Real-time forecasting in football matches (and in other sports)

Seminar at Newcastle University

Soudeep Deb

Indian Institute of Management Bangalore
Bannerghatta Road, Bengaluru 560069, India.
Email: soudeep@iimb.ac.in.

December 5, 2025

Outline

Acknowledgements

Chinmay Divekar
PhD Student, IIM Bangalore.

Rishideep Roy
Lecturer, Univ of Essex

Kapil Gupta
Asst. Prof., IIM Kozhikode.

Saikat Sengupta
M.Stat student, ISI Kolkata

Outline

Motivation

- Types of sports:
  - Discrete events (cricket, baseball etc.)
  - Dynamic / Continuous events (football, basketball etc.)

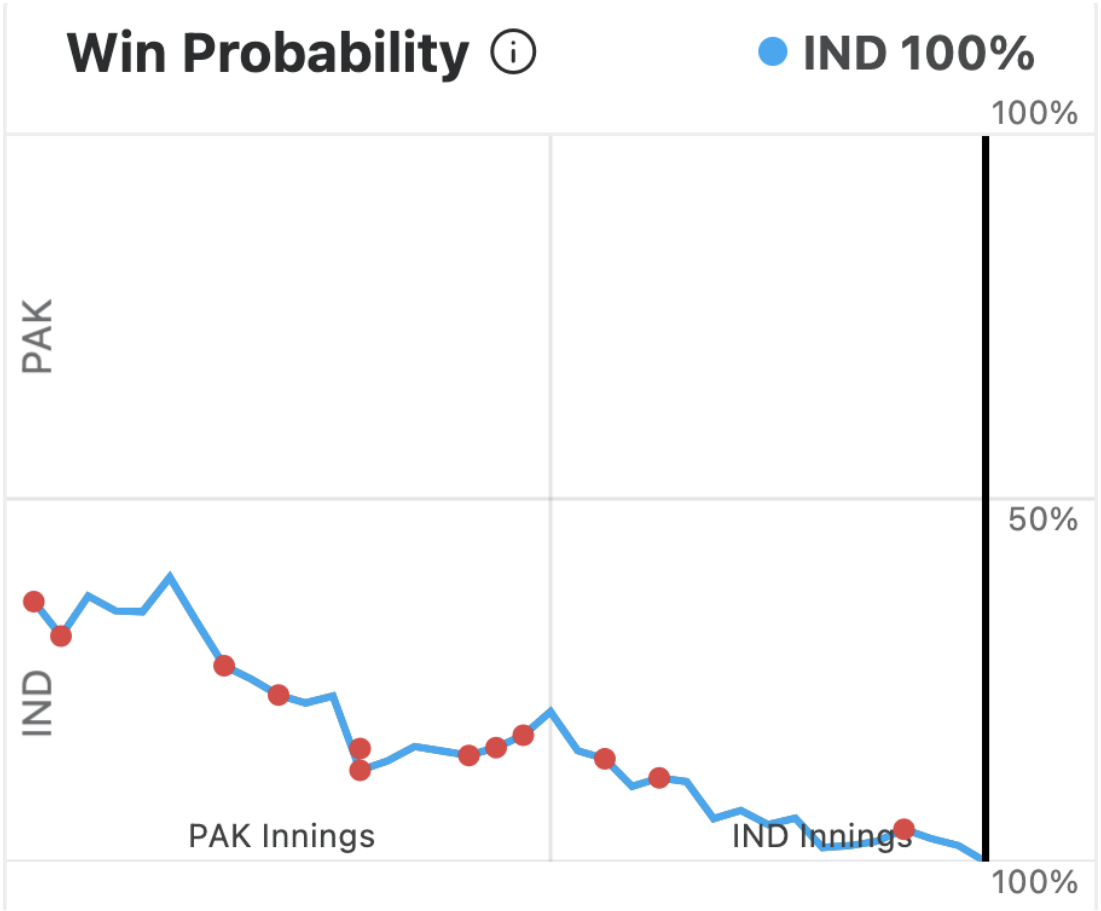
Motivation

- Types of sports:
  - Discrete events (cricket, baseball etc.)
  - Dynamic / Continuous events (football, basketball etc.)
- Need for within-game forecasting in sports:
  - Betting markets
  - Performance improvement and strategy building
  - Broadcasting, content optimization

Motivation

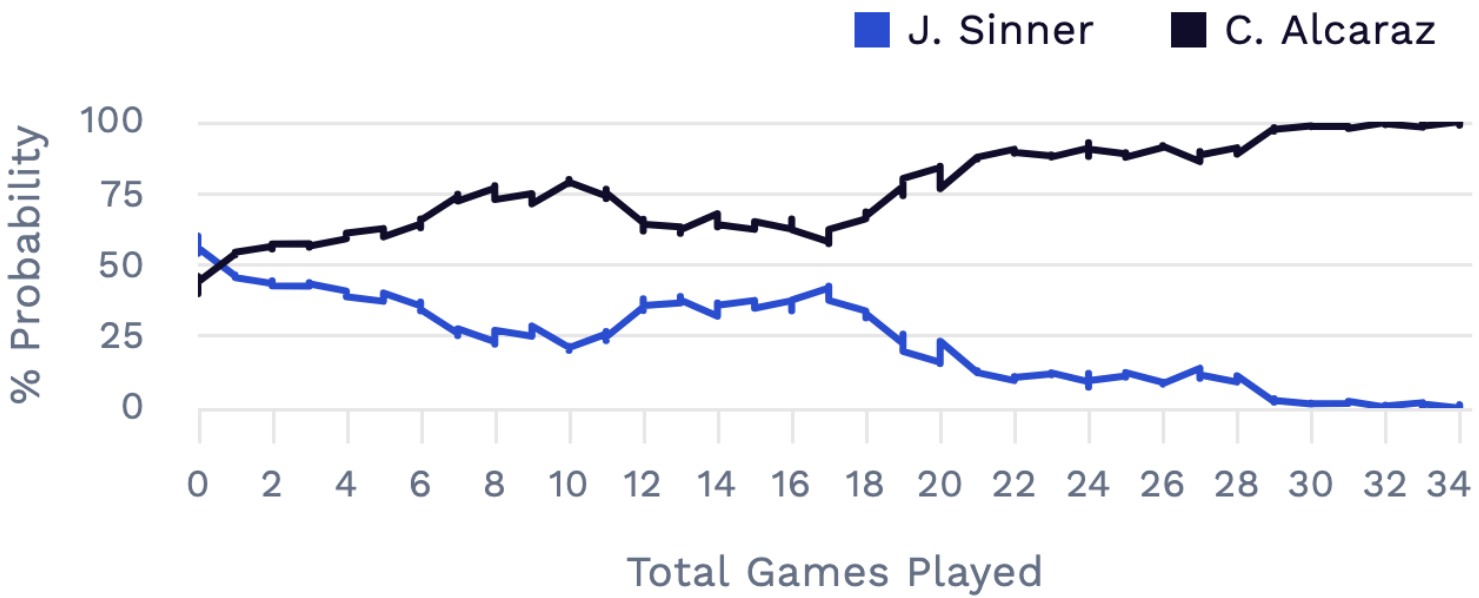
- Types of sports:
  - Discrete events (cricket, baseball etc.)
  - Dynamic / Continuous events (football, basketball etc.)
- Need for within-game forecasting in sports:
  - Betting markets
  - Performance improvement and strategy building
  - Broadcasting, content optimization
- Football is unique as a sport for within-game forecasting

Example of in-game forecasting in cricket



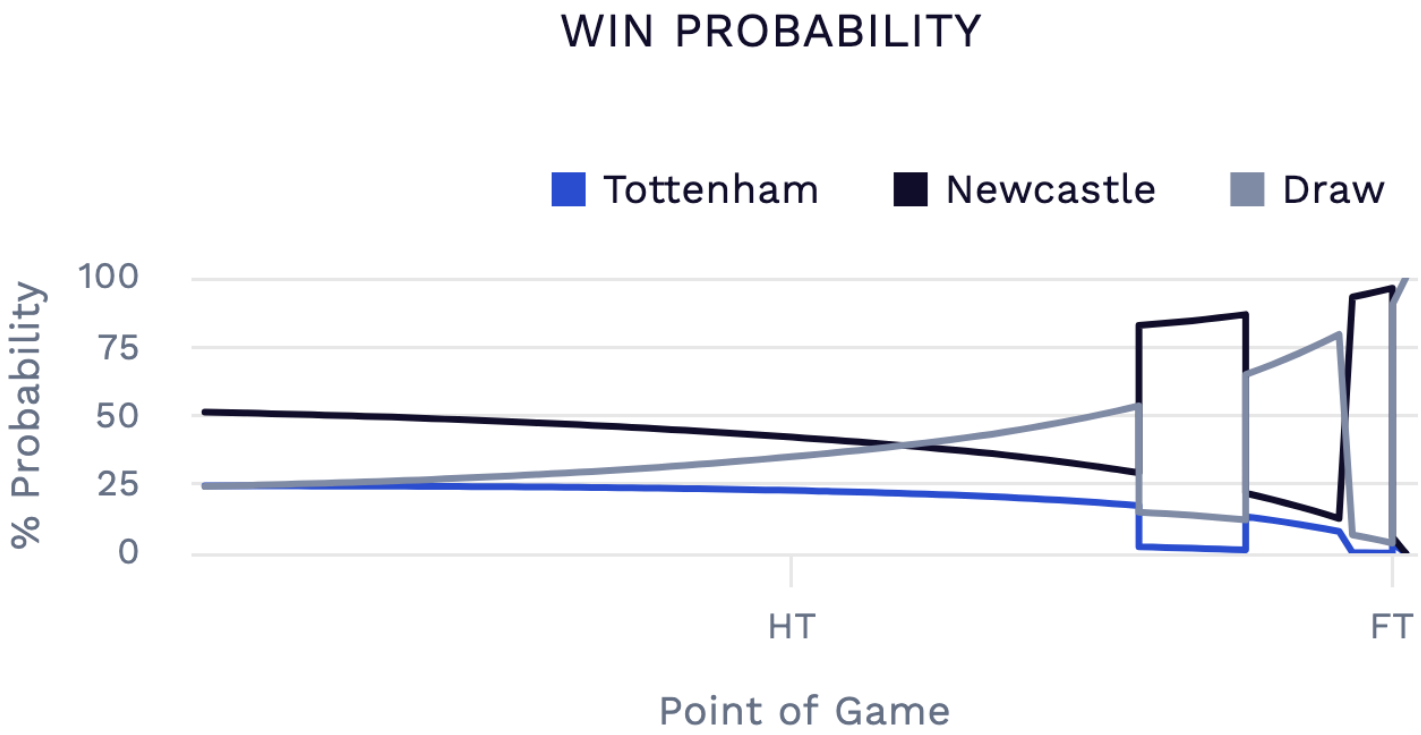
India vs Pakistan, Asia Cup 2025 (source: https://www.espncricinfo.com/)

Example of in-game forecasting in tennis



J Sinner vs C Alcaraz, US Open 2025 (source: https://www.dimers.com/)

Example of in-game forecasting in football



Tottenham vs Newcastle, EPL 2025 (source: https://www.dimers.com/)

Our contribution

- We develop a Bayesian latent variable model for analyzing and forecasting soccer match outcomes in real-time.
- Our method considers different types of events, not only goals.
- Time-varying effects of various events are estimated.
- The predictions are accompanied with a credible interval to reflect how confident one should be.
- Details can be seen in the paper by Divekar, Deb & Roy (2024).

Outline

Data

- The data is obtained for 3040 matches played over 8 seasons from 2008-2016 in the English Premier League.
- Along with goals, we consider shots-off and shots-on target, red and yellow cards, crosses, corners, fouls.
- Here's how the data may look:

	Event type k						
Game ID	1^{st} min (H)	1^{st} min (A)	2^{nd} min (H)	2^{nd} min (A)	...	90^{th} min (H)	90^{th} min (A)
1	0	1	2	0	...	2	1
2	2	0	0	0	...	1	0
			⋮				
3040	0	0	3	1	...	0	0

- A few time-independent covariates, such as team strengths before the match starts, are also used as features.

Proposed model (I)

- We model the outcome of a game from the perspective of the team playing at home.
- The dependent variable is an ordered multinomial random variable with three categories – loss, draw or win.
- The focus of the model is on forecasting the outcome in real-time, that is, after every minute of the match.

Proposed model (I)

- We model the outcome of a game from the perspective of the team playing at home.
- The dependent variable is an ordered multinomial random variable with three categories – loss, draw or win.
- The focus of the model is on forecasting the outcome in real-time, that is, after every minute of the match.
- Let Y_i denote the outcome for the home team in the i^{th} match. Then define a latent variable Π_i and cut-offs δ_1, δ_2 such that,

$$Y_i = \begin{cases} -1, & \text{if } \Pi_i < \delta_1 \\ 0, & \text{if } \delta_1 \leqslant \Pi_i \leqslant \delta_2 \\ 1, & \text{if } \Pi_i > \delta_2. \end{cases} \tag{1}$$

Proposed model (II)

- For the i^{th} match till the t^{th} time point we write

$$\Pi_i^{(t)} = \mathbf{z}_i^\top \boldsymbol{\gamma} + \sum_{k \in [K]} \mathbf{x}_{ik}^{(t)^\top} \boldsymbol{\beta}_k^{(t)} + \epsilon_i^{(t)}, \tag{2}$$

  where, $\mathbf{x}_{ik}^{(t)}$ is the set of time-dependent features, $\boldsymbol{z}_i$ is the set of time-independent features, and $\boldsymbol{\epsilon}^{(t)} \sim \mathcal{N}(0, \sigma_y^2 \boldsymbol{I}_n)$.
- The vector $\boldsymbol{\gamma}$ captures the effects of time-independent features, and $\boldsymbol{\beta}_k^{(t)}$ captures the time-varying effects of different types of events.
- We use 90 different models, one after every time point $t \in [1, 90]$.

Bayesian estimation

- The model is implemented via a complete Bayesian framework, using the principles of Gibbs sampling, Geweke statistic etc.
- For the computation, we specify Gaussian priors for γ and β_k as

$$\begin{aligned} \gamma &\sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p), \\ \beta_k^{(t)} &\sim \mathcal{N}_{2t}(\mathbf{0}, \boldsymbol{\Sigma}_k^{(t)}). \end{aligned} \tag{3}$$

- The structure of the covariance matrix $\boldsymbol{\Sigma}_k^{(t)}$ is defined through exponentially decaying function, i.e., correlation between the time-varying effects reduces at an exponential rate.
- We also use Gaussian priors for the cut-offs in the latent process:

$$\delta_j \sim \mathcal{N}(0, \tau^2),\ j \in \{1, 2\}. \tag{4}$$

Posterior computation (I)

- For estimation, we combine the covariates into a single matrix $\boldsymbol{M}$.
- Likewise, we denote the parameters by $\boldsymbol{\nu} = [\gamma \ \boldsymbol{\beta}_1 \ ... \ \boldsymbol{\beta}_K]$.
- Due to conjugacy in the priors we obtain Gaussian conditional posteriors for γ, β and Π:

$$\begin{aligned} \Pi_i \mid \boldsymbol{M}, \boldsymbol{\nu}, \boldsymbol{\delta} &\sim \mathcal{TN}(\boldsymbol{M}_i' \boldsymbol{\nu}, \sigma_y^2, \delta_{j-1}, \delta_j) \\ \boldsymbol{\nu} \mid \boldsymbol{M}, \Pi, \mathbf{y}, \boldsymbol{\delta} &\sim \mathcal{N}_{p+2Kt}\left(\boldsymbol{\mu}_\nu, \tilde{\boldsymbol{\Sigma}}\right) \end{aligned} \tag{5}$$

where $\tilde{\boldsymbol{\Sigma}} = (\sigma_y^{-2} \boldsymbol{M}' \boldsymbol{M} + \boldsymbol{\Sigma}_0^{-1})^{-1}$ and $\boldsymbol{\mu}_\nu = \sigma_y^{-2} \tilde{\boldsymbol{\Sigma}} (\boldsymbol{M}' \Pi)$.

Posterior computation (II)

- We establish a correspondence between the multinomial categories through a transformation of a Dirichlet(α_1,α_2,α_3) distribution to obtain,

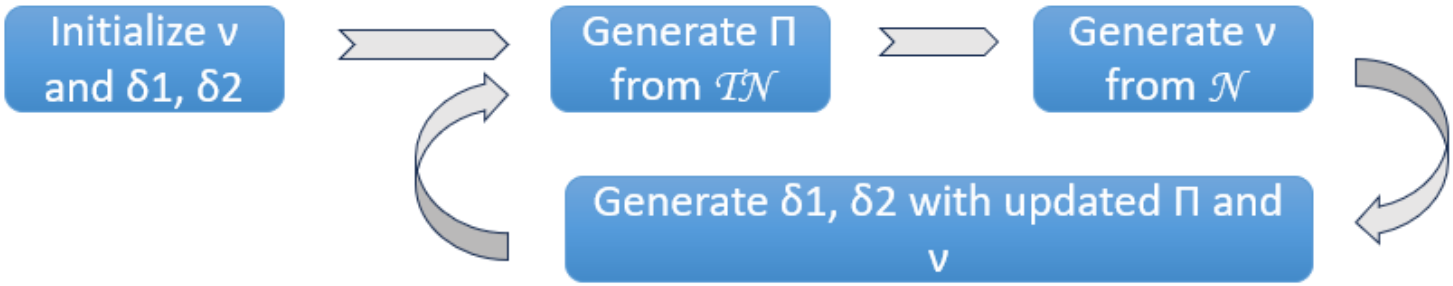
$$\begin{aligned}\Phi(\delta_1 \mid \delta_2, \boldsymbol{M}, \Pi, \mathbf{y}, \boldsymbol{\nu}) &\sim \Phi(\delta_2)\mathrm{Beta}(\alpha_1, \alpha_2),\\ \Phi(\delta_2 \mid \delta_1, \boldsymbol{M}, \Pi, \mathbf{y}, \boldsymbol{\nu}) &\sim [1-\Phi(\delta_1)]\mathrm{Beta}(\alpha_2, \alpha_3) + \Phi(\delta_1).\end{aligned} \tag{6}$$

where

$$\begin{aligned}\Phi(\max\{\Pi_i : Y_i = -1\}) \leqslant \Phi_{\delta_1} \leqslant \Phi(\min\{\Pi_i : Y_i = 0\}),\\ \Phi(\max\{\Pi_i : Y_i = 0\}) \leqslant \Phi_{\delta_2} \leqslant \Phi(\min\{\Pi_i : Y_i = 1\}).\end{aligned}$$

Posterior computation (III)

- Gibbs-sampling procedure is used to obtain the joint posterior distribution of the parameters from their conditional posteriors.



- To assess convergence, we use the Geweke statistic.
- We draw a sample for the posterior distribution after the chains converge, and utilize thinning in this step.

Predictive analysis

- We thus obtain $\hat{\Pi}_i^{(t)}$ and $\hat{Y}_i^{(t)}$ for every time point t.
- The probabilities of the different categories for the outcome variable, corresponding to the home team, can be calculated as

$$\begin{aligned} P\,(\text{Win}) &= 1 - \Phi_{\hat{\Pi}_i}(\hat{\delta}_2), \\ P\,(\text{Draw}) &= \Phi_{\hat{\Pi}_i}(\hat{\delta}_2) - \Phi_{\Pi_i}(\hat{\delta}_1), \\ P\,(\text{Loss}) &= \Phi_{\hat{\Pi}_i}(\hat{\delta}_1), \end{aligned} \tag{7}$$

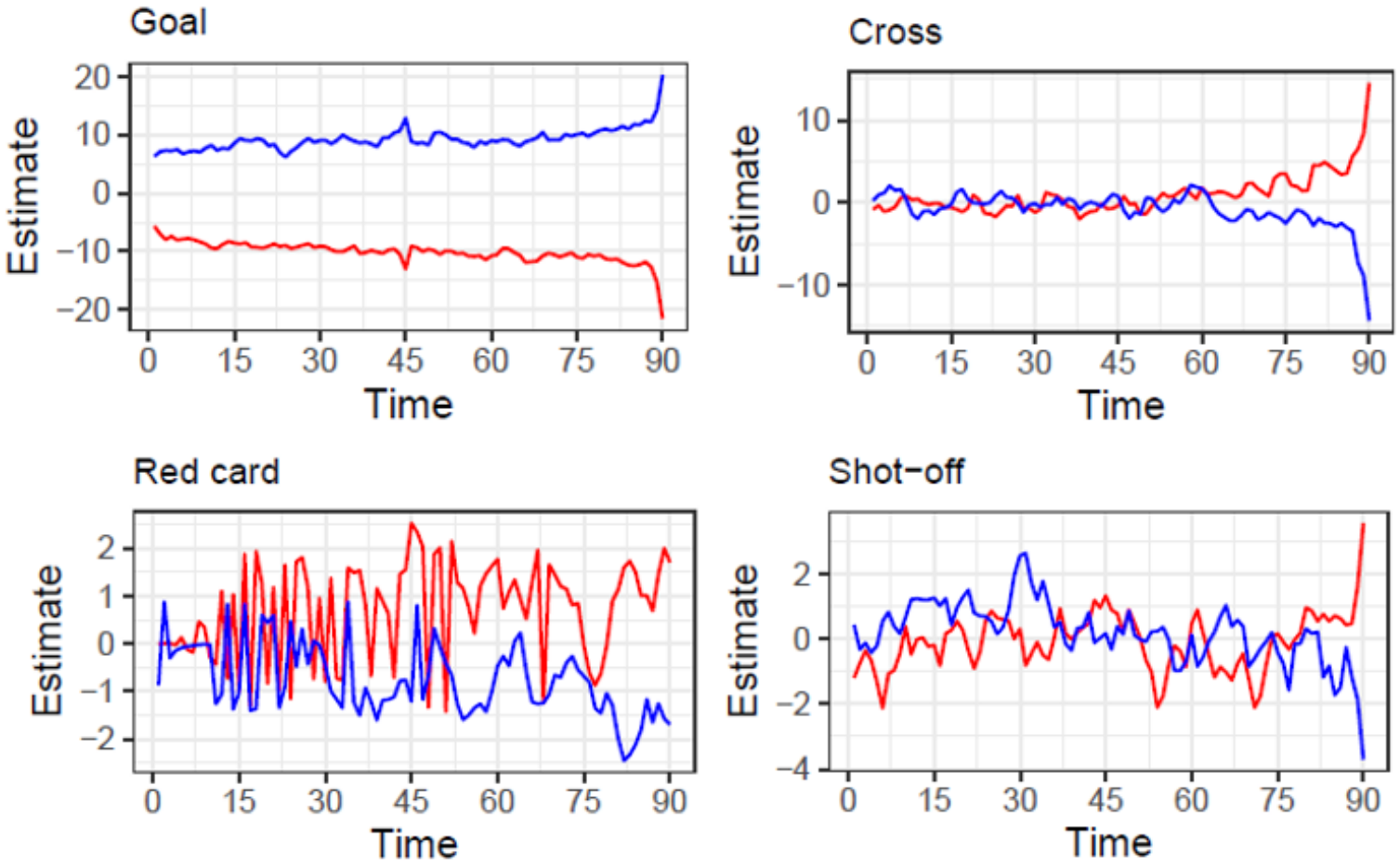
where $\Phi_{\hat{\Pi}_i}$ is the Gaussian cumulative distribution function of $\hat{\Pi}_i$.

Outline

Summary of the results

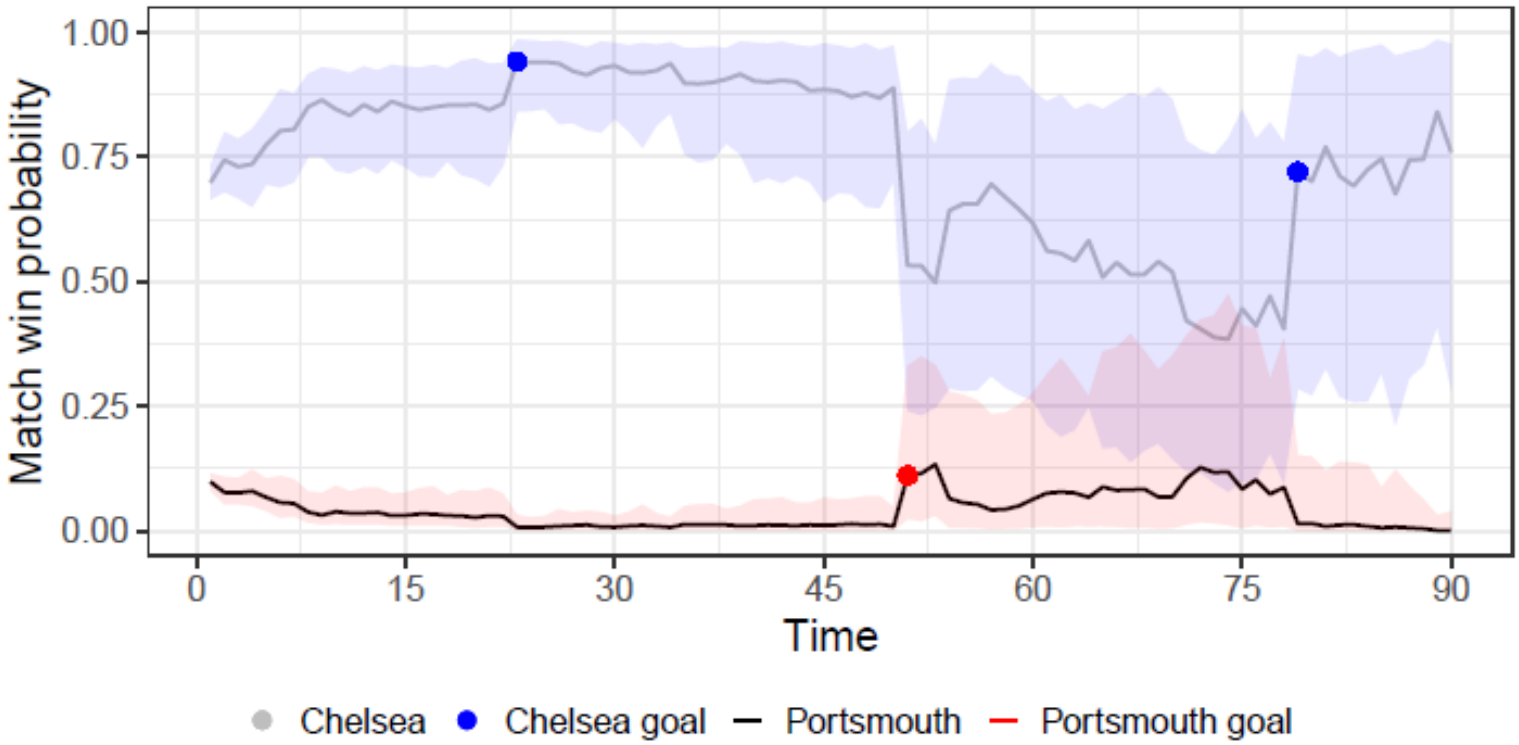
- All events have differential impact on the win probability depending on the time of occurrences.
- Goals and red cards have maximal impact; while crosses and shots have significant impact as well.
- The effect of home strength and away strength are relatively equal and of opposite signs.
- We employ a 90:10 (train:test) split for assessing predictive ability. Our method records great accuracy for the second half of the game.

Estimated time-varying effects



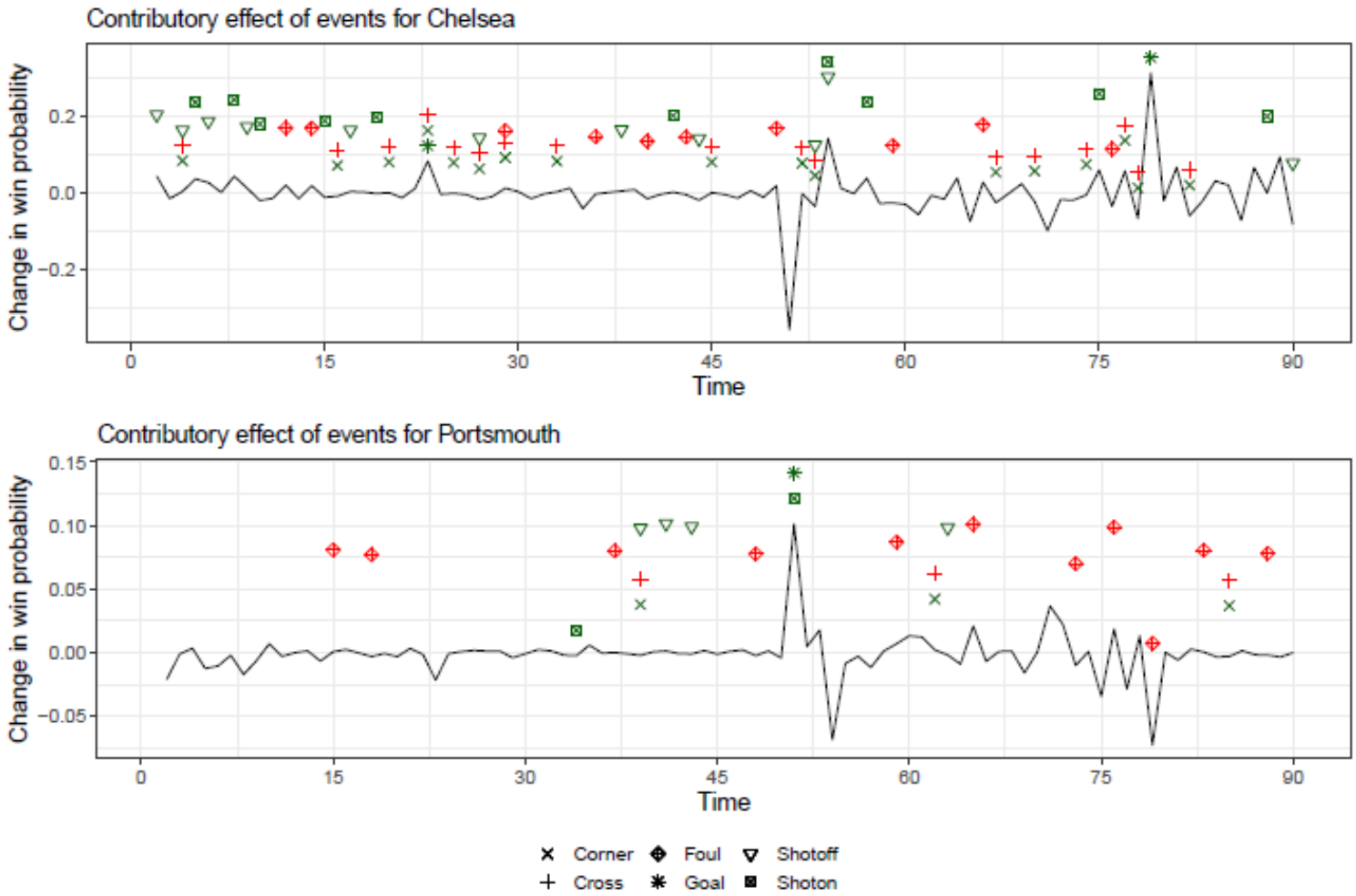
Effect of some events on the outcome of the match captured at every time-point for both teams. Such effects are estimated for all types of events.

An application of our method



Minute-by-minute forecast of the win probability, along with the credible interval, for the Chelsea vs Portsmouth match.

Further insights



The probabilistic effect of the time-varying events for the Chelsea vs Portsmouth match.

Model competitors

- Since there is no existing work, we rely on different algorithms that can be used for predictive modelling.

Model competitors

- Since there is no existing work, we rely on different algorithms that can be used for predictive modelling.
- The first one is a GLM in a probit framework.

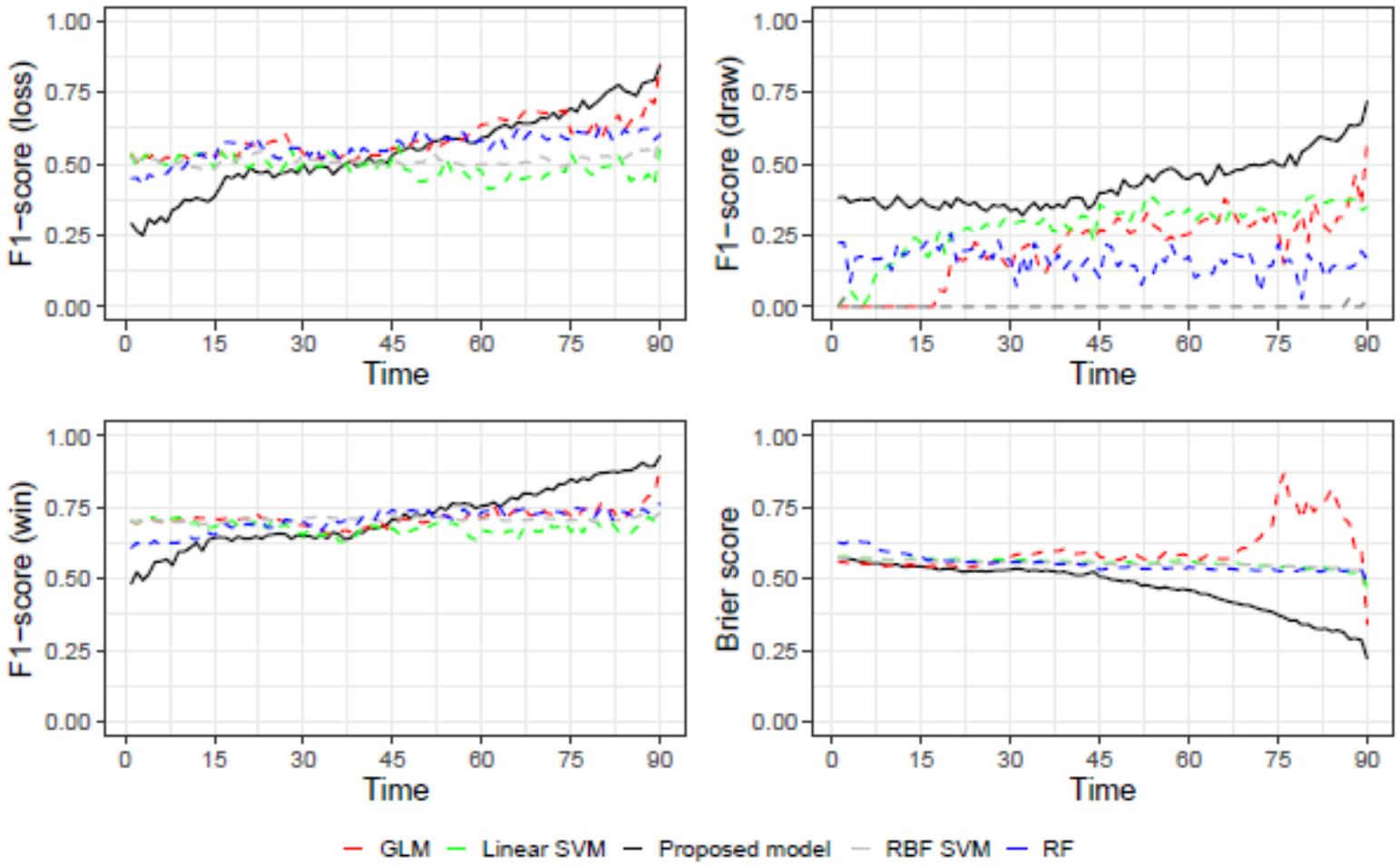
Model competitors

- Since there is no existing work, we rely on different algorithms that can be used for predictive modelling.
- The first one is a GLM in a probit framework.
- In context of football, machine learning algorithms have been used as predictive algorithms, which can be extended for our purpose. On those lines, we employ a linear SVM as a comparative model, which assumes linearly separable classes (L-SVM).
- For the next model, we incorporate non-linearity in the SVM in the form of a Gaussian radial basis function (R-SVM).

Model competitors

- Since there is no existing work, we rely on different algorithms that can be used for predictive modelling.
- The first one is a GLM in a probit framework.
- In context of football, machine learning algorithms have been used as predictive algorithms, which can be extended for our purpose. On those lines, we employ a linear SVM as a comparative model, which assumes linearly separable classes (L-SVM).
- For the next model, we incorporate non-linearity in the SVM in the form of a Gaussian radial basis function (R-SVM).
- As the fourth model in the comparative discussions, we modify the standard random forest (RF) algorithm.

Comparative performance



F1-score and Brier score (averaged over the test set) for different models, with respect to their within-game forecasting accuracy as a soccer match progresses

Robustness checks

- Teams are split into two groups – 'big six' and 'others'. Results remain similar for both groups.
- We predict the outcomes of the matches for a certain team, by training the model on the dataset excluding all matches of that team. Results display a consistent pattern.

Outline

Contribution

- We developed a Bayesian latent variable model for analyzing and forecasting soccer match outcomes in real-time.
- The article is the first in academic literature to furnish real-time predictions in football.
- The proposed model can assist with in-game adjustments and decision making for the managers.

Other applications

- In-game forecasting can be successfully utilized in various interesting applications, such as substitution strategy, within-match betting markets, advertisement in broadcasting, etc.

Other applications

- In-game forecasting can be successfully utilized in various interesting applications, such as substitution strategy, within-match betting markets, advertisement in broadcasting, etc.
- Our method is adaptable to other situations where the key variable is categorical in nature.

Other applications

- In-game forecasting can be successfully utilized in various interesting applications, such as substitution strategy, within-match betting markets, advertisement in broadcasting, etc.
- Our method is adaptable to other situations where the key variable is categorical in nature.
- We developed a modified algorithm for a multinomial time series for predicting winner of an election, when the counting is in progress (Deb, Roy, & Das, 2024).

Outline

Within-match forecasting in Cricket[1]

- A novel approach to within-game cricket forecasting is introduced by treating each match as a dynamic time series.
- We develop a new time series classification method, which leverages trend parallelism. Additionally, the aim is to study the changing influence of covariates within a match on the final outcome.

[1] This is a work in progress.

Within-match forecasting in Cricket[1]

- A novel approach to within-game cricket forecasting is introduced by treating each match as a dynamic time series.
- We develop a new time series classification method, which leverages trend parallelism. Additionally, the aim is to study the changing influence of covariates within a match on the final outcome.
- For main analysis, we consider all completed T20 matches between 2005 to 2025 with detailed information.
- Covariates are categorized into two groups: pre-match covariates and within-match covariates.

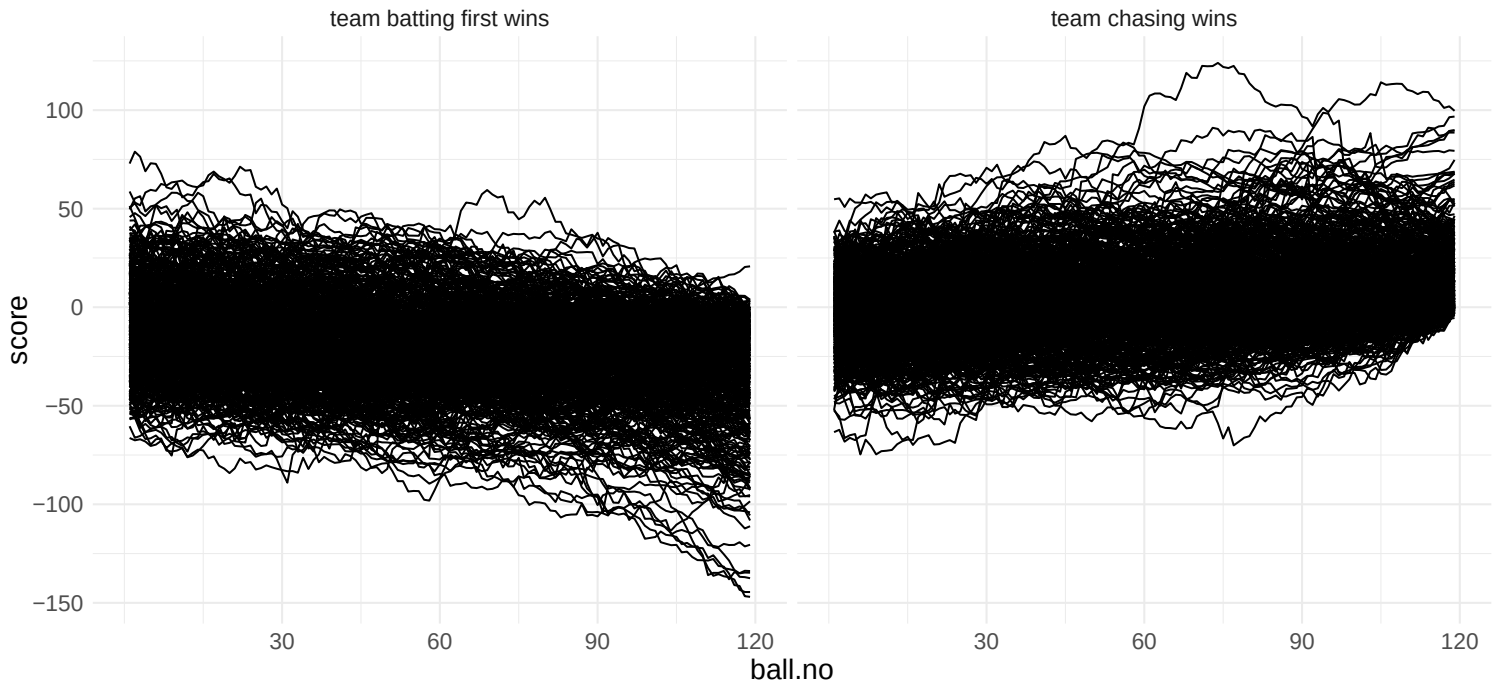
[1] This is a work in progress.

Motivation behind the classification algorithm

- We define difference of (modelled and adjusted for other factors) scores as a time series of length $t \leqslant 120$.

Motivation behind the classification algorithm

- We define difference of (modelled and adjusted for other factors) scores as a time series of length $t \leqslant 120$.
- This produces a distinct pattern for matches where the team batting first won, to the matches where the team batting second won.



Proposed classification algorithm

Input: $(x_{k,i})_{i=1}^{n}$: new time series to be classified
Output: Classification of $(x_{k,i})_{i=1}^{n}$
Start with $\mathcal{M}$ classes $C_1, C_2, \ldots, C_{\mathcal{M}}$;
for *each class* C_j **do**
    Estimate trend functions of time series within C_j using local linear estimation;
    Use the estimated trends to compute $\hat{\mu}_{C_j}(.)$ and $\hat{c}_k$, $k \in C_j$;
    Compute $RSS_{C_j} = \sum_{k \in C_j} \sum_{i=1}^{n} \{y_{k,i} - \hat{\mu}_{C_j}(i/n) - \hat{c}_k\}^2$;

for $l = 1$ *to* $\mathcal{M}$ **do**
    Assign $(x_{k,i})_{i=1}^{n}$ to C_l keeping rest of the classes unchanged;
    Recalculate RSS_{C_l};
    $RSS_l = \sum_{i=1}^{\mathcal{M}} RSS_{C_i}$;
    Reset C_l by removing $(x_{k,i})_{i=1}^{n}$ from it;

Classify $(x_{k,i})_{i=1}^{n}$ to C_t, where RSS_t is least;

Motivation is taken from Zhang (2013), Chazin et al. (2018).

Within-match forecasting in Tennis

- We work with point-by-point data from Wimbledon single's matches.
- Main objective is to understand whether first set offers key insights about how the match is going to end.
- LASSO-induced logistic regression model with random effects is developed to draw insights and to predict the outcomes.
- Details are in the paper by Gupta, Krishnamurthy & Deb (2024).

Model

Regressors

- Differences in the player's game statistics (except SEED and HAND).

$$g_{ij} = s_{ij} - r_{ij},\ 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p.$$

- Interaction terms (z_{ij}).

Proposed Model

- Variables selection is through LASSO, which avoids multicollinearity and overfitting issues.
- Random intercepts (corresponding to both players' seed categories) help to explain unexplained variance.
- Estimates are obtained from logistic regression setup, which helps with debiasing (Meinhausen, 2007).

Algorithm : LILRM_SEED_RI estimators

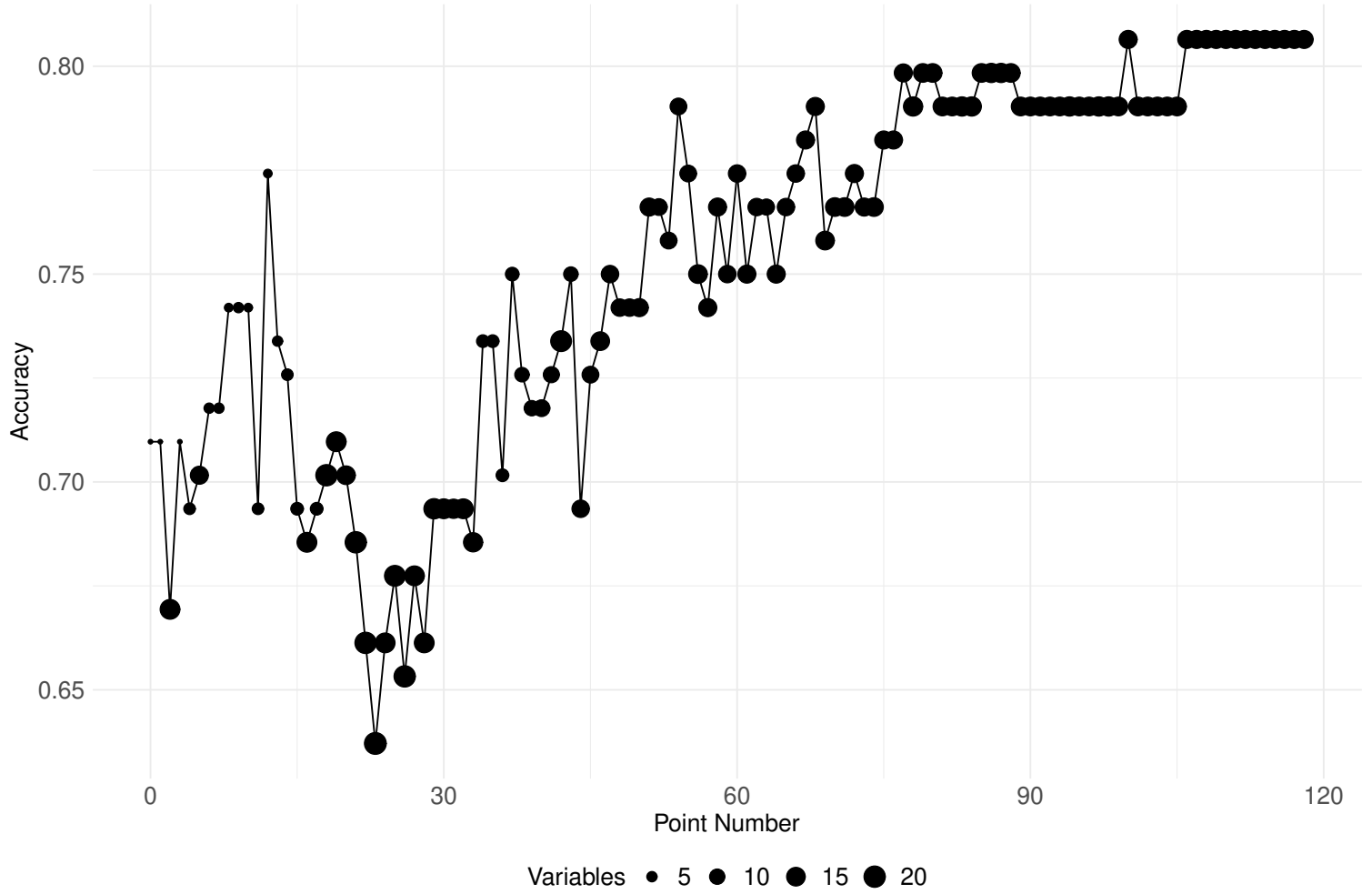
Input: Dataset $(Y, \mathcal{X})$, where Y is response, $\mathcal{X}$ is information on regressors.
Output: $\hat{S}$ = set of selected variables, $\hat{\Theta}$ = estimated coefficient vector.
Steps :

1. Create the design matrix X. Let m be the total number of variables.
2. Perform LASSO on (Y, X). Use $\hat{\phi}$ to denote the LASSO estimators.
3. Obtain index set of non-zero LASSO estimators: $\hat{S} = \{1 \leqslant k \leqslant m \mid \hat{\phi}_k \neq 0\}$.
4. Use seed wise random intercept models with the index set as covariates.
5. Consider the negative binomial log likelihood and use adaptive quadrature of gauss-hermite (AQGH) algorithm to find the estimates for all the variables in the index set $\hat{S}$. These estimates are called LILRM_SEED_RI estimators $(\hat{\Theta})$.

return $(\hat{S}, \hat{\Theta})$.

Average predictive performance after each point

Gender, age and rating difference variables are mandatory variables.



Outline

References (I)

- Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association, 88(422):669–679.
- Baboota, R. & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2):741–755.
- **Chazin, H., Deb, S., Falk, J., & Srinivasan, A. (2019). New Statistical Approaches to Intra-individual Isotopic Analysis and Modelling of Birth Seasonality in Studies of Herd Animals. Archaeometry, 61(2), 478-493.**
- **Deb, S., Roy, R., & Das, S. (2024). Forecasting elections from partial information using a Bayesian model for a multinomial sequence of data. Journal of Forecasting, 43(6), 1814-1834.**
- Dechi, B. O. (2019). Bayesian Analysis of Ordinal Outcomes Through Latent Variable Approach. The University of Texas at El Paso.
- **Divekar, C., Deb, S., & Roy, R. (2024). Real-time forecasting within soccer matches through a Bayesian lens. Journal of the Royal Statistical Society Series A: Statistics in Society, 187(2):513–540.**

References (II)

- Goddard, J. (2005). Regression models for forecasting goals and match results in association football. International Journal of forecasting, 21(2):331–340.
- **Gupta, K., Krishnamurthy, V., & Deb, S. (2024). What elements of the opening set influence the outcome of a tennis match? An in-depth analysis of Wimbledon data. IIMB Management Review.**
- McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2):109–127.
- McHale, I. & Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. Statistica Neerlandica, 61(4):432–445.
- Meinshausen, N. (2007). Relaxed lasso. Computational Statistics & Data Analysis, 52(1), 374-393.
- Vecer, J. (2014). Crossing in soccer has a strong negative impact on scoring: Evidence from the English Premier League, the German Bundesliga and the World Cup 2014. Available at SSRN 2225728.
- Zhang, T. (2013). Clustering high-dimensional time series based on parallelism. Journal of the American Statistical Association, 108(502), 577-588.

Thank you!

Contact: soudeep@iimb.ac.in
Webpage: https://soudeepd.github.io/