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

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

[1]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, 96 Euston Road, London, England, NW1 2DB, UK

October 17, 2022

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 (9) 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 (7) 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 $\boldsymbol{\eta}$, corresponding to a single state, is a set of multinomial parameters, one for each location, that add up to 1.

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 S1: Observed transition counts $y_{i\to j}$ from the first 5 states to zones 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

Table S2: Posterior means of the multinomial transition probabilities $\eta_{i\to j}$ 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

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

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

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

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

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

Table S4: Transition counts $c_{i\to j}$ 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

S1.3 Baseline Markov chain model for the marks

The probability mass function for the marks specified in expression (15) 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 $\boldsymbol{c} = \{c_{i\to j}\}$, 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

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

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

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

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 $\theta_{i\to j}$ corresponding to the first 5 states are given in Table S5.

S2 Dealing with 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

Table S5: Posterior means of the multinomial parameters $\theta_{i \to j}$ 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

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 Mβ 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 ⇒ 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 $A = \{1, \dots, M\}$ be the set of M distinct event types.*
- *Let $B = \{b_{sn}\}$, where $b_{sn} \in A$ for $s \in \{1, \dots, S\}$ and $n \in \{1, \dots, N_s\}$, 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 $D = \{d_1, \dots, d_C\}$, 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 $d_i = \{x_{i1}, \dots, x_{iW}, y_i\}$, where $x_{ij}, y_i \in B$ for $i \in \{1, \dots, C\}$ and $j \in \{1, \dots, W\}$. We call $\{x_{i1}, \dots, x_{iW}\}$ as the transient events of the*

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

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.*

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,

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

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,

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

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,

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

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

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

S2.3.2 Lift

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

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

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 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.

- 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.

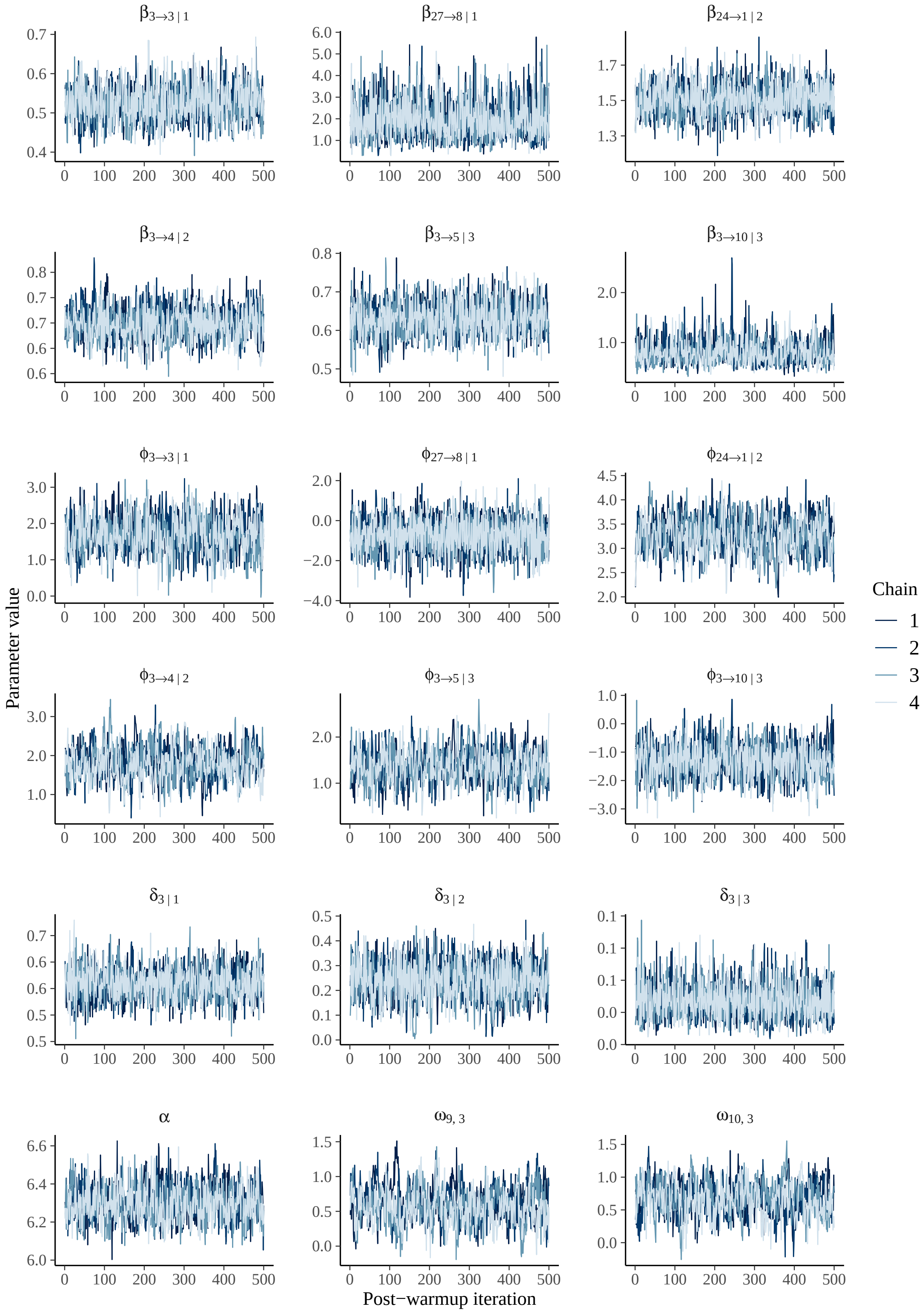

Figure S1: Chain-wise trace plots for some of the parameters of the MβA model with $W = 5$ and $N = 100$.