A Bayesian Examination of Win Probability and Timeout Usage in NCAA Men's Basketball

Luke Benz
Yale University Department of Applied Mathematics

Introduction

In a game played between Team A and Team B, an in-game win probability model estimates Team A's chances of beating Team B after each play completed in the game. Examining an in-game win probability model for men's college basketball allows one to explore the role that several factors, including timeouts and patterns of timeout usage, have in determining which team wins a given game. To date, no known in-game win probability models for any sport have an underlying framework that is exclusively Bayesian, in nature. The Bayesian model presented in this work allows for the examination of posterior correlation between covariates as well as uncertainty estimates for win probability curves.

Data

The data used in this project are complete play-by-play data consisting of 7,342,416 plays from 11,411 games over the 2016-17 and 2017-18 seasons, scraped from ESPN using the **ncaahoopR** package[1]. The following covariates are considered in the model.

- `favored_by`: Points team is favored by prior to the game, using Vegas point spread. The point spread is imputed when missing.
- `score_diff`: The current score differential.
- `timeout_remaining`: The number of timeouts remaining a team has at its disposal.
- `timeout_ind`: A binary indicator whether or not team took a timeout in the previous 60 seconds of game time.
- `home`: A binary indicator if the team in question is playing on its home court.

The model response, Y, available for every play of every game, is a binary indicator as to whether the team in question won the game from which the play comes. Since the above covariates have a non-linear dependence on time remaining in the game, the data are partitioned into the 31 data slices, where it is assumed the covariates remain constant time over each interval.

Statistical Model

Data slice k of time interval $(a_k, b_k]$ includes all plays (covariates and responeses) such that the `time_remaining`, t (in seconds) satisfied $a_k < t \leq b_k$. For each data slice $1 \leq k \leq 31$ of size n_k, the following model is assumed:

$$Y_{k,i}|p_{k,i} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{k,i}) \text{ for } 1 \leq i \leq n_k$$
$$\text{logit}(p_{k,i}) = \boldsymbol{X}_{k,i}\boldsymbol{\beta}_k$$

$$\text{Priors: } \beta_{k,j} \overset{\text{iid}}{\sim} N(0, 100^2) \text{ for } j = 1, ..., 5$$

The idea to use time-sliced logistic regression to model win probability in college basketball comes from Burke [2] and Torvik [3], though neither utilized a Bayesian framework for the model. The above statistical model was fit using the Gibbs sampling method for logistic regression first proposed by Polson et al. [4].

Posterior Distributions and Credible Intervals

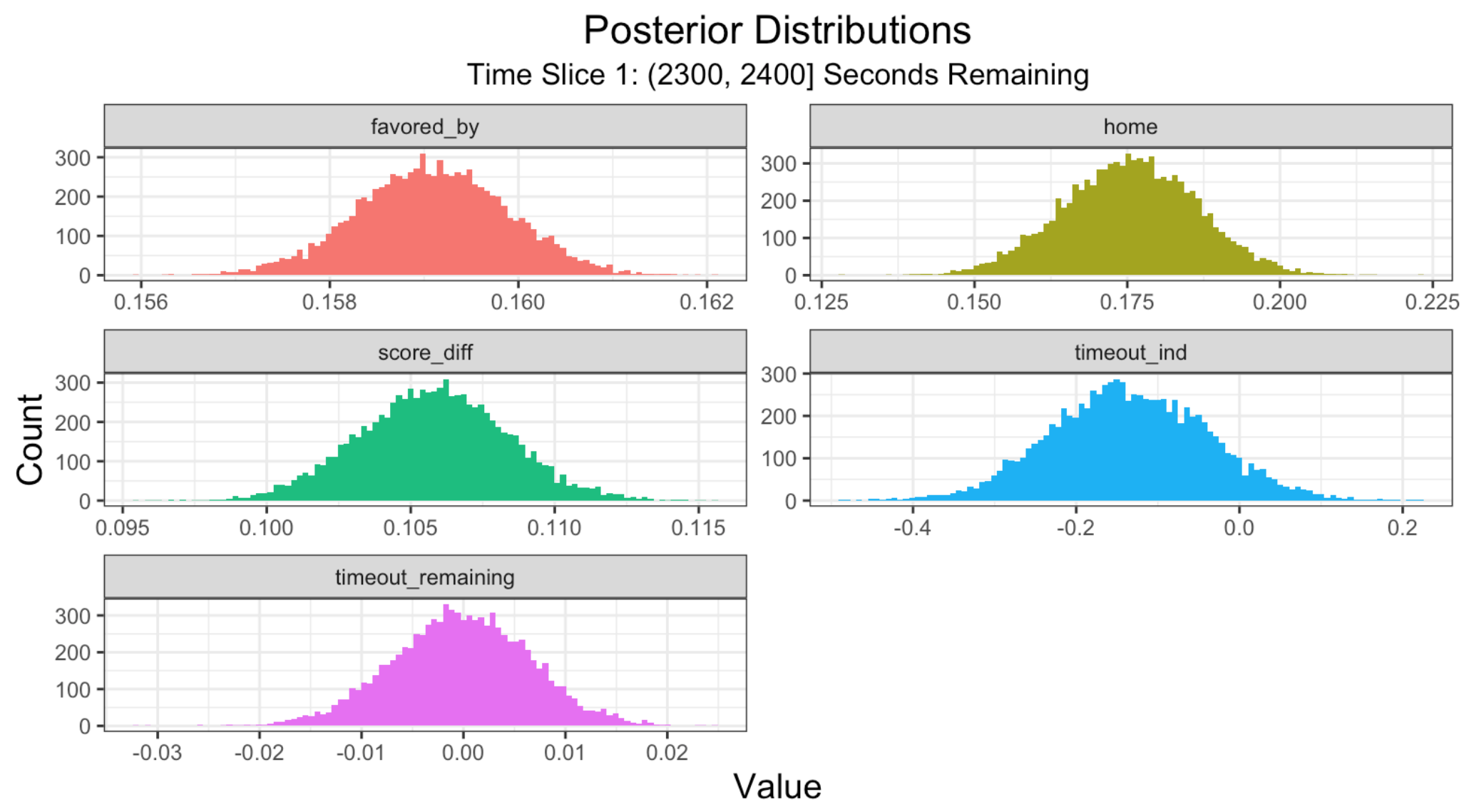

Figure 1: Posterior Distributions for $\boldsymbol{\beta}$ during the first time slice, $2300 <$ `time_remaining` ≤ 2400.

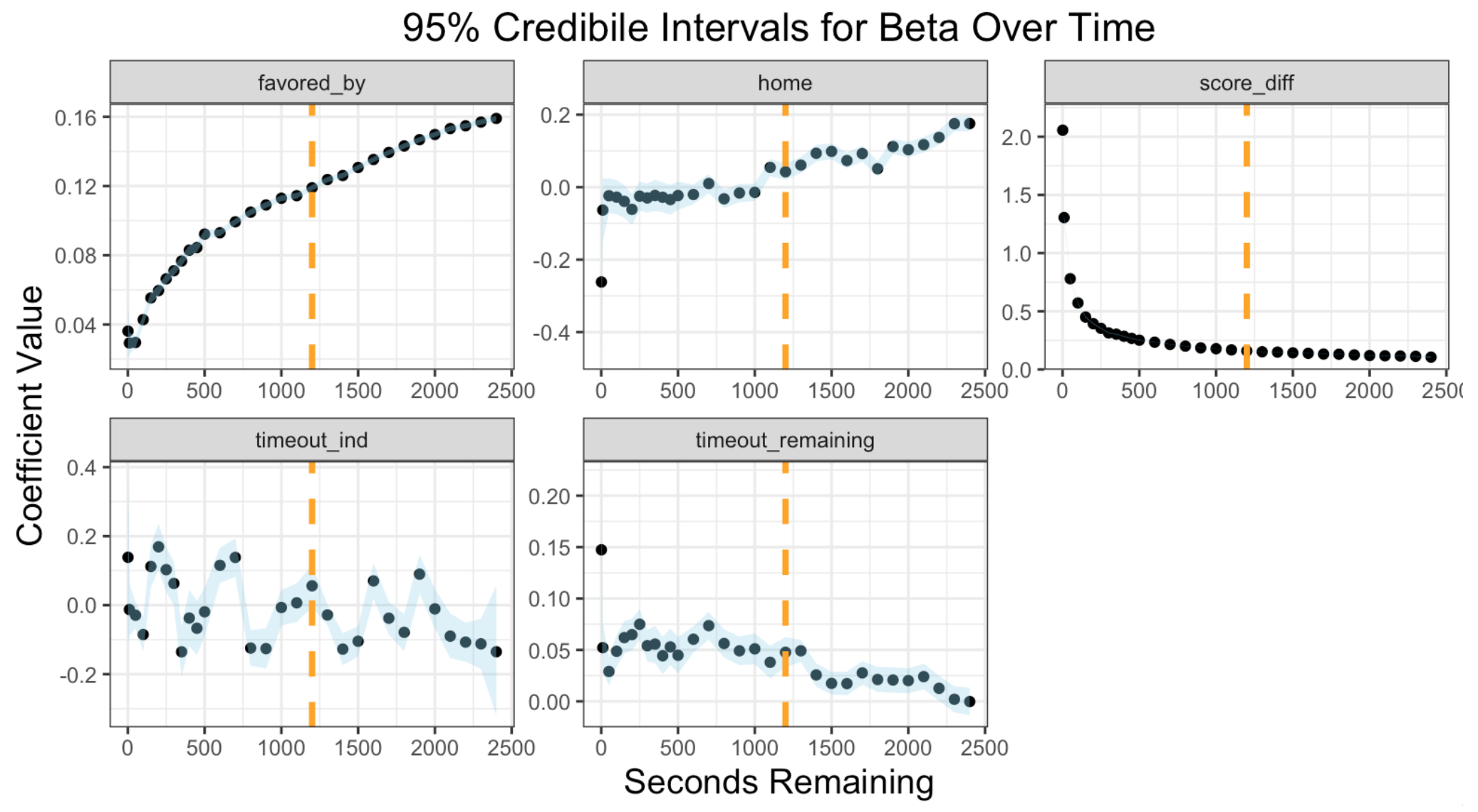

Figure 2: Posterior means for $\boldsymbol{\beta}$ over time are shown in black with 95% credible intervals in light blue.

Timeout Usage

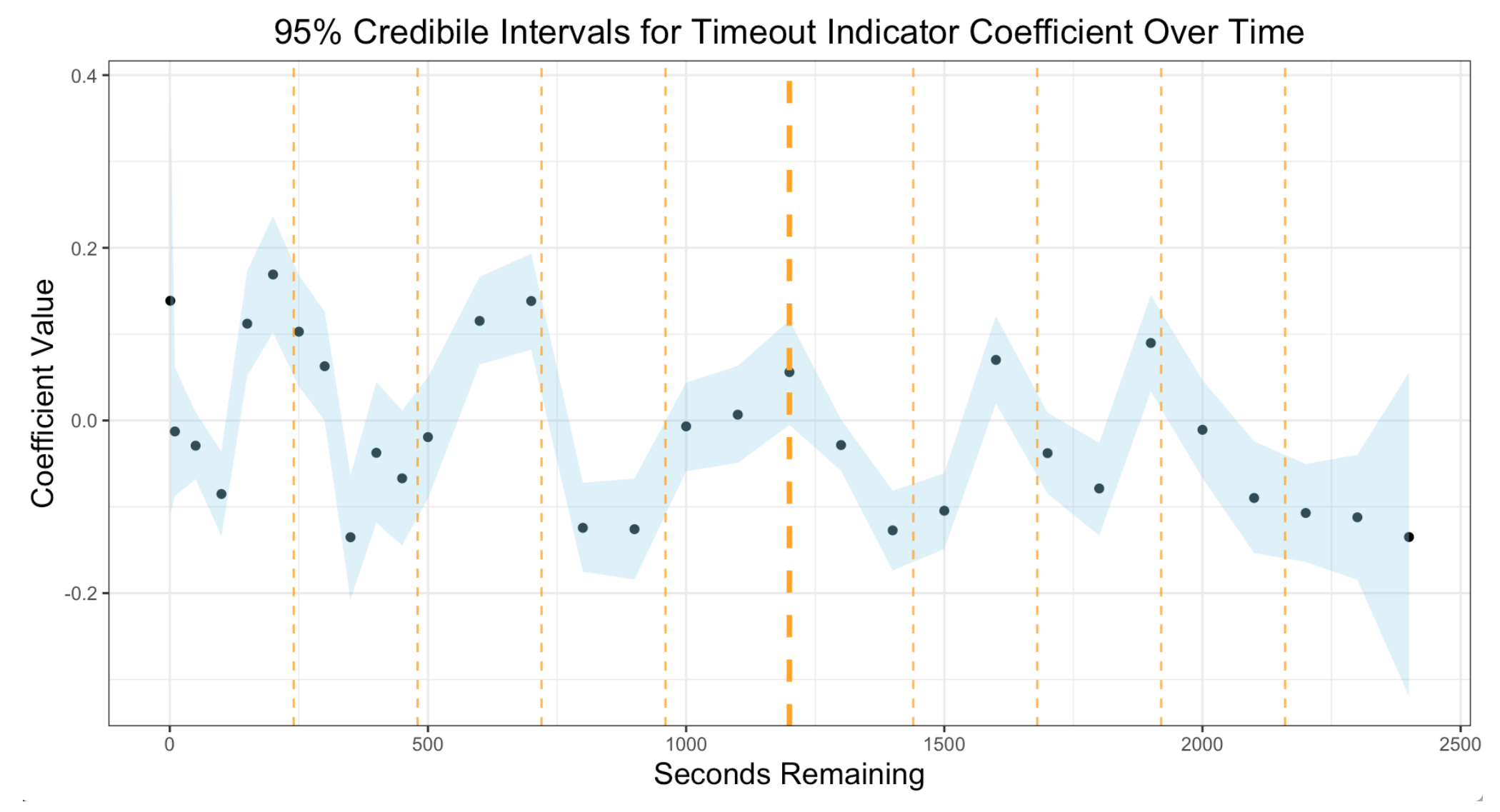

Figure 3: Posterior means for $\beta_{\texttt{timeout\_ind}}$ over time are shown in black with 95% credible intervals in light blue. Small orange lines denote typical spots for media timeouts.

At certain times of the game, taking a timeout is associated with positive changes in the log-odds (increases in win probability) and at other times of the game, taking a timeout is associated with negative changes in the log-odds (decreases in win probability). Such a pattern seems linked with times at which media timeouts occur, suggesting that taking a timeout to replace the media timeout may be better than using a timeout just before a media timeout is set to occur. There is also evidence to suggest the utility of using the so-called "use-it-or-lose-it" timeout immediately prior to halftime.

Posterior Correlations

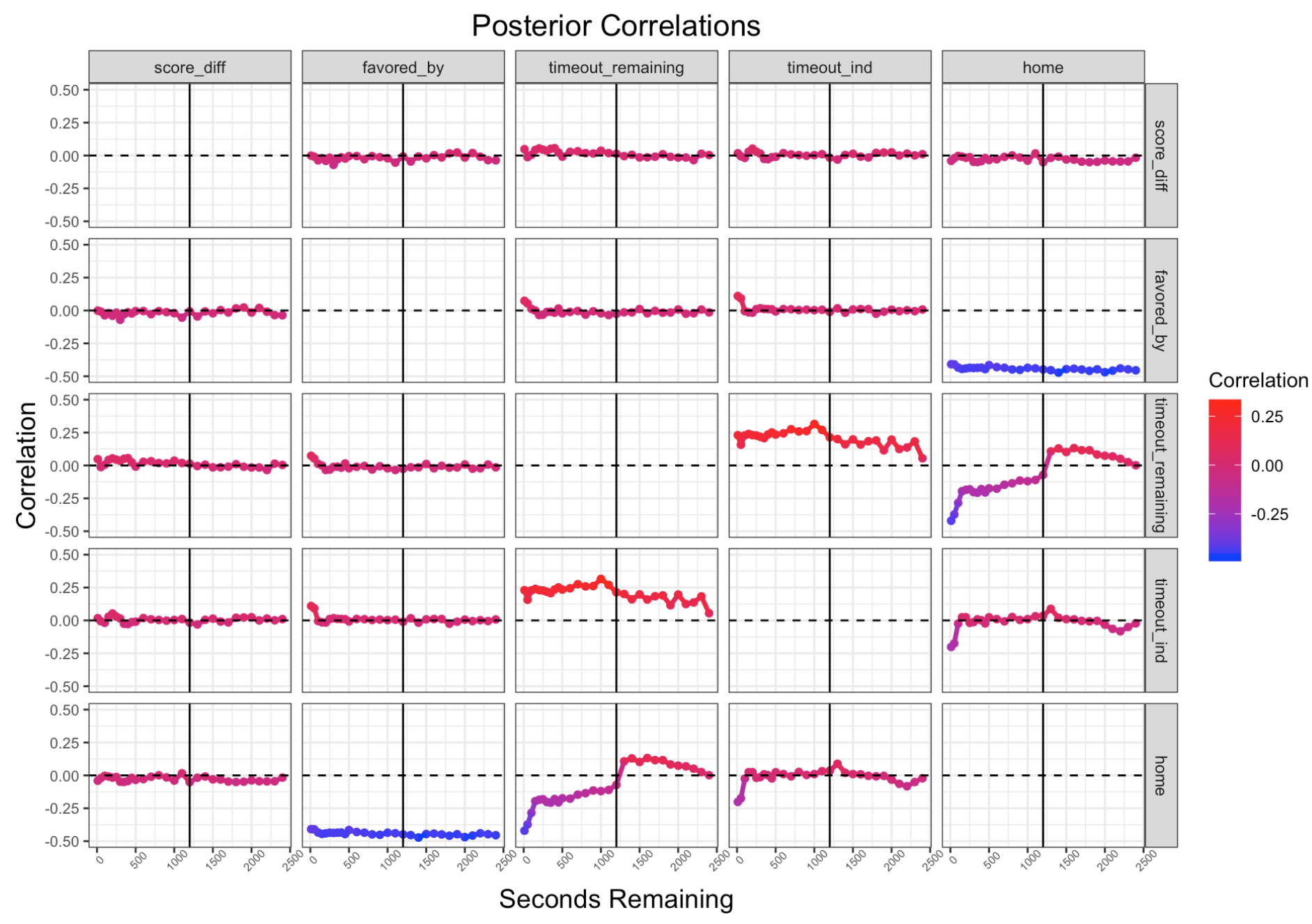

Figure 4: Posterior correlations $\rho(\beta_{k,i}, \beta_{k,j})|\boldsymbol{Y}$. There appears to be increased utility to taking the use-it-or-lose-it timeout as the home team. The switch in correlation between $\beta_{\texttt{home}}$ and $\beta_{\texttt{timeout\_remaining}}$ from positive to negative is perhaps explained by the hypothesis that home teams are likelier to have more timeouts remaining later in the game than away teams.

Win Probability Uncertainty Bounds

For an unseen game, uncertainty bounds on in-game win probability can be obtained by using our MCMC $\boldsymbol{\beta}$ posterior samples to derive 95% credible intervals for $p_{k,i}$. To best estimate the variance in $p_{k,i}$, one can consider the `favored_by` covariate $\sim N(\theta, 4)$ in MCMC sampling, where θ is the Vegas point spread.

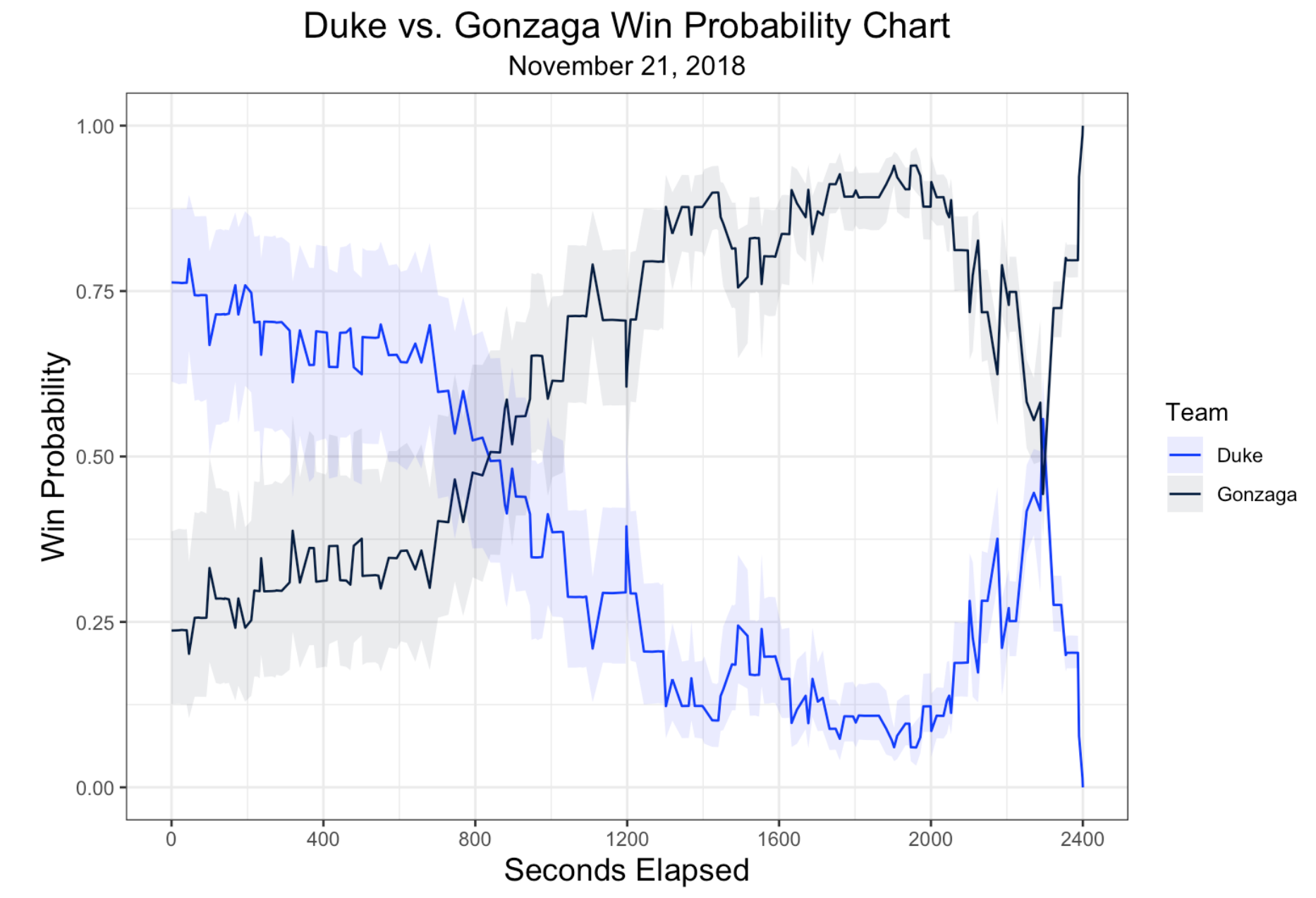

Figure 5: Win probability chart with 95% credible intervals for Duke vs. Gonzaga in the 2018 Maui Invitational Championship.

Future Work

Future work will seek use a Bayesian changepoint analysis to explore whether using a timeout has any quantifiable effect on momentum. Eventually, the hope is to build an application that could tell coaches exactly when timeouts should and shouldn't be used.

References

[1] L. S. Benz.
ncaahoopR: An R package for working with NCAA Basketball Play-by-Play Data.
`https://github.com/lbenz730/ncaahoopR`, 2018.

[2] B. Burke.
Modeling Win Probability for a College Basketball Game: A Guest Post from Brian Burke.
`https://bit.ly/1FJIjd4`, 2009.

[3] B. Torvik.
How I Built a (Crappy) Basketball Win Probability Model.
`http://adamcwisports.blogspot.com/2017/07/how-i-built-crappy-basketball-win.html`, 2018.

[4] N. G. Polson, J. G. Scott, and J. Windle.
Bayesian Inference for Logistic Models Using Pólya-Gamma Latent Variables.
Journal of the American Statistical Association, 108(54):1339–1349, August 2013.