do not necessarily cluster, as off-the-shelf marked Hawkes processes imply. Hence, the freedom to use a more flexible conditional distribution for the occurrence times, such as a Gamma process, is a rather attractive prospect. Furthermore, as discussed in Section 4.3, team information can be incorporated in the model in a direct way as covariate information to drive the cross-excitation based on the relative abilities of the teams.

   Overall, the framework of Section 4 can be used to provide valuable explanatory tools into the underlying dynamics of the game for the coaching staff and inform strategic decision making. It can also produce predictions of events, such as goals, in a specified time horizon, and of game outcomes that can enhance, among other things, the viewing experience in televised games.

5.2 Excitation-based models

Assume data from S realisations of independent spatio-temporal point processes, with the sth realisation involving n_s events. Denote by t_{si}, m_{si}, and z_{si} the occurrence time, mark and location of the i-th event in the sth realisation, respectively. Each of the S independent spatio-temporal marked point processes has a likelihood as in (6) with conditional probability mass function for the marks as in (4). The product of the S likelihoods is the overall likelihood for all S game periods.

The probability density functions for the occurrence times within each period are set to

$$g(t_{si} \mid \mathcal{F}_{st_{si-1}}; \boldsymbol{\zeta}) = p(t_{si} - t_{si-1} \mid m_{si-1}, \boldsymbol{a}, \boldsymbol{b})$$
$$t_{si} - t_{si-1} \mid m_{si-1}, \boldsymbol{a}, \boldsymbol{b} \sim \text{Gamma}(a_{m_{si-1}}, b_{m_{si-1}}), \tag{7}$$

where $\mathcal{F}_{st_{si}}$ denotes the filtration of the sth process up to time t_{si}.

By this specification, the time to next event is modelled using a gamma distribution with shape and rate parameters that are specific to the mark of the last observed event. In this way, we wish to capture the differences in the expected time to the next event across the different event types. For example, we expect a shorter time to the next event after an in-play event like a Pass, compared to that of an out-of-play event like a Foul. Even within the group of out-of-play events, we expect a shorter delay following a Throw-in as compared to a Goal event.

For a discrete set of locations $\{1, \dots, Z\}$, the conditional probability mass function for the current location is set to

$$h(z_{si} \mid t_{si}, \mathcal{F}_{st_{si-1}}; \boldsymbol{\eta}) = \eta_{(z_{si-1}, m_{si-1}) \to z_{si}}, \tag{8}$$

where $\eta_{(z_{si-1}, m_{si-1}) \to z_{si}}$ is the probability of transitioning into location z_{si} given the location z_{si-1} and the mark m_{si-1} of the last observed event. Expression (8) models the sequence of locations as a discrete first-order Markov chain (see, for example, Norris, 1997) with a transition probability matrix $\boldsymbol{\eta}$. The current state of the Markov chain is determined by the combination of the location and the mark of the last observed event, and the probability of transitioning into the next location depends only on the current state. The state space of the Markov chain is given by the Cartesian product $\{1, \dots, Z\} \times \{1, \dots, M\}$.

We consider four alternative parameterisations for the conditional probability mass function for the marks. The Sβ (scalar β) spatio-temporal marked point process results from (7), (8) and a conditional mark distribution of the form (4), that is

$$f(m_{si} \mid t_{si}, \mathcal{F}_{st_{si-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_{si}} + \sum_{t_{sj} < t_{si}} e^{\alpha - \beta(t_{si} - t_{sj})} \gamma_{m_{sj} \to m_{si}}}{1 + \sum_{t_{sj} < t_{si}} e^{\alpha - \beta(t_{si} - t_{sj})}}, \tag{9}$$

where $\alpha = \log(\alpha^*)$ to avoid the non-negative constraint on α^* in (4). The Vβ (vector β) model results from (7), (8) and

$$f(m_{si} \mid t_i, \mathcal{F}_{st_{si-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_{si}} + \sum_{t_{sj} < t_{si}} e^{\alpha - \beta_{m_{sj}}(t_{si} - t_{sj})} \gamma_{m_{sj} \to m_{si}}}{1 + \sum_{t_{sj} < t_{si}} e^{\alpha - \beta_{m_{sj}}(t_{si} - t_{sj})}}, \tag{10}$$

where β_m is the exponential decay rate of the excitation caused by an event of mark m. Vβ allows the decay rates to depend on the mark of the event causing the excitation. Hence, Sβ is formally nested in Vβ and results when $\beta = \beta_1 = \dots = \beta_M$. The M$\beta$ (matrix β) process results from (7), (8) and

$$f(m_{si} \mid t_{si}, z_{si}, \mathcal{F}_{st_{si-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_{si} \mid z_{si}} + \sum_{t_{sj} < t_{si}} e^{\alpha - \beta_{m_{sj} \to m_{si} \mid z_{si}}(t_{si} - t_{sj})} \gamma_{m_{sj} \to m_{si} \mid z_{si}}}{\sum_{m=1}^{M} \left[\delta_{m \mid z_{si}} + \sum_{t_{sj} < t_{si}} e^{\alpha - \beta_{m_{sj} \to m \mid z_{si}}(t_{si} - t_{sj})} \gamma_{m_{sj} \to m \mid z_{si}} \right]}, \tag{11}$$

where $\beta_{m \to m' \mid z}$ is the decay rate of the excitation caused by an event of mark m on an event of mark m' at location z. Specification (11) allows the decay rates to vary both with the pair of marks involved in the excitation and across locations, and allows the background mark probabilities $\boldsymbol{\delta}$ and event conversion rates $\boldsymbol{\gamma}$ to vary across location. The Mβ model can be used to account for scenarios like those where a Corner event excites a Pass event in the immediate future and a Shot event later on ($\beta_{\text{Corner} \to \text{Pass} \mid 3} > \beta_{\text{Corner} \to \text{Shot} \mid 3}$). It also allows us to capture effects such

13

as how a team is more likely to make more passes and retain possession of the ball in the defensive zone, while attempting more shots on goal in the attacking zone ($\gamma_{\mathsf{Pass}\to\mathsf{Pass}|1} > \gamma_{\mathsf{Pass}\to\mathsf{Pass}|3}$ and $\gamma_{\mathsf{Pass}\to\mathsf{Shot}|3} > \gamma_{\mathsf{Pass}\to\mathsf{Shot}|2}$). The final model we consider is the MβA (matrix β with abilities) where the baseline category logits of the conversion rates in (5) are driven by team information as

$$\log\left(\frac{\gamma_{m_{sj}\to m|z}(c)}{\gamma_{m_{sj}\to M|z}(c)}\right) = \phi_{m_{sj}\to m|z} + \omega_{cm} \quad (m = 1, \ldots, M-1; c = 1, \ldots, C). \tag{12}$$

In the above expression, $\phi_{m\to m'|z}$ is a location-dependent baseline conversion, and c is the team in possession of the ball attempting the event conversion. The parameter ω_{cm}, then, reflects the ability of a team to complete a conversion to an event of mark m.

5.3 Prior distributions

The shape and rate parameters of the Gamma distributions for the inter-arrival times in (7) are assigned independent exponential priors with rates a' and b', respectively. The probability mass function for the locations specified in (8), models the locations as a multinomial distribution given the current state of the Markov chain. The conjugate prior for the multinomial distribution is the Dirichlet distribution (see, for example, Gelman et al., 2013, Section 3.4) and therefore we assign a Dirichlet prior on the multinomial probabilities $\boldsymbol{\eta}$ with a common concentration rate parameter ν. The background mark probability vector $\boldsymbol{\delta}$ in the Sβ, Vβ, Mβ and MβA models is also assigned a Dirichlet prior with concentration hyper-parameter δ'. The location-specific mark probability vectors $(\delta_{1|1}, \ldots, \delta_{M|1})^\top, \ldots, (\delta_{1|Z}, \ldots, \delta_{M|Z})^\top$ in the Mβ and MβA models are assigned independent Dirichlet priors with concentration hyper-parameter δ'. The excitation factor α is assigned a normal prior with mean 0 and standard deviation σ_α. The decay rate parameter β in the Sβ model, the parameters $\beta_1, \ldots, \beta_M$ in the Vβ model, and their location-specific counterparts in the Mβ and MβA models are assigned independent exponential priors with a common rate β'. The parameters $\phi_{m\to m'|z}$ and ω_{cm} $(m, m' = 1, \ldots, M; z = 1, \ldots, Z; c = 1, \ldots, C)$ in the MβA model are assigned independent Normal priors with mean 0 and standard error σ_γ. The conversion rate parameters $(\gamma_{m\to 1}, \ldots, \gamma_{m\to M})^\top$ in the Sβ and Vβ models, and their location-specific counterparts in the Mβ model are assigned independent Dirichlet priors with a common concentration rate parameter γ'.

5.4 Posterior distributions

The time and location conditional distributions corresponding to (7) and (8) share no parameters with each other, and no parameters with any of the conditional mark distributions for the Sβ, Vβ, Mβ, and MβA models. Furthermore, the likelihood in (6) can be factorised into a term depending only on the time parameters $\boldsymbol{\zeta}$, a term depending only on the location parameters $\boldsymbol{\eta}$ and a term depending only on the mark parameters $\boldsymbol{\theta}$. Given that the priors for $\boldsymbol{\zeta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are also independent, the derivation of, or sampling from, the posterior distributions can be performed separately for each of those parameters.

The priors for the location parameters $\boldsymbol{\eta}$ are conjugate, so the posterior for $\boldsymbol{\eta}$ is readily obtained. If $\boldsymbol{y} = \{y_{i\to j}\}$, for $j \in \{1, \ldots, Z\}$, be the observed counts of transitions originating from the state i where $i \in \{1, \ldots, Z\} \times \{1, \ldots, M\}$, then the posterior distribution of each row of the transition matrix $\boldsymbol{\eta_i}$ is a Dirichlet distribution with concentration parameters $\boldsymbol{y_i} + \nu$.

Posterior sampling for the parameter vectors $\boldsymbol{a}$, $\boldsymbol{b}$ in (7) of the conditional distributions for the times, and the parameters $\boldsymbol{\theta}$ of the conditional distributions for the marks in each of the Sβ, Vβ, Mβ, and MβA models is carried out using the variant of the Hamiltonian Monte Carlo procedure (Duane et al., 1987) that is implemented in Stan (Stan Development Team, 2020).

We have also implemented posterior sampling using a Metropolis-within-Gibbs procedure, which, though, proved to mix poorly in artificial data sets for the Sβ, Vβ, Mβ, and MβA, rendering it computationally infeasible. As in the case of Hawkes process, the poor mixing stems from the presence of strong correlations between the model parameters as well as the flatness of the likelihood function (Veen and Schoenberg, 2008). Stan, on the other hand, implements the No-U-Turn Sampler (Hoffman and Gelman, 2014) that automatically calibrates tuning parameters in a warm-up phase and can efficiently sample from complex posterior distributions.

5.5 Model complexity

The conditional mark distribution in the Mβ and MβA models involves a large number of parameters. There are M^2Z decay rate parameters and $M(M-1)Z$ baseline conversion rate parameters which makes posterior sampling a computationally challenging task. We have developed a screening procedure based on association rule learning (see, for example, Agrawal et al., 1993) that operates on the data involved in the likelihood and eliminates parameters prior to posterior sampling.

The screening procedure retains only those parameters that capture the most significant event interactions and depends on two constants that need to be chosen. The first is the window size W for the number of transient events, and any event is allowed to be triggered by only one of the W events leading up to it. The other is the number of event pairs N considered in each of the three zones and sets a threshold on the number of significant event interactions that are identified. Full details on the association rule learning screening procedure are given in Section S2 of the Supporting Materials.

5.6 Model evaluation

Let $\boldsymbol{X}^{(train)}$ be the set of the training data on which the likelihood is based on, consisting of $n^{(train)}$ events, and let $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(R)}$ be R samples from the posterior distribution. Denote by $\boldsymbol{X}^{(test)}$ the set of held-out test data, consisting of $n^{(test)}$ events.

One method to evaluate the predictive accuracy of each model is to use the log point-wise predictive density (Vehtari et al., 2017) computed on the test data, using the posterior samples

$$\widehat{lpd} = \sum_{(t,z,m)\in\boldsymbol{X}^{(test)}} \log\left(\frac{1}{R}\sum_{r=1}^{R} L(t,z,m \mid \mathcal{F}_{t^-}, \boldsymbol{\zeta}^{(r)}, \boldsymbol{\eta}^{(r)}, \boldsymbol{\theta}^{(r)})\right), \tag{13}$$

where $L(t,z,m \mid \mathcal{F}_{t^-}, \boldsymbol{\zeta}^{(r)}, \boldsymbol{\eta}^{(r)}, \boldsymbol{\theta}^{(r)})$ is the likelihood of (t,z,m) given the filtration $\mathcal{F}_{t^-}$ at the posterior sample $\boldsymbol{\zeta}^{(r)}, \boldsymbol{\eta}^{(r)}, \boldsymbol{\theta}^{(r)}$. Large values of $\widehat{lpd}$ indicate better predictive accuracy.

Apart from Sβ, Vβ, Mβ, and MβA, we also evaluate the predictive accuracy of two simpler baseline models that do not include Hawkes-like excitation effects. The first baseline model, termed FOMC model, is based on the factorisation of the likelihood of marked spatio-temporal processes in expression (6) with models for the times and locations as in (7) and (8), respectively, but with the conditional probability mass function for the marks being a first-order Markov chain

$$f(m_i \mid t_i, z_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \theta_{(z_i, m_{i-1})\to m_i}. \tag{14}$$

In this specification, $\theta_{(z,m)\to m'}$ is the probability of the event mark m' given that last observed event has location z and mark m. The second baseline model, termed MSTHP, is the marked spatio-temporal homogeneous Poisson process (Daley and Vere-Jones 2003, Section 7.3), which has likelihood

$$\mathcal{L}^{(P)}(\boldsymbol{q} \mid \boldsymbol{\rho}) = \prod_{m=1}^{M}\prod_{z=1}^{Z} \rho_{mz}^{q_{mz}} \exp\left\{-T\rho_{mz}\right\}, \tag{15}$$

Table 5: Zone-wise frequencies for each event type in the training data.

| | Home | | | | | Away | | | |
| | mark | | zone | | | | mark | | zone | | |
m	label	1	2	3	m	label	1	2	3
1	Home_Win	236	257	41	16	Away_Win	25	204	301
2	Home_Dribble	17	96	96	17	Away_Dribble	76	65	19
3	Home_Pass_S	1699	4633	1658	18	Away_Pass_S	1427	4390	2030
4	Home_Pass_U	541	825	725	19	Away_Pass_U	542	811	702
5	Home_Shot	0	0	292	20	Away_Shot	193	2	0
6	Home_Keeper	155	0	0	21	Away_Keeper	0	0	192
7	Home_Save	106	0	0	22	Away_Save	0	0	149
8	Home_Clear	557	141	32	23	Away_Clear	27	124	660
9	Home_Lose	122	287	368	24	Away_Lose	323	349	142
10	Home_Goal	0	0	22	25	Away_Goal	20	0	0
11	Home_Foul	62	126	64	26	Away_Foul	46	112	69
12	Home_Out_Throw	97	184	163	27	Away_Out_Throw	143	173	110
13	Home_Out_GK	149	0	0	28	Away_Out_GK	0	0	220
14	Home_Out_Corner	0	0	112	29	Away_Out_Corner	76	0	0
15	Home_Pass_O	7	19	20	30	Away_Pass_O	13	11	5

Table 6: Posterior summaries and convergence diagnostics from 2000 posterior samples for selected parameters from the MβA model after screening with $(W, N) = (5, 100)$.

parameter	mean	sd	$\hat{R}$	$N^{(eff)}$	parameter	mean	sd	$\hat{R}$	$N^{(eff)}$
$\beta_{3\rightarrow3|1}$	0.52	0.04	1.00	1309.81	$\phi_{3\rightarrow4|2}$	1.81	0.41	1.01	540.41
$\beta_{27\rightarrow8|1}$	1.97	0.86	1.00	1953.81	$\phi_{3\rightarrow5|3}$	1.38	0.35	1.01	576.16
$\beta_{24\rightarrow1|2}$	1.51	0.09	1.00	2042.84	$\phi_{3\rightarrow10|3}$	-1.31	0.59	1.01	1098.83
$\beta_{3\rightarrow4|2}$	0.65	0.03	1.01	913.52	$\delta_{3|1}$	0.56	0.03	1.00	1805.80
$\beta_{3\rightarrow5|3}$	0.63	0.04	1.01	1933.43	$\delta_{3|2}$	0.24	0.08	1.00	1356.85
$\beta_{3\rightarrow10|3}$	0.81	0.24	1.01	882.31	$\delta_{3|3}$	0.03	0.01	1.00	2207.39
$\phi_{3\rightarrow3|1}$	1.70	0.50	1.01	792.46	α	6.30	0.09	1.01	866.96
$\phi_{27\rightarrow8|1}$	-0.74	0.87	1.00	2159.96	$\omega_{9,3}$	0.59	0.26	1.03	334.15
$\phi_{24\rightarrow1|2}$	3.29	0.37	1.02	539.25	$\omega_{10,3}$	0.67	0.25	1.01	341.24

where ρ_{mz} is the Poisson rate parameter and q_{mz} is the number of event occurrences for mark m at location z over a total observation time T in the data.

The FOMC and MSTHP models have conjugate prior distributions and therefore their posteriors are readily obtained. Details on those prior distributions and the derivation of their posterior distributions are given in Section S1 of the Supporting Materials.

6 Explanatory modelling

6.1 Training

Samples from the posterior distributions for the parameters of the Sβ, Vβ, Mβ, and MβA models of Section 5.2, and of the FOMC and MSTHP baseline models of Section 5.6 are obtained using all event sequences from the first 20 games of the league season, all games played between 17/08/2013 and 26/08/2013, which constitute $\boldsymbol{X}^{(train)}$. The training data involves $S = 40$ game periods involving of $27,660$ events. Each of the 20 teams participating in the league plays in two of the 20 games, one at their home and one at their away venue. As is also described in Section 2.1, there are $M = 30$ marks and $Z = 3$ zones. Table 5 gives the zone-wise event frequencies across marks in the training data used for modelling, reflecting the large variability
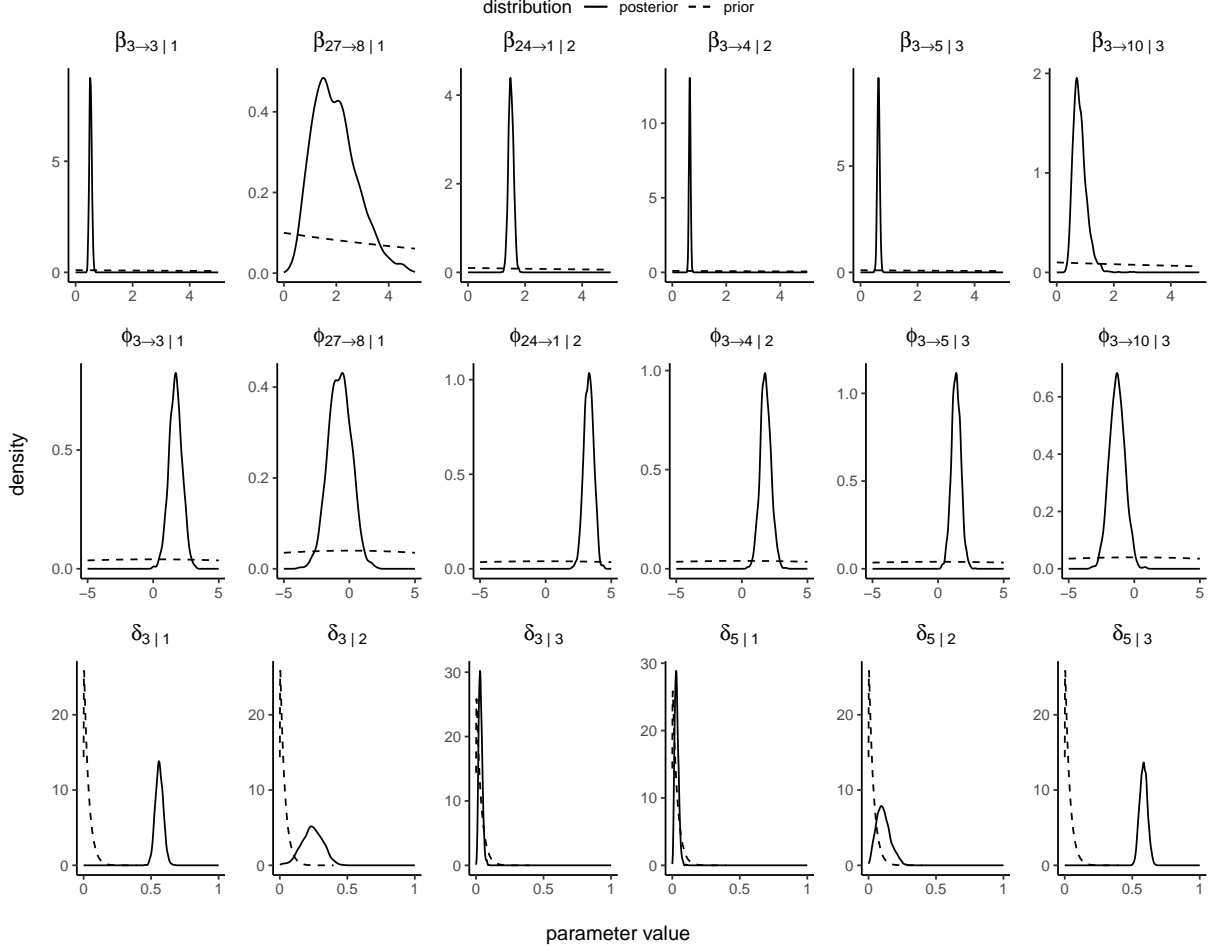
Figure 5: Visualising the impact of prior specifications by overlaying the posterior and prior densities for selected model parameters for the MβA model with $(W, N) = (5, 100)$.

in the frequencies both across marks and zones.

For the Mβ and MβA models, the association rule learning screening procedure of Section 5.5 is employed for all combinations of $W \in \{5, 10\}$ and $N \in \{50, 100\}$, to eliminate some of the model parameters and reduce the model complexity before posterior sampling.

The hyper-parameters for the prior distributions specified in Section 5.3 are as follows. The hyper-parameters a' and b' for the exponential priors on the parameters of the Gamma distributions in (7) are both set to 0.01. The Dirichlet prior on the background mark probabilities $\boldsymbol{\delta}$ has concentration hyper-parameter $\delta' = 1$. The exponential prior on the decay rates $\boldsymbol{\beta}$ has a rate hyper-parameter $\beta' = 0.1$. The Normal priors on the excitation factor α and the baseline-category logit model parameters $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ have hyper-parameters $\sigma_\alpha, \sigma_\gamma = 10$.

The ability parameters in the baseline category logit specification for the conversion rates of the MβA are not directly identifiable. In order to make them so, we set the abilities ω_{cm} for West Ham United to 0 ($m = 1, \ldots, 30$). Then, $\omega_{cm} > 0$ indicates that for team c, a previous event is more likely to trigger an event of mark m when compared to the reference team.

Samples from the posterior distributions are obtained by running four parallel chains using the Hamiltonian Monte Carlo procedures implemented in Stan. The Stan templates we used are all provided in the Supporting Materials. Each chain is initialised with different starting values and run for a total of 500 iterations post the warm-up phase. Table 6 gives posterior summaries along with convergence diagnostics for some of the parameters of the MβA model with $W = 5$

Table 7: Cumulative log posterior densities $\widehat{lpd}$ over 10 game periods in the test data for all fitted models along with the number of estimated parameters $d^{(par)}$ in each model. For the Mβ models, W is the number of transient events and N is the number of significant event pairs identified in the rule-based framework for reducing model complexity.

model	abbreviation	$d^{(par)}$	$\widehat{lpd}$
Homogeneous Poisson process (Baseline)	MSTHP	90	-35469.50
Matrix β $(W, N) = (10, 50)$	Mβ2	538	-22288.64
Matrix β $(W, N) = (5, 50)$	Mβ1	538	-22152.08
First order Markov chain (Baseline)	FOMC	870	-21898.31
Scalar β	Sβ	902	-21838.04
Vector β	Vβ	915	-21829.81
Matrix β with abilities $(W, N) = (5, 100)$	MβA	1539	-21599.81
Matrix β $(W, N) = (10, 100)$	Mβ4	988	-21496.56
Matrix β $(W, N) = (5, 100)$	Mβ3	988	-21342.57

and $N = 100$; the corresponding chain-wise trace plots are provided in the Supporting Materials.

The convergence of the algorithm is assessed using the potential scale reduction factor $\hat{R}$ proposed by Gelman et al. (1992), which is the ratio of the average variance within each chain to the variance of the aggregated samples across chains. If the chains have converged to the stationary distribution, the expected value of $\hat{R}$ is 1. All parameters have $\hat{R} < 1.1$, which, as recommended in Gelman et al. (1992), is evidence for convergence. Table 6 also gives the effective sample size (see, for example, Gelman et al., 2013, Section 11.5) for the samples for each of the posterior marginals, which indicate that the sampler returned samples with satisfactory autocorrelation. For some parameters the effective sample size is larger than the sample size due to negative autocorrelations. This, as pointed out in Vehtari et al. (2021), is a consequence of the HMC algorithm used in Stan being an antithetic Markov chain which has negative autocorrelations on odd lags. The impact of the prior distributions in Section 5.3 on the posterior samples is minimal as seen in Figure 5 indicating that the posterior distributions of the parameters have concentrated after accounting for the likelihood.

6.2 Model evaluation

The Sβ, Vβ, Mβ, and MβA models are compared with each other and with the FOMC and MSTHP baseline models of Section 5.6 in terms of their log point-wise predictive density (13) computed on test data. The test data $\boldsymbol{X}^{(test)}$ includes all events from the 5 games immediately following the games in the training data, played between 31/08/2013 and 01/09/2013. The test data involves $S = 10$ game periods involving of $27,660$ events.

Table 7 gives the log predictive densities $\widehat{lpd}$, summed over all the events in the 10 game periods in the test data. Table 7 also provides the number of parameters in each model as a measure of their complexity. The three top performing models are the Mβ model after screening with $(W, N) = (5, 100)$, followed by Mβ model after screening with $(W, N) = (10, 100)$, and MβA with $(W, N) = (5, 100)$. Notably, the Mβ model after screening with $(W, N) = (5, 100)$ performs the best among the list of fitted models, significantly outperforming also models of similar complexity, such as the Sβ, Vβ and FOMC models. The slightly poorer performance of the MβA model with $(W, N) = (5, 100)$ is most probably due to the fact that, in the training data, each team plays just one game at their home and one at their away venue. Nevertheless, in order to illustrate the full explanatory potential of the modelling framework in Section 4, we focus on inferences based on the posterior samples from the MβA model.

Table 8: Posterior means and standard deviations (in parenthesis) of the zone dependent background mark probabilities $\delta_{m|z}$ for $z \in \{1, 2, 3\}$ from the MβA model. The dots ($\cdot$) denote means and standard deviations less than 0.005.

mark label	1	2	3	mark label	1	2	3
Home_Win	.	.	.	Away_Win	.	.	.
Home_Dribble	.	0.01 (0.01)	.	Away_Dribble	.	0.02 (0.02)	.
Home_Pass_S	0.56 (0.03)	0.24 (0.08)	0.03 (0.01)	Away_Pass_S	0.03 (0.01)	0.11 (0.05)	0.58 (0.03)
Home_Pass_U	0.06 (0.01)	0.04 (0.02)	0.03 (0.01)	Away_Pass_U	0.02 (0.01)	0.07 (0.04)	0.05 (0.01)
Home_Shot	.	.	.	Away_Shot	.	.	.
Home_Keeper	0.04 (0.01)	.	.	Away_Keeper	.	.	0.05 (0.01)
Home_Save	.	.	.	Away_Save	.	.	.
Home_Clear	0.05 (0.02)	.	.	Away_Clear	.	.	.
Home_Lose	.	0.04 (0.02)	.	Away_Lose	0.02 (0.01)	0.07 (0.03)	0.06 (0.01)
Home_Goal	.	.	.	Away_Goal	.	.	.
Home_Foul	0.03 (0.01)	0.09 (0.04)	0.02 (0.01)	Away_Foul	0.02 (0.01)	0.06 (0.03)	.
Home_Out_Throw	.	.	.	Away_Out_Throw	.	.	.
Home_Out_GK	.	.	.	Away_Out_GK	.	.	.
Home_Out_Corner	.	.	.	Away_Out_Corner	.	.	.
Home_Pass_O	.	0.08 (0.02)	.	Away_Pass_O	.	0.05 (0.02)	.

6.3 Background mark probabilities

Table 8 gives the posterior means of the background probabilities $\delta_{m|z}$ for all marks $m \in \{1, \ldots, 30\}$ and all locations $z \in \{1, 2, 3\}$. The background mark probabilities for the home and away team events in zone 1 are almost equal to those in zone 3 for the away and home team events, respectively. This is as expected because the attacking zone for the home team is the defensive zone for the away team and vice-versa.

The similar probabilities for the home and away background mark probabilities could indicate that the background process of the game is not influenced by home advantage. To confirm this, we fit another MβA model after constraining all the corresponding home and away background mark probabilities to be equal, for example, $\delta_{\mathsf{Home\_Pass\_S}|1} = \delta_{\mathsf{Away\_Pass\_S}|3}$, $\delta_{\mathsf{Home\_Foul}|3} = \delta_{\mathsf{Away\_Foul}|1}$, $\delta_{\mathsf{Home\_Dribble}|2} = \delta_{\mathsf{Away\_Dribble}|2}$ and so on. The constrained MβA model has 45 fewer parameters to be estimated as compared to the full MβA model.

The formal method to test our hypothesis is to calculate the following Bayes factor, defined as the ratio of the marginal likelihood of the constrained MβA model to the marginal likelihood of the full MβA model. Then a Bayes factor greater than 1 would indicate that there is no evidence in favour of the full MβA model and therefore, the background mark probabilities do not capture home advantage. However, as a consequence of both MβA models being high dimensional (~ 1500), calculating their marginal likelihoods proved computationally infeasible.

As an alternative, for the constrained MβA model, we calculate its out-of-sample log predictive density on the same test data as carried out for all the other fitted models in Section 6.2. In fact, the constrained MβA model ($\widehat{lpd} = -21589.28$) turns out with better predictive performance than the full MβA model ($\widehat{lpd} = -21599.81$), supporting our claim that the background process of the game is not influenced by home advantage.

We also observe that the successful Pass events account for the majority of the background probability mass, while events like Shots and Goals have nearly zero probability. This sug-

Table 9: Posterior means and standard deviations (in parenthesis) of the event conversion probabilities $\gamma_{m_j \to m_i | z_i}$ corresponding to the location $z = 2$ from the MβA model for Manchester united. The $\gamma_{m \to m' | z}$'s are computed by setting the team identifier $c = 11$, (corresponding to Manchester United), for both the home as well as the away events. The highlighted cells (in bold) illustrate the superior ability of Manchester United when playing at home compared to away. The dots ($\cdot$) denote means and standard deviations less than 0.005.

	Home_Win	Home_Dribble	Home_Pass_S	Home_Pass_U	Home_Shot	Home_Keeper	Home_Save	Home_Clear	Home_Lose	Home_Goal	Home_Foul	Home_Out_Throw	Home_Out_GK	Home_Out_Corner	Home_Pass_O	Away_Win	Away_Dribble	Away_Pass_S	Away_Pass_U	Away_Shot	Away_Keeper	Away_Save	Away_Clear	Away_Lose	Away_Goal	Away_Foul	Away_Out_Throw	Away_Out_GK	Away_Out_Corner	Away_Pass_O
Home_Win	.	.	**0.41** (0.1)	0.03 (0.02)	.	.	.	.	0.01 (0.01)	.	.	.	.	.	.	0.01 (0.02)	.	.	0.03 (0.02)	.	.	.	.	0.02 (0.02)	.	.	.	0.44 (0.11)	.	.
Home_Dribble	.	.	**0.72** (0.14)	0.07 (0.05)	.	.	.	.	0.05 (0.03)	.	.	.	.	.	.	0.07 (0.09)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.08 (0.05)
Home_Pass_S	.	.	**0.8** (0.06)	0.08 (0.01)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.03 (0.04)	.	.	.	.	.	.	.	.	.	0.02 (0.03)	.	.
Home_Pass_U	0.02 (0.02)	.	0.02 (0.02)	0.01 (0.01)	.	.	0.02 (0.02)	.	.	0.02 (0.02)	.	.	0.17 (0.04)	0.22 (0.03)	0.07 (0.01)	.	.	.	0.15 (0.03)	.	.	0.23 (0.05)	.	.	.	.	.	.	.	.
Home_Shot	.	.	0.49 (0.24)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.31 (0.21)	.	.	.	.	.	.	.	.	.	.	.	.	0.2 (0.12)
Home_Keeper	.	.	0.4 (0.23)	0.15 (0.14)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.15 (0.15)	.	.	.	.	0.17 (0.16)	.	.	.	.	.	.	0.13 (0.07)
Home_Save	.	.	0.4 (0.23)	.	.	.	.	.	.	.	.	.	.	.	.	0.4 (0.23)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.2 (0.12)
Home_Clear	.	.	0.04 (0.03)	0.02 (0.01)	.	.	.	.	0.04 (0.03)	.	.	.	.	.	.	0.02 (0.01)	0.13 (0.03)	0.04 (0.02)	.	.	.	.	0.02 (0.02)	.	.	0.64 (0.09)	.	.	.	.
Home_Lose	0.02 (0.02)	.	0.02 (0.02)	0.01 (0.01)	.	.	0.02 (0.02)	.	0.02 (0.02)	.	.	.	.	.	.	0.59 (0.08)	0.07 (0.02)	.	.	.	.	.	.	0.1 (0.03)	.	0.1 (0.04)	.	.	.	.
Home_Goal	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.55 (0.22)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.45 (0.22)
Home_Foul	.	.	0.48 (0.15)	0.24 (0.11)	.	.	.	.	.	.	.	.	.	.	.	.	0.09 (0.09)	.	.	.	.	.	.	.	.	.	.	.	.	0.19 (0.1)
Home_Out_Throw	.	.	0.86 (0.05)	0.09 (0.03)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Home_Out_GK	.	.	.	0.06 (0.05)	.	.	.	0.07 (0.09)	.	0.04 (0.05)	0.25 (0.2)	.	.	.	.	.	.	0.14 (0.13)	.	.	.	.	.	0.09 (0.06)	0.07 (0.08)	.	0.2 (0.18)	.	.	0.07 (0.03)
Home_Out_Corner	.	.	0.7 (0.18)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.3 (0.18)
Home_Pass_O	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.4 (0.19)	0.25 (0.15)	.	.	.	.	.	.	.	.	.	.	.	.	0.35 (0.16)
Away_Win	0.05 (0.04)	.	.	0.01 (0.01)	.	.	.	0.02 (0.02)	0.03 (0.02)	.	0.59 (0.1)	.	.	.	.	**0.22** (0.07)	0.02 (0.01)	.	.	.	.	.	.	.	.	.	.	.	.	0.03 (0.01)
Away_Dribble	0.31 (0.24)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**0.49** (0.23)	0.06 (0.05)	.	.	.	.	0.03 (0.03)	.	.	.	.	.	.	.	0.11 (0.06)
Away_Pass_S	0.03 (0.04)	.	0.06 (0.09)	0.02 (0.02)	.	.	.	0.02 (0.03)	0.02 (0.03)	.	0.06 (0.09)	.	.	.	.	.	**0.68** (0.11)	0.06 (0.01)	.	.	.	.	.	.	.	.	.	.	.	.
Away_Pass_U	0.25 (0.05)	.	0.2 (0.03)	0.07 (0.01)	.	.	.	0.14 (0.03)	.	.	0.28 (0.05)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Away_Shot	.	.	0.44 (0.22)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.36 (0.21)	.	.	.	.	.	.	.	.	.	.	.	.	0.2 (0.12)
Away_Keeper	.	.	.	0.16 (0.15)	.	.	.	0.19 (0.16)	.	.	.	.	.	.	.	.	0.42 (0.22)	0.11 (0.11)	.	.	.	.	.	.	.	.	.	.	.	0.13 (0.07)
Away_Save	.	.	0.47 (0.22)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.31 (0.2)	.	.	.	.	.	.	.	.	.	.	.	.	0.22 (0.13)
Away_Clear	0.02 (0.01)	.	0.15 (0.04)	0.05 (0.02)	.	.	.	0.02 (0.01)	.	.	0.69 (0.07)	.	.	.	.	0.03 (0.02)	.	.	.	.	.	.	.	.	.	.	.	0.01 (0.01)	.	.
Away_Lose	0.69 (0.06)	.	0.05 (0.01)	.	.	.	.	.	0.08 (0.02)	.	0.1 (0.03)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Away_Goal	.	.	0.73 (0.17)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.27 (0.17)
Away_Foul	.	.	.	0.13 (0.12)	.	.	.	.	.	.	.	.	.	.	.	0.5 (0.15)	0.18 (0.09)	.	.	.	.	.	.	.	.	.	.	.	.	0.18 (0.1)
Away_Out_Throw	0.02 (0.02)	.	0.02 (0.02)	.	.	.	.	0.01 (0.01)	.	.	0.01 (0.02)	.	.	.	.	0.86 (0.06)	0.05 (0.02)	.	.	.	.	.	.	.	.	.	.	.	.	.
Away_Out_GK	.	.	.	0.09 (0.08)	.	.	.	0.14 (0.08)	0.13 (0.12)	.	0.19 (0.14)	.	.	.	.	.	0.11 (0.08)	.	.	.	.	.	.	0.1 (0.09)	.	0.06 (0.06)	0.1 (0.09)	.	.	0.08 (0.03)
Away_Out_Corner	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.6 (0.2)	.	.	.	.	.	.	.	.	.	.	.	.	.	0.4 (0.2)
Away_Pass_O	.	.	0.74 (0.15)	0.1 (0.08)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.16 (0.11)

gests that the Shot and Goal events are highly unlikely to originate solely from the background component, but are instead triggered by excitations from previous events.

6.4 Excitation factor

The excitation factor α in expression (11) is a scaling factor applied to the contributions from the previous occurrences to the event mark probability. In (11), the background component has a weight of 1, while previous occurrences are weighted by $\exp(\alpha)$.

The 95% highest posterior density interval for $\exp(\alpha)$ is $(451.35, 642.54)$, meaning the contributions from previous occurrences carry substantially higher weight relative to the background component. In other words, this indicates that event sequences in football have a significant dependence on their history.
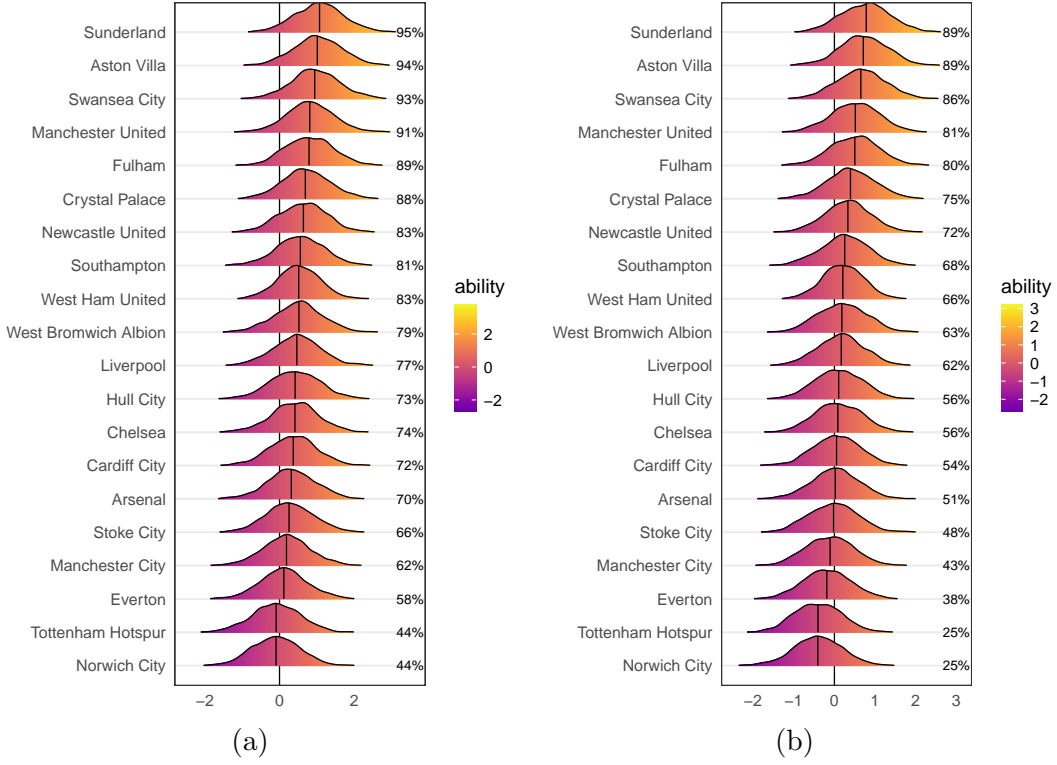
Figure 6: Posterior distribution of $\theta_{\mathsf{Home\_Win}\to\mathsf{Home\_Pass\_S}|2} + \omega_{c,\mathsf{Home\_Pass\_S}}$ - $\theta_{\mathsf{Away\_Win}\to\mathsf{Away\_Pass\_S}|2}$ - $\omega_{c,\mathsf{Away\_Pass\_S}}$ in (a) and $\theta_{\mathsf{Home\_Pass\_S}\to\mathsf{Home\_Pass\_S}|2} + \omega_{c,\mathsf{Home\_Pass\_S}}$ - $\theta_{\mathsf{Away\_Pass\_S}\to\mathsf{Away\_Pass\_S}|2}$ - $\omega_{c,\mathsf{Away\_Pass\_S}}$ in (b), from the baseline logit specification for incorporating team abilities in (12). Interpreted as the relative ability of a team to convert a Win to a Successful Pass when playing at home compared to away in (a) and similarly from one Successful Pass to another Successful Pass in (b). Teams are ranked in the decreasing order of the means of their respective posterior distributions shown by the overlayed vertical lines.

6.5 Decay rates

As mentioned in Section 4.2 the decay rate $\beta_{m\to m'|z}$ in expression (11) is the exponential decay rate of the excitation caused by an event of mark m on an event of mark m' at location z. By allowing the decay rates to depend on the pair of marks involved in the excitation, we had hoped to account for scenarios like a Corner event exciting a Pass_S event in the short term and a Shot event in the longer term. Indeed, the 95% highest posterior density interval for $\beta_{\mathsf{Home\_Corner}\to\mathsf{Home\_Pass\_S}|3}$ is $(1.34, 2.36)$ and $\beta_{\mathsf{Home\_Corner}\to\mathsf{Home\_Shot}|3}$ is $(0.16, 0.44)$ illustrating that the Corner $\to$ Shot excitation decays at a much slower rate compared to the Corner $\to$ Pass_S excitation.

6.6 Conversion rates

The parameter $\gamma_{m\to m'|z}$ in expression (11) is the probability the excitation from an event of mark m triggers an event of mark m' at location z.

Table 9 gives the posterior means and standard deviations of $\gamma_{m\to m'|z}$ in the midfield region ($z = 2$) for Manchester United. The probabilities for the Home_Win $\to$ Home_Pass_S, Home_Dribble $\to$ Home_Pass_S and Home_Pass_S $\to$ Home_Pass_S conversions are higher compared to their away team counterparts, indicating that the home team is superior in retaining possession of the ball.

Figure 6 provides a ridge-line plot of the log odds ratio for home versus away ability of a
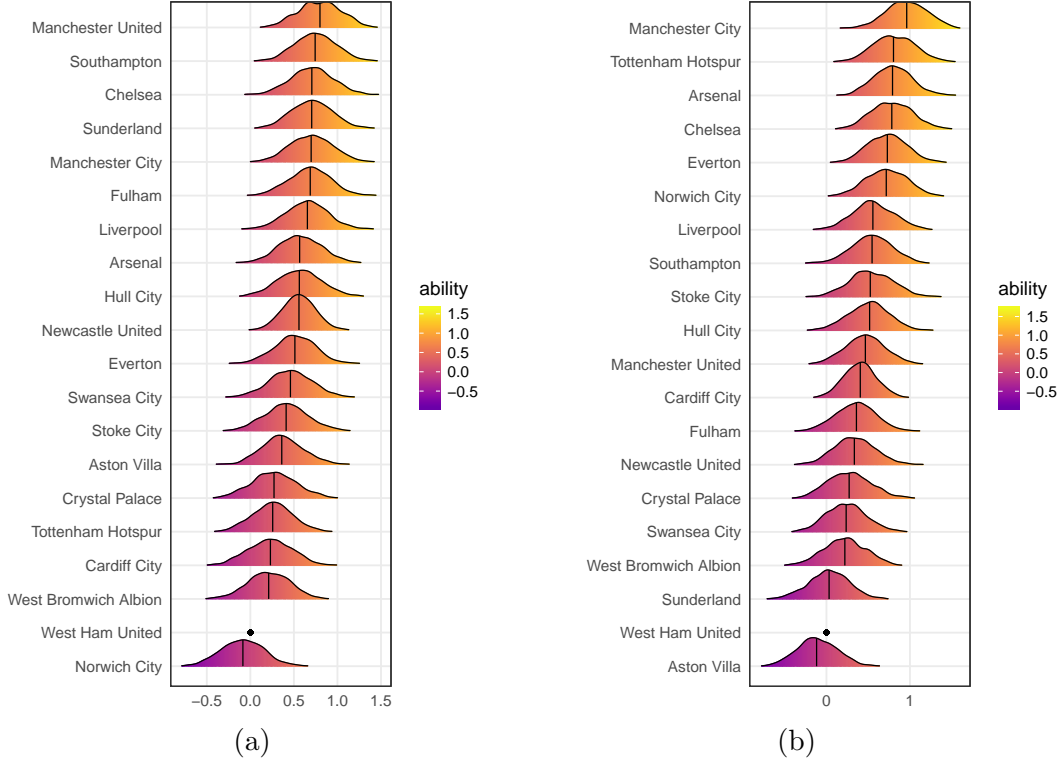
Figure 7: Posterior distribution of the parameters $\omega_{c,\mathsf{Home\_Pass\_S}}$ in (a) and $\omega_{c,\mathsf{Away\_Pass\_S}}$ in (b), from the baseline logit specification for incorporating team abilities in expression (12). Teams are ranked in the decreasing order of the means of their respective posterior distributions shown by the overlayed vertical lines.

team to convert a $\mathsf{Win} \rightarrow \mathsf{Pass\_S}$ (Figure 6a) and $\mathsf{Pass\_S} \rightarrow \mathsf{Pass\_S}$ (Figure 6b). The teams are listed in the decreasing order of the means of their respective posterior log odds ratios which are indicated by vertical lines. The percentage values by each plot, indicate the fraction of the distribution greater than 0. All but two teams in (Figure 6a) and five teams in (Figure 6b) have greater than 50% of their distribution greater than 0, confirming that the vast majority of teams possess a higher ability to retain possession while playing at home.

In this way, we not only confirm the well-known home advantage effect, but also quantify team performance for games played at home as well as away.

6.7 Team abilities

Figure 7 provides a ridge-line plot of the posterior distribution of the parameters $\omega_{c,\mathsf{Home\_Pass\_S}}$ and $\omega_{c,\mathsf{Away\_Pass\_S}}$. The teams are listed in the decreasing order of the means of their respective posterior distributions which are indicated by vertical lines. We observe that Manchester United, the team with the highest ability to retain possession in home games (Figure 7a), drop significantly down in the rankings for the away games (Figure 7b). This is evidence that when Manchester United plays away they seem to deviate from the possession-based strategy they seem to adopt in the home games.

Figure 8a provides a ridge-line plot of the posterior distribution of the cumulative ability of a team to attempt a shot on goal. A higher $\omega_{c,\mathsf{Home\_Shot}}$, for example, indicates that for the team c, an event like $\mathsf{Home\_Pass\_S}$ is more likely to trigger a $\mathsf{Home\_Shot}$. We do not expect the cumulative abilities $\omega_{c,\mathsf{Home\_Shot}} + \omega_{c,\mathsf{Away\_Shot}}$ of the dominant teams to be high, as they might prefer to make additional passes to create better goal scoring opportunities. A weaker team,
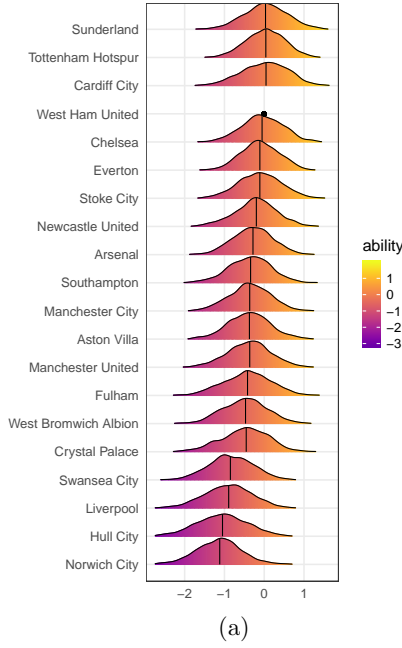
Team	Shots	Passes	S/P
Cardiff City	18	115	0.16
Sunderland	28	195	0.14
Tottenham Hotspur	37	261	0.14
Aston Villa	20	153	0.13
Newcastle United	22	180	0.12
Crystal Palace	18	148	0.12
Stoke City	24	199	0.12
Everton	40	342	0.12
Chelsea	29	248	0.12
Fulham	19	163	0.12
West Ham United	22	194	0.11
West Bromwich Albion	18	159	0.11
Swansea City	23	214	0.11
Arsenal	30	280	0.11
Southampton	23	223	0.10
Liverpool	28	292	0.10
Manchester United	22	238	0.09
Manchester City	30	330	0.09
Norwich City	19	227	0.08
Hull City	13	191	0.07

(a)   (b)

Figure 8: (a) Posterior distribution of $\omega_{c,\mathsf{Home\_Shot}} + \omega_{c,\mathsf{Away\_Shot}}$, the cumulative ability of a team c, relative to West Ham (baseline), to attempt a shot on goal. (b) The number of shots, passes completed in the attacking third and shots per pass completed in the attacking third (S/P) for each team in the training data.

on the other hand, typically has fewer opportunities to attack and therefore, is more likely to attempt a shot on goal when possible. Indeed, this is what we observe in Figure 8, where we compare the team rankings based on their cumulative ability $\omega_{c,\mathsf{Home\_Shot}} + \omega_{c,\mathsf{Away\_Shot}}$ with the number of shots per pass completed in the attacking third (S/P column in Figure 8b) in the training data. The comparison between Cardiff City and Norwich City is an interesting example of two teams that appear to be similar with 18 and 19 shots on goal attempted, respectively, in their two games in the training data. However, the two teams are at the opposite ends of the ranking based on their cumulative ability $\omega_{c,\mathsf{Home\_Shot}} + \omega_{c,\mathsf{Away\_Shot}}$, capturing the clear difference between their attacking styles.

Figure 9a shows the team rankings based on the cumulative ability to trigger five different event types. For example, the Pass column ranks teams in the decreasing order of their posterior means of $\omega_{c,\mathsf{Home\_Pass\_S}} + \omega_{c,\mathsf{Away\_Pass\_S}}$. The teams are ordered in Figure 9a by the rankings based on their cumulative passing ability. Despite training on just the first 20 out of 380 games of the 2013/14 season, the rankings based on the passing ability is a good predictor of the positions the teams finished in the final league table of the 2013/14 season in Figure 9b.

6.8 Event genealogy

The branching structure u_{si} indicates whether the ith event in sth sequence is an "immigrant" ($u_{si} = 0$) or an "offspring" of a previous event with index j ($u_{si} = j$). Given an observed event sequence $\mathcal{F}_{st_{sn_s}}$, the conditional branching structure probabilities $\mathbb{P}(u_{si} \mid \mathcal{F}_{st_{si}})$ based on the

Team	Pass	Shot	Goal	Win	Save
Manchester City	1	11	1	15	11
Chelsea	2	5	11	20	4
Arsenal	3	9	3	5	7
Southampton	4	10	8	7	19
Manchester United	5	13	4	2	18
Everton	6	6	17	11	14
Liverpool	7	18	12	9	5
Hull City	8	19	15	1	10
Tottenham Hotspur	9	2	14	3	6
Fulham	10	14	7	14	3
Stoke City	11	7	9	12	8
Newcastle United	12	8	19	16	17
Sunderland	13	1	13	8	2
Swansea City	14	17	18	17	12
Cardiff City	15	3	2	19	15
Norwich City	16	20	10	4	20
Crystal Palace	17	16	16	13	16
West Bromwich Albion	18	15	20	10	1
Aston Villa	19	12	5	6	13
West Ham United	20	4	6	18	9

(a)

1		Manchester City
2		Liverpool FC
3		Chelsea FC
4		Arsenal FC
5		Everton FC
6		Tottenham Hotspur
7		Manchester United
8		Southampton FC
9		Stoke City
10		Newcastle United
11		Crystal Palace
12		Swansea City
13		West Ham United
14		Sunderland AFC
15		Aston Villa
16		Hull City
17		West Bromwich Albion
18		Norwich City
19		Fulham FC
20		Cardiff City

(b)

Figure 9: (a) Team rankings based on the cumulative ability to trigger a particular event type. For example, the column Pass ranks teams in the decreasing order of their posterior means of $\omega_{c,\text{Home\_Pass\_S}} + \omega_{c,\text{Away\_Pass\_S}}$. The teams are ordered in (a) by the rankings based on their passing ability which the best predictor of the final position in the league table of the 2013/14 season in (b) from www.whoscored.com.
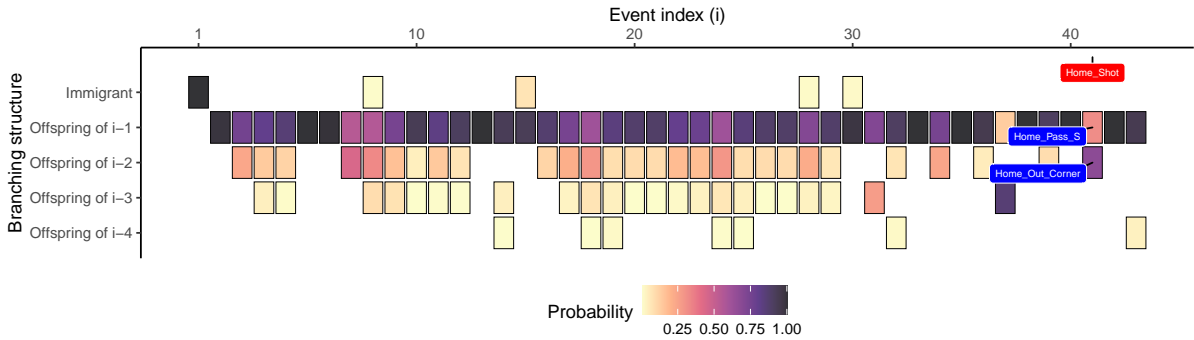

Figure 10: Branching structure probabilities for events in the first 4 minutes of the game between Chelsea and Hull City on 18/08/2013. The highlighted event Home_Shot has a higher probability of being an offspring of the event Home_Out_Corner than being an offspring of the more recent Home_Pass_S event.

model specification in expression (11) are

$$\mathbb{P}(u_{si} = 0 \mid \mathcal{F}_{st_{si}}) = \frac{\delta_{m_{si}|z_{si}}}{\delta_{m_{si}|z_{si}} + \sum_{t_{sk}<t_{si}} e^{\alpha - \beta_{m_{sk}\rightarrow m_{si}|z_{si}}(t_{si}-t_{sk})} \gamma_{m_{sk}\rightarrow m_{si}|z_{si}}},$$

$$\mathbb{P}(u_{si} = j \mid \mathcal{F}_{st_{si}}) = \begin{cases} \dfrac{e^{\alpha - \beta_{m_{sj}\rightarrow m_{si}|z_{si}}(t_{si}-t_{sj})} \gamma_{m_{sj}\rightarrow m_{si}|z_{si}}}{\delta_{m_{si}} + \sum_{t_{sk}<t_{si}} e^{\alpha - \beta_{m_{sk}\rightarrow m_{si}|z_{si}}(t_{si}-t_{sk})} \gamma_{m_{sk}\rightarrow m_{si}|z_{si}}} & \text{for } t_{sj} < t_{si} \\ 0 & \text{for } t_{sj} \geq t_{si} \end{cases}. \quad (16)$$

The branching structure probabilities in (16) quantify the relative contributions of the background and previous occurrences in the mark probability of the ith event in the sth sequence.
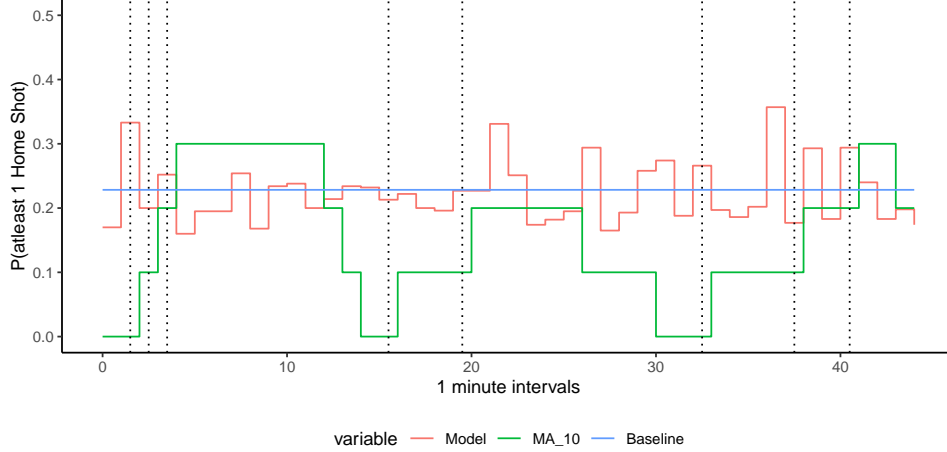
Figure 11: Forecasting the probability of observing at least one Home Shot event in 1-minute intervals over the first half of the game between Arsenal and Tottenham Hotspur (01/09/2013) in the test data. Intervals with observed Home Shot events are highlighted using dotted lines. MA_10 is a 10-step moving average model used as a benchmark for comparison.

Figure 10 shows the branching structure probabilities for all events in the first four minutes of the game between Chelsea and Hull City on 18/08/2013. To illustrate the flexibility of the model to account for dependence between events over arbitrary durations of time, we highlight the event Home_Shot showing a higher probability of being an offspring of the event Home_Out_Corner than being an offspring of the more recent Home_Pass_S event.

7 Model-based predictions

Finally, we illustrate how the mechanistic modelling framework presented in this paper can be used to simulate event sequences in football and obtain predictions of event probabilities in real-time. We split the first half of the game between Arsenal and Tottenham Hotspur (01/09/2013) in the test data into 1-minute intervals starting at time 0 before any events have occurred. For each interval, given the history of events up to but not including the interval, we simulate events over the next one minute $Q = 100$ times for each of the $R = 500$ posterior samples from the MβA model with the tuning parameter setting ($W = 5$, $N = 100$).

In Figure 11, we plot the proportion of all simulations within each interval where at least one Home_Shot event was simulated, and use dotted lines to denote the intervals where an Home_Shot event was actually observed. The homogeneous Poisson model (baseline) being memoryless predicts a constant probability over all intervals. We also include a 10-step moving average model (MA_10) as a benchmark for comparison. A quick inspection reveals that in 5 of the 8 intervals in which a Home_Shot is observed, the model predicts a shot probability greater than the baseline homogeneous Poisson model. The model also predicts a higher shot probability than the 10-step moving average model in all 8 intervals.

In Figure 12, we formally validate the performance of the model against three moving average models for the classification task of whether a shot will be observed in an interval. The area under the Receiver Operating Characteristic (ROC) curve is a performance measure that evaluates the performance of a classification model over all classification thresholds. The area under the curve (AUC) values are given in the legend and clearly confirm the superior performance of the model.
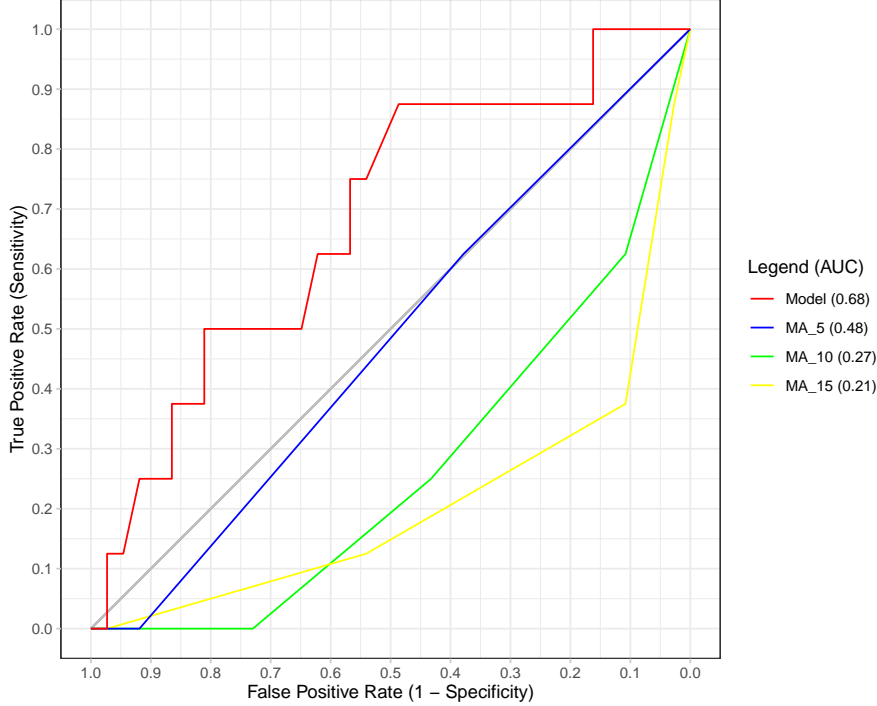
Figure 12: Validating the performance of the model against three moving average models for the classification task of whether a shot will be observed in an interval where MA_10 is a 10-step moving average model and so on. The ROC curve evaluates the performance of a classification model over all classification thresholds. The area under the curve (AUC) values in the legend clearly confirm the superior performance of the model.

8 Discussion and concluding remarks

Building on the decomposition of a multivariate distribution function, we showed how the joint modelling in classical point process models like Hawkes processes, can be decoupled. The introduced flexible modelling framework can, for example, retain the characteristic property of excitation in Hawkes processes in the model for the marks while avoiding the clustering of event times. A comprehensive Bayesian approach for the modelling of flexible marked spatio-temporal point processes was developed including an approach to evaluate the predictive accuracy of the fitted Bayesian models using the out-of-sample log predictive density.

We presented a case study showing how the flexible modelling framework developed in this paper can be tailored to separately model the components of the events in football, namely, the times, the locations and the event types. We were also able to incorporate team information into the model in a direct way that captured the relative abilities of the teams for each event type. We developed a method based on association rules to reduce the increased model complexity introduced by model extensions. The rule-based approach identified significant event interactions within sequences in a clever way by placing thresholds on some measures of significance. We then evaluated the accuracy of the excitation based models by comparing against two baseline models and confirmed the superior performance of the models with excitation effects.

We provided a detailed parameter description showing how the model parameters can be used to gain valuable insight into football. The excitation framework of the best performing model captured both the magnitudes and the durations of all pairwise event interactions across different locations. From the conversion rate parameters, we were able to quantify the well-known home advantage effect. We also discussed how the team ability parameters can be

used to obtain rankings for the teams by event type, that can be used as predictors for team performance. The team ability parameters also captured some interesting differences in the playing styles of the teams, that were not immediately apparent. In this way, the model along with its parameters can be used to develop a deeper understanding of the game-play by the coaching staff and inform strategic decision making. The proposed model can also be used to simulate the sequence of events in a game to obtain real-time predictions of event probabilities. We believe these predictions would enhance, among others, the viewing experience of televised games.

The methodology developed in this paper can be readily applied to other team sports like rugby, hockey, basketball etc. As none of the methods have been tailored specifically to football or even sports for that matter, they can also be applied to a wide range of applications that generate event data streams. Specifically, the ability parameters in the modelling framework can be used for other multi-agent environments, for example, accidents by car type, countries in the analyses of financial events, and so on.

9 Acknowledgments

References

Agrawal, R., T. Imieliundefinedski, and A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2), 207–216.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley.

Borrie, A., G. K. Jonsson, and M. S. Magnusson (2002). Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. *Journal of sports sciences 20*(10), 845–852.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics 141*(2), 876–912.

Clemente, F. M., M. S. Couceiro, F. M. L. Martins, and R. S. Mendes (2015). Using network metrics in soccer: a macro-analysis. *Journal of human kinetics 45*(1), 123–134.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Daley, D. J. and D. Vere-Jones (2003). *An introduction to the theory of point processes. Vol. I* (2nd ed.). New York: Springer-Verlag.

Decroos, T., L. Bransen, J. Van Haaren, and J. Davis (2018). Actions Speak Louder Than Goals: Valuing Player Actions in Soccer. arXiv:1802.07127.

Decroos, T., V. Dzyuba, J. V. Haaren, and J. Davis (2017). Predicting soccer highlights from spatio-temporal match event streams. In S. P. Singh and S. Markovitch (Eds.), *AAAI*, pp. 1302–1308.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns* (3rd ed.). Boca Raton, Florida: CRC Press.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics letters B 195*(2), 216–222.

Duch, J., J. S. Waitzman, and L. A. N. Amaral (2010). Quantifying the performance of individual players in a team activity. *PLOS One 5*(6), e10937+.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton, Florida: CRC Press.

Gelman, A., D. B. Rubin, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

González, J. A., F. J. Rodríguez-Cortés, O. Cronie, and J. Mateu (2016). Spatio-temporal point process statistics: a review. *Spatial Statistics 18*, 505–544.

Grund, T. U. (2012). Network structure and team performance: The case of english premier league soccer teams. *Social Networks 34*(4), 682–690.

Gudmundsson, J. and M. Horton (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR) 50*(2), 1–21.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika 58*(1), 83–90.

Hoffman, M. D. and A. Gelman (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Mackay, N. (2017). Predicting goal probabilities for possessions in football. Master's thesis, Vrije Universiteit Amsterdam.

Mohler, G. O., M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association 106*(493), 100–108.

Norris, J. R. (1997). *Markov Chains*. Cambridge: Cambridge University Press.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics 50*(2), 379–402.

Passos, P., K. Davids, D. Araújo, N. Paz, J. Minguéns, and J. Mendes (2011). Networks as a novel tool for studying team ball sports as complex social systems. *Journal of Science and Medicine in Sport 14*(2), 170–176.

Pena, J. L. and H. Touchette (2012). A network theory analysis of football strategies. arXiv:1206.6904.

Rasmussen, J. G. (2013). Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability 15*(3), 623–642.

Robberechts, P., J. Van Haaren, and J. Davis (2019). Who will win it? an in-game win probability model for football. arXiv:1906.05029.

Routley, K. and O. Schulte (2015). A markov game model for valuing player actions in ice hockey. In M. Meila and T. Heskes (Eds.), *UAI*, pp. 782–791.

Stan Development Team (2020). Cmdstan: the command-line interface to stan. Version 2.22.1.

Van Haaren, J., S. Hannosset, and J. Davis (2016). Strategy discovery in professional soccer match data. In *Proceedings of the KDD-16 Workshop on Large-Scale Sports Analytics*, pp. 1–4.

Veen, A. and F. P. Schoenberg (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association 103*(482), 614–624.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing 27*(5), 1413–1432.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, P.-C. Bürkner, et al. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*.

Wang, Q., H. Zhu, W. Hu, Z. Shen, and Y. Yao (2015). Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2197–2206.

Supporting materials for
Bayesian modelling of marked point processes with applications to event sequences from football
by

Santhosh Narayanan[1,4], Ioannis Kosmidis[1,4], and Petros Dellaportas[2,3,4]

[1]University of Warwick, Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
[2]Department of Statistical Science, University College London, Gower St. London WC1E 6BT, UK
[3]Department of Statistics, Athens University of Economics and Business, 76 Patission Str., Athens 10434, Greece
[4]The Alan Turing Institute

March 8, 2021

S1   Derivation of posterior distributions using conjugate priors

S1.1   Markov chain model for the locations

The probability mass function for the locations specified in expression (8) of the main text, models the locations as a multinomial distribution given the current state (defined by the location and the mark of the last observed event). Similar to the model for the inter-arrival times, the model for the locations is another component of the complete model specification in expression (6) of the main text. We are able to perform inference for this model separately as it does not share any parameters with the other components. Each row of the transition probability matrix η, corresponding to a single state, is a set of multinomial parameters, one for each location, that add up to 1.

Table S1: Observed transition counts $y_{i \to j}$ from the first 5 states to each zone in the training data.

state i		next zone j		
zone	mark label	1	2	3
1	Home_Win	195	38	1
1	Home_Dribble	12	5	0
1	Home_Pass_S	845	797	51
1	Home_Pass_U	75	304	160
1	Home_Shot	0	0	0

Let $\boldsymbol{y} = \{y_{i \to j}\}$, for $j \in \{1, \ldots, Z\}$, be the observed counts of transitions originating from the state i where $i \in \{1, \ldots, Z\} \times \{1, \ldots, M\}$. Table S1 gives the observed transition counts from the first 5 states in the training data. Out of a total of 90 states, 23 are never observed in the dataset, for example, it is nearly impossible for a Home_Shot event to occur in the defensive third (zone = 1) of the home team.

The likelihood of $\boldsymbol{y}_i$ given the multinomial probabilities $\boldsymbol{\eta}_i$ is

$$p(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i) \propto \prod_{j=1}^{Z} \eta_{i \to j}^{y_{i \to j}} \,,$$

where $\sum_{j=1}^{Z} \eta_{i \to j} = 1$. The conjugate prior for the multinomial distribution is the Dirichlet distribution (see, for example, Gelman et al., 2013, Section 3.4),

$$p(\boldsymbol{\eta}_i \mid \boldsymbol{\nu}_i) \propto \prod_{j=1}^{Z} \eta_{i \to j}^{\nu_{i \to j}-1} \,,$$

Table S2: Posterior means of the multinomial probabilities $\eta_{i \to j}$ for transitions from the first 5 states.

state i		next zone j		
zone	mark label	1	2	3
1	Home_Win	0.83	0.16	0.01
1	Home_Dribble	0.65	0.30	0.05
1	Home_Pass_S	0.50	0.47	0.03
1	Home_Pass_U	0.14	0.56	0.30
1	Home_Shot	0.33	0.33	0.33

where $\boldsymbol{\nu}_i > 0$ are the hyperparameters. The posterior distribution of $\boldsymbol{\eta}_i$ is therefore a Dirichlet with parameters $\boldsymbol{\nu}_i + \boldsymbol{y}_i$. To have a non-informative prior we set the hyperparameters $\boldsymbol{\nu}_i = \nu = 1$ and the resulting posterior means of the parameters $\eta_{i \to j}$ are given in Table S2.

S1.2 Baseline homogeneous Poisson process model

The likelihood for the homogeneous Poisson model for marked spatio-temporal data as specified in Section 5.6 of the main text is

$$\mathcal{L}^{(P)}(\boldsymbol{q} \mid \boldsymbol{\rho}) = \prod_{m=1}^{M} \prod_{z=1}^{Z} \rho_{mz}^{q_{mz}} \exp\left\{-T\rho_{mz}\right\} \,, \tag{S1}$$

where ρ_{mz} is the Poisson rate parameter and q_{mz} is the number of event occurrences for mark m at location z over a total observation time T in the data. Table 5 of the main text gives the observed counts $N_{m,z}$ in the training data. The conjugate prior for the Poisson process likelihood is a Gamma distribution

Table S3: Posterior means of the homogeneous Poisson rates $\rho_{m,z}$, for the first 5 marks in each zone.

mark		zone		
m	label	1	2	3
1	Home_Win	0.0035	0.0038	0.0006
2	Home_Dribble	0.0003	0.0014	0.0014
3	Home_Pass_S	0.0251	0.0683	0.0244
4	Home_Pass_U	0.0080	0.0122	0.0107
5	Home_Shot	0.0000	0.0000	0.0043

$$p(\boldsymbol{\rho} \mid \kappa, \tau) \propto \prod_{m=1}^{M} \prod_{z=1}^{Z} \rho_{m,z}^{\kappa-1} \exp\left(-\tau \rho_{m,z}\right) \,,$$

where $\kappa > 0$ and $\tau > 0$ are the hyperparameters for the shape and rate of the Gamma distribution respectively. Therefore, the posterior distribution of $\boldsymbol{r}$ is a Gamma distribution

$$\kappa' = \kappa + N_{m,z} \qquad \tau' = \tau + T \,,$$

where κ' and τ' are the updated hyperparameters. We set the values, $\kappa = 1$ and $\tau = 0$ that correspond to a non-informative prior.

The resulting posterior means of the Poisson rates $\rho_{m,z}$, for the first 5 marks in each zone, are given in Table S3. We use the rgamma function from the R package stats, which implements the method proposed by Ahrens and Dieter (1982), for simulating from a Gamma distribution.

Table S4: Transition counts $c_{i \rightarrow j}$ from the first 5 states to the first 5 marks in the training data. We abbreviate the prefix Home to H in the mark labels.

state i		label of next mark j				
mark label	zone	H_Win	H_Dribble	H_Pass_S	H_Pass_U	H_Shot
H_Win	1	0	0	80	25	0
H_Win	2	0	8	138	18	0
H_Win	3	0	4	28	11	4
H_Dribble	1	1	1	8	3	0
H_Dribble	2	0	5	39	11	0

Table S5: Posterior means of the multinomial parameters $\theta_{i \rightarrow j}$ corresponding to the first 5 states. We abbreviate the prefix Home to H in the mark labels.

state i		label of next mark j				
mark label	zone	H_Win	H_Dribble	H_Pass_S	H_Pass_U	H_Shot
H_Win	1	0.01	0.01	0.74	0.24	0.01
H_Win	2	0.01	0.05	0.82	0.11	0.01
H_Win	3	0.02	0.10	0.56	0.23	0.10
H_Dribble	1	0.11	0.11	0.50	0.22	0.06
H_Dribble	2	0.02	0.10	0.67	0.20	0.02

S1.3 Baseline Markov chain model for the marks

The probability mass function for the marks specified in expression (14) of the main text, models the marks as a multinomial distribution given the current state (defined by the current location and the mark of the last observed event). Each row of the transition probability matrix $\boldsymbol{\theta}$, corresponding to a single state, is a set of multinomial parameters, one for each mark, that add up to 1.

Similar to the model for locations in Section S1.1, let $\boldsymbol{c} = \{c_{i \rightarrow j}\}$, for $j \in \{1, \ldots, M\}$, be the count of observations of the transitions from the state i where $i \in \{1, \ldots, M\} \times \{1, \ldots, Z\}$. Table S4 gives the observed counts of transitions from the first 5 states in the training data.

The likelihood of $\boldsymbol{c}$ given the multinomial parameters $\boldsymbol{\theta}$ is

$$p(\boldsymbol{c}_i \mid \boldsymbol{\theta}_i) \propto \prod_{j=1}^{M} \theta_{i \rightarrow j}^{c_{i \rightarrow j}},$$

where the sum of the probabilities, $\sum_{j=1}^{M} \theta_{i \rightarrow j} = 1$. The conjugate prior for the multinomial distribution is the Dirichlet distribution,

$$p(\boldsymbol{\theta}_i \mid \mathbf{u}_i) \propto \prod_{j=1}^{M} \theta_{i \rightarrow j}^{u_{i \rightarrow j}-1},$$

where $\mathbf{u}_i > 0$ are the hyperparameters. The posterior distribution of $\boldsymbol{\theta}_i$ is therefore a Dirichlet with parameters $\mathbf{u}_i + \boldsymbol{c}_i$. We set $\mathbf{u}_i$ to 1 and the resulting posterior means of the parameters $\theta_{i \rightarrow j}$ corresponding to the first 5 states are given in Table S5.

S2 Dealing with model complexity

The conditional mark distribution in the $M\beta$ and $M\beta A$ models involves a large number of parameters. There are $M^2 Z$ decay rate parameters and $M(M-1)Z$ baseline conversion rate parameters which makes posterior sampling a computationally challenging task. We have developed a screening procedure that operates on the data involved in the likelihood and eliminates parameters prior to posterior sampling.

In the $M\beta$ model, the decay rate parameters $\boldsymbol{\beta}$ and conversion rate parameters $\boldsymbol{\gamma}$ capture the duration and magnitude of the excitation effects between all pairs of event types. However, it is reasonable to assume that the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are sparse, because the excitation effects between all event pairs are not equally significant. To be precise, we expect most elements of the $\boldsymbol{\beta}$ matrix to be infinite, meaning the corresponding excitations decay almost instantaneously. For the $\boldsymbol{\gamma}$ matrix, we expect most its values to be zero, meaning the corresponding event conversions have probability zero. For example, a successful Pass event by one team cannot significantly excite a Pass event for the opposite team, as this would make the commonplace occurrence of a string of passes by a single team very unlikely. If we are able to identify the most significant pairs of event interactions, we can thereby limit the number of elements within the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that we need to estimate.

S2.1 Association rule learning

Association rule learning is a method for discovering strong relationships between variables in large databases (see, for example, Agrawal et al., 1993). For example, the association rule Bread $\Rightarrow$ Butter identified from a supermarket sales database would indicate that if a customer buys bread, they are also likely to buy butter. The objective of association rule learning is to identify rules that are interesting based on some measure of significance.

S2.2 Definition for event sequences

Inspired by the original definition in Agrawal et al. (1993, Section 2), we define the problem of association rule learning in the context of event sequences as

Definition S2.1.

- *Let $A = \{1, \ldots, M\}$ be the set of M distinct event types.*

- *Let $B = \{b_{sn}\}$, where $b_{sn} \in A$ for $s \in \{1, \ldots, S\}$ and $n \in \{1, \ldots, N_s\}$, be the training data consisting of S event sequences with N_s number of observed events in the sequence s.*

- *Construct a database of subsequences $D = \{d_1, \ldots, d_C\}$, where $C = \sum_1^S N_s$, such that each event b in B has a corresponding subsequence of length $W + 1$ in D, made up of b and the W events preceding b.*

- *Each subsequence in D is denoted by $d_i = \{x_{i1}, \ldots, x_{iW}, y_i\}$, where $x_{ij}, y_i \in B$ for $i \in \{1, \ldots, C\}$ and $j \in \{1, \ldots, W\}$. We call $\{x_{i1}, \ldots, x_{iW}\}$ as the transient events of the subsequence before the terminal event y_i. Depending on W, the elements of the subsequence corresponding to the initial events of a sequence can be empty, because they have shorter histories.*

- *Given a set of event types A and a database of subsequences D, a rule is defined as an implication of the form: $x \Rightarrow y$, where $x, y \in A$. The association rule has the interpretation that the event type x is likely to be a transient event in subsequences terminating with event type y.*

Table S6: Support $P(x \cap y)$ for selected event pairs in the training data, where the rows denote the transient event x and columns are the terminal event y.

	Home_Win	Home_Dribble	Home_Pass_S	Home_Pass_U
Home_Win	0.0015	0.0027	0.0663	0.0124
Home_Dribble	0.0006	0.0008	0.0158	0.0030
Home_Pass_S	0.0111	0.0099	0.5925	0.0962
Home_Pass_U	0.0163	0.0026	0.1036	0.0289

In other words, the rule $x \Rightarrow y$, would indicate that the event type x excites the occurrence chance of an event with type y.

S2.3   Measures of significance

To identify interesting association rules, we place constraints on two measures of significance (Brin et al., 1997), namely support and lift.

S2.3.1   Support

The support of x with respect to a rule $x \Rightarrow y$ and a database D is defined as the proportion of subsequences d in the database which contain x as a transient event,

$$P(x) = \frac{|\{d \in D; x \in \mathsf{trans}(d)\}|}{|D|} \, , \tag{S2}$$

where $|\cdot|$ denotes the cardinality of a set and $\mathsf{trans}(d)$ is the set of transient events in the subsequence d. Similarly, the support of y with respect to a rule $x \Rightarrow y$ is defined as the proportion of subsequences d which terminate with y,

$$P(y) = \frac{|\{d \in D; y \in \mathsf{term}(d)\}|}{|D|} \, , \tag{S3}$$

where $\mathsf{term}(d)$ is the terminal event in the subsequence d.

The support of a rule $x \Rightarrow y$ is defined as, the proportion of subsequences d which contain x as a transient event and terminate in y,

$$P(x \cap y) = \frac{|\{d \in D; x \in \mathsf{trans}(d); y \in \mathsf{term}(d)\}|}{|D|} \, . \tag{S4}$$

Table S6 gives the support $P(x \cap y)$ for selected event pairs in the training data.

S2.3.2   Lift

The lift of a rule $x \Rightarrow y$ is defined as

$$\mathsf{lift}(x \Rightarrow y) = \frac{P(x \cap y)}{P(x) \cdot P(y)} \, . \tag{S5}$$

If the lift of a rule equals 1, it would indicate that the occurrence of y is independent of that of x. If the rule has lift > 1, then the event x excites the occurrence chance of y and lift < 1 indicates x inhibits the occurrence of y. Table S7 gives the $\mathsf{lift}(x \Rightarrow y)$ for selected event pairs in the training data.

We implement the following steps to place constraints on the lift and support measures and identify significant dependence between pairs of events.

Table S7: lift$(x \Rightarrow y)$ for selected event pairs in the training data, where the rows denote the transient event x and columns are the terminal event y.

	Home_Win	Home_Dribble	Home_Pass_S	Home_Pass_U
Home_Win	0.4141	2.2669	0.9793	0.9766
Home_Dribble	0.7176	2.7990	0.9588	0.9698
Home_Pass_S	0.3845	1.0141	1.0879	0.9450
Home_Pass_U	2.3031	1.0860	0.7782	1.1609

- Create a database of subsequences as defined in Definition S2.1, for $W = 5$ and $W = 10$, where W is the number of transient events in each subsequence.

- For each W, calculate lift for all event pairs and retain only those pairs that have lift > 1.

- Set a threshold on the support $P(x \cap y) > \epsilon$, such that when $\epsilon = \epsilon_1$ exactly $N = 50$ event pairs remain, and when $\epsilon = \epsilon_2$, $N = 100$ event pairs remain.

In this way, we select the specific elements of the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, corresponding to the identified significant event pairs, for parameter estimation. The elements of the matrices corresponding to the discarded event pairs are fixed, to the value 10^6 in the case of the decay rates $\boldsymbol{\beta}$, and 10^{-6} for the conversion rates $\boldsymbol{\gamma}$. A large value for the decay rate causes the excitation to die out almost instantaneously, and a very small value for the conversion rate makes the event conversion extremely unlikely. The results of evaluating four separate models, that are fitted based on the specific choices of the tuning parameters given above for the length of subsequence window W and the number of identified event pairs N, are discussed in Section 6.2 of the main text.

References

Agrawal, R., T. Imieliundefinedski, and A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2), 207–216.

Ahrens, J. H. and U. Dieter (1982). Generating gamma variates by a modified rejection technique. *Communications of the ACM 25*(1), 47–54.

Brin, S., R. Motwani, J. D. Ullman, and S. Tsur (1997). Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec. 26*(2), 255–264.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton, Florida: CRC Press.
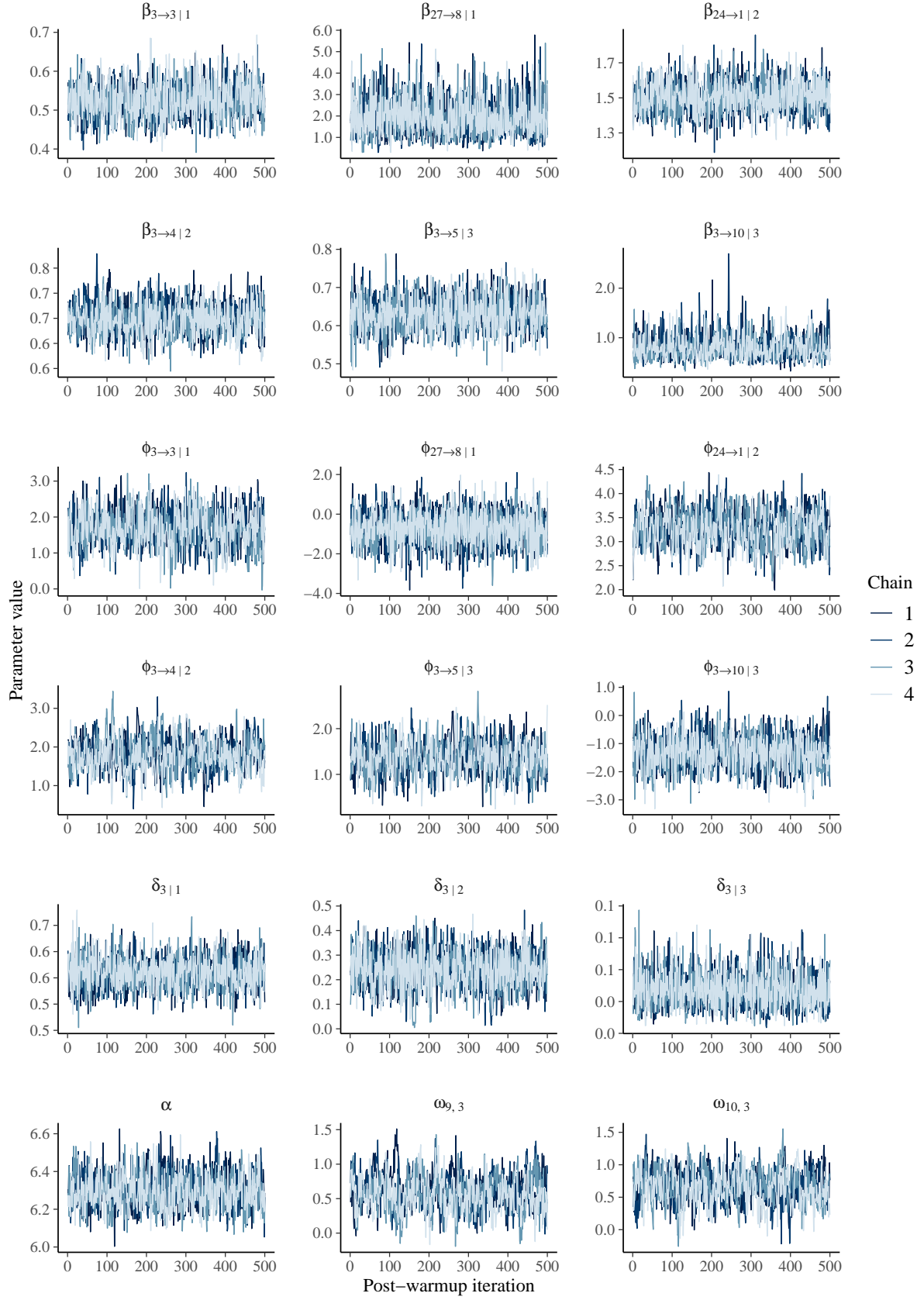
Figure S1: Trace plots.