Noname manuscript No.
(will be inserted by the editor)

Having a Ball: evaluating scoring streaks and game excitement using in-match trend estimation

Andreas Kryger Jensen and Claus Thorn Ekstrøm
(Andreases Orcid), (0000-0003-1191-373X)
Biostatistics, Department of Public Health,
University of Copenhagen
aeje@sund.ku.dk, ekstrom@sund.ku.dk

25 November, 2020

Address(es) of author(s) should be given

Abstract Many popular sports involve matches between two teams or players where each team have the possibility of scoring throughout the match. While the overall winner and result is interesting it conveys little information about the underlying scoring trends throughout the match. Modeling approaches that accommodate a finer granularity of the score difference throughout the match is needed in order to evaluate game strategies, discuss scoring streaks, teams strengths, and other aspects of the game.

We propose a latent Gaussian process to model the score difference between two teams and introduce the Trend Direction Index as an easily interpretable probabilistic measure of the current trend in the match as well as a measure of post-game trend evaluation. In addition we propose the Excitement Trend Index — the expected number of monotonicity changes in the trend of the running score difference — as a measure of overall game excitement.

We apply the our proposed methodology to analyze all 1143 matches from the 2019–2020 National Basketball Association (NBA) season, and show how the trends can be interpreted in individual games and how the excitement score can be used to cluster teams according to how exciting they are to watch.

Keywords: Bayesian Statistics, Gaussian Processes, Sports Statistics, Trends, APBRmetrics

1 Declarations

Funding: The research was funded by the University of Copenhagen.

Conflicts of interest/Competing interests: None

Availability of data and material: All data are available at `https://github.com/aejensen/Having-a-Ball`

Code availability: All code are available at `https://github.com/aejensen/Having-a-Ball` and `https://github.com/aejensen/TrendinessOfTrends`.

2 Introduction

Sports analytics receive increasing attention within statistics and not just for match prediction or betting but also for game evaluation, in-game and post-game coaching purposes, and for setting strategies and tactics in future matches.

Many popular sports such as football (soccer), basketball, boxing, table tennis, volleyball, American football, and handball involve matches between two teams or players where each team have the possibility of scoring throughout the match. Several research papers seek to predict match results (e.g., Karlis and Ntzoufras (2003); Groll et al. (2019); Gu and Saaty (2019)) or match winners for single matches (Cattelan, Varin, and Firth (2013)) in order to infer the match winner and potentially the winner of a tournament (Ekstrøm et al. (2020); Baboota and Kaur (2018)). While the overall result is highly interesting it conveys very little information about the individual development and trends throughout the match and modeling approaches that allow a finer granularity of the score difference throughout the match are needed.

The trend in score difference between the two teams is a proxy for their underlying strengths. In particular, sustained periods of time where the score difference increases suggests that one team outperforms the other whereas periods where the teams are constantly catching up to each other suggest that the team's strengths in those periods are similar. Modeling the local trend of the score difference will therefore reflect several aspects of the game in particular the team strengths and game dynamics and momentum as they develop through the match.

Figure 1 shows the development of the score difference for the final match of the playoffs in the 2020 National Basketball Association (NBA) series between Los Angeles Lakers and Miami Heat. Positive numbers indicate that LA Lakers are leading and the running score difference shows that the Lakers pulled ahead until the third quarter where Miami Heat started to keep up the scoring pace before overtaking the Lakers and reducing the lead.

In this paper we will consider the score difference between two teams as a latent Gaussian process and use the Trend Direction Index (TDI) from Jensen and Ekstrøm (2020b) as an accommodating measure to evaluate the local probability of the *monotonicity* of the latent process at a given time point during a match. The Trend Direction Index uses a Bayesian framework to provide a direct answer to questions such as "What is the probability that the latent process (i.e., that one team is doing better than another) is increasing at a given time-point?" This will allow real-time evaluation of the score difference trend at the current time-point in the game and will provide post-game inference about the "hot" periods of a match where one team out-performed the other. Furthermore we will present the Excitement Trend Index (ETI) as an objective measure of spectator excitement in a given match. The ETI is defined as the expected number of times that the score difference changes monotonicity during a match. If the score difference changes monotonicity often then that echos a game where both teams frequently score whereas a game with a low ETI will represent a one-sided match where one team is doing consistently better than the other over sustained periods of time.

Other authors have considered using continuous processes to model the score difference of matches. Gabel and Redner (2012) shows that NBA basketball score difference is

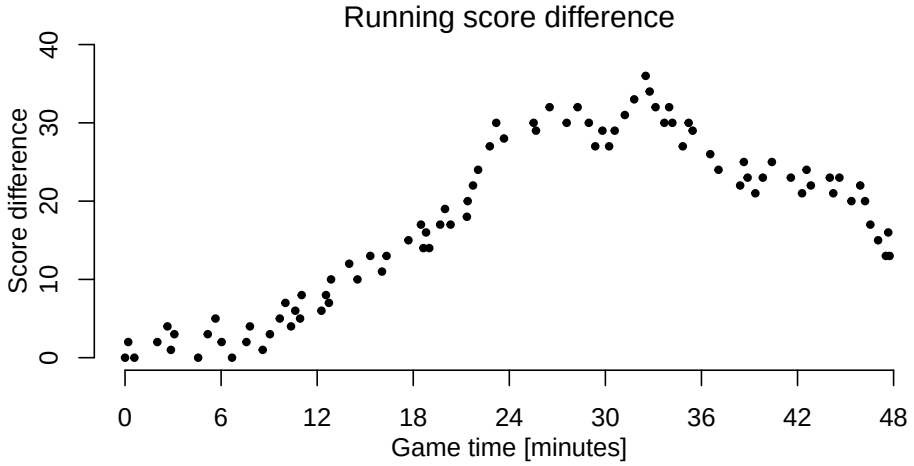

Fig. 1: Game development in the final match of the NBA 2019–2020 season between Los Angeles Lakers and Miami Heat. October 11, 2020. Positive values indicate that LA Lakers are leading.

well described by a continuous-time anti-persistent random walk which suggests that a latent Gaussian process might be viable.

- Y. Chen, Dawson, and Müller (2020)
- T. Chen and Fan (2018)

The paper is structured as follows. In the next section we introduce trend modeling of score differences through a latent Gaussian process and define the Trend Direction Index and Excitement Trend Index that captures the local trends in monotonicity and game excitement, respectively. In Section 4 we apply the our proposed methodology to analyze both the final match of the playoff as well as evaluating the game excitement distribution of the season by considering the ETI from all 1143 matches from the 2019–2020 National Basketball Association (NBA) season. We show how this distribution can be used to assess relative match excitement and how the ETI can be used to classify teams according to their average level of match excitement. We conclude with a discussion in Section 5. Materials to reproduce this manuscript and its analyses can be found at Jensen and Ekstrøm (2020a).

Unused stuff for now:

- Philosophical question: What exactly is the sampling model for a single match?

3 Methods

Our model is based on the observed score differences D_m in a given match indexed by m. For each match we observe the random variables $\mathcal{D}_m = (t_{mi}, D_{mi})_{0<i\leq J_m}$ where $t_{m1} < t_{m2} < t_{mi} < \ldots < t_{mJ_m}$ are the ordered time points at which any teams scores, $D_{mi} = D_m(t_{mi})$ is the associated difference in scores at time t_{mi}, and J_m is the total number of scorings during the match. We use the convention that D_m is

the difference in scores of the away team with respect to the home team, so that $D_m(t) > 0$ means that the away team is leading at time t.

We assume that the observed match data from a given match are noisy realizations of a latent smooth, random function defined in continuous time evaluated at the random time points where goals occur. Let d_m be the latent functions from which the realizations $\mathcal{D}_m$ are generated. Our objective is to infer d_m and its time dynamics from $\mathcal{D}_m$. In pursuance of this ambition we propose the following model where d_m is a Gaussian process defined on a compact subset of the real line $\mathcal{I}_m$ corresponding to the duration of the m'th game, and the observed data conditional on the scoring times and the values of the latent process at these times are independently normally distributed random variables with a match specific variance. This model for a given match can be stated hierarchically as

$$\begin{aligned} \text{✿}_m \mid \boldsymbol{\Psi_m}, \mathbf{t}_m &\sim H(\text{✿}_m \mid \boldsymbol{\Psi}_m) \\ d_m(t) \mid \text{✿}_m &\sim \mathcal{GP}(\mu_{\boldsymbol{\beta}_m}(t), C_{\boldsymbol{\theta}_m}(s,t)) \\ D_m(t_{mi}) \mid d_m(t_{mi}), t_{mi}, \text{✿}_m &\overset{iid}{\sim} N\left(d_m(t_{mi}), \sigma_m^2\right) \end{aligned} \tag{1}$$

where $\text{✿}_m = (\boldsymbol{\beta}_m, \boldsymbol{\theta}_m, \sigma_m^2)$ is a vector of hyper-parameters governing the dynamics of the latent Gaussian process with a prior distribution H indexed by parameters $\boldsymbol{\Psi}_m$, and $\mathbf{t}_m = (t_{m1}, \ldots, t_{mJ_m})$ is the vector of time points where goals occurs in the match. The functions $\mu_{\boldsymbol{\beta}_m}$ on $\mathcal{I}_m$ and $C_{\boldsymbol{\theta}_m}$ on $\mathcal{I}_m \times \mathcal{I}_m$ are the prior mean and covariance functions of the latent Gaussian process, and σ_m^2 is the variance characterizing the magnitudes deviations between for the observed score differences and the values of the latent process.

A Gaussian process is characterized by the multivariate joint normality of all of the joint distributions resulting from evaluating the process at any finite set of time points (Rasmussen and Williams (2006)). Specifically, for any finite set $\mathbf{t}^* \subset \mathcal{I}_m$ it follows that that the vector $d(\mathbf{t}^*) \mid \text{✿}_m$ is distributed as $N(\mu_{\boldsymbol{\beta}_m}(\mathbf{t}^*), C_{\boldsymbol{\theta}_m}(\mathbf{t}^*, \mathbf{t}^*))$ where $\mu_{\boldsymbol{\beta}_m}(\mathbf{t}^*)$ is the vector generated by evaluating the prior mean function $\mu_{\boldsymbol{\beta}_m}(t)$ at $\mathbf{t}^*$ and $C_{\boldsymbol{\theta}_m}(\mathbf{t}^*, \mathbf{t}^*))$ is the covariance matrix generated by evaluating the prior covariance function $C_{\boldsymbol{\theta}_m}(s,t)$ at $\mathbf{t}^* \times \mathbf{t}^*$. Using the properties of multivariate normal distributions the posterior distribution of $d_m(\mathbf{t}^*) \mid \mathcal{D}_m, \text{✿}_m$ is also a multivariate normal distribution. This facilitates Bayesian estimation of the distribution of the latent process governing the score difference given the observed data from each match.

In addition to obtaining inference for the latent process we may also estimate its time dynamics. This follows since a Gaussian process along with its time derivatives (provided they exist) are distributed as a multivariate Gaussian process (Cramer and Leadbetter (1967)). We can therefore augment the model in Equation (1) with an additional latent structure of the first and second derivatives of d_m with respect to time as

$$\begin{bmatrix} d_m(s) \\ d'_m(t) \\ d''_m(u) \end{bmatrix} \mid \text{✿}_m \sim \mathcal{GP}\left(\begin{bmatrix} \mu_{\boldsymbol{\beta}_m}(s) \\ \mu'_{\boldsymbol{\beta}_m}(t) \\ \mu''_{\boldsymbol{\beta}_m}(u) \end{bmatrix}, \begin{bmatrix} C_{\boldsymbol{\theta}_m}(s,s') & \partial_2 C_{\boldsymbol{\theta}_m}(s,t) & \partial_2^2 C_{\boldsymbol{\theta}_m}(s,u) \\ \partial_1 C_{\boldsymbol{\theta}_m}(t,s) & \partial_1\partial_2 C_{\boldsymbol{\theta}_m}(t,t') & \partial_1\partial_2^2 C_{\boldsymbol{\theta}_m}(t,u) \\ \partial_1^2 C_{\boldsymbol{\theta}_m}(u,s) & \partial_1^2\partial_2 C_{\boldsymbol{\theta}_m}(u,t) & \partial_1^2\partial_2^2 C_{\boldsymbol{\theta}_m}(u,u') \end{bmatrix} \right) \tag{2}$$

where $'$ and $''$ denote the first and second time derivatives, ∂_j^k is the k'th order partial derivative with respect to the j'th variable. Combining the models in Equations (1)

and (2) we obtain explicit expressions for the posterior distributions $d'_m \mid \mathcal{D}_m, \soccerball_m$ and $d''_m \mid \mathcal{D}_m, \soccerball_m$. The posterior joint distributions of the latent processes can be expressed as the multivariate Gaussian process

$$\begin{bmatrix} d_m(s) \\ d'_m(t) \\ d''_m(u) \end{bmatrix} \mid \mathcal{D}_m, \soccerball_m \sim \mathcal{GP}\left(\begin{bmatrix} \mu_{d_m}(s) \\ \mu_{d'_m}(t) \\ \mu_{d''_m}(u) \end{bmatrix}, \begin{bmatrix} \Sigma_{d_m}(s,s') & \Sigma_{d_m d'_m}(s,t) & \Sigma_{d_m d''_m}(s,u) \\ \Sigma_{d'_m d_m}(t,s) & \Sigma_{d'_m}(t,t') & \Sigma_{d'_m d''_m}(t,u) \\ \Sigma_{d''_m d_m}(u,s) & \Sigma_{d''_m d'_m}(u,t) & \Sigma_{d''_m}(u,u') \end{bmatrix} \right) \tag{3}$$

where explicit expressions for the posterior mean and covariance functions are given in the supplementary material. In practice this means that we can sample from the posterior joint distribution at any finite number of time points as this will be equal to sampling from a high-dimensional normal distribution. We use the posteriors of the first and second time derivatives of the latent process to characterize the dynamical properties of each match through the Trend Direction Index (TDI) and the Excitement Trend Index (ETI).

We define the Trend Direction Index (TDI) of a particular match m as the local posterior probability that d_m is an increasing function for all time points $t \in \mathcal{I}_m$. Under our model this is equal to

$$\begin{aligned} \mathrm{TDI}_m(t \mid \soccerball_m) &= P\left(d'_m(t) > 0 \mid \mathcal{D}_m, \soccerball_m\right) \\ &= \frac{1}{2} + \frac{1}{2}\,\mathrm{Erf}\left(\frac{\mu_{d'_m}(t)}{2^{1/2}\Sigma_{d'_m}(t,t)^{1/2}}\right) \end{aligned} \tag{4}$$

where $\mathrm{Erf}\colon x \mapsto 2\pi^{-1/2}\int_0^x \exp(-u^2)\mathrm{d}u$ is the error function and $\mu_{d'_m}$ and $\Sigma_{d'_m}$ are the posterior mean and covariance functions of the time derivative defined in Equation (3). The interpretation of the TDI is that it quantifies the probability that one team is currently increasing the differences in scores or equivalently that they are changing the trend in their favor. A TDI equal to 50% means that the game is in a stagnant state. We note that the TDI is symmetric with respect to the reference team in the definition of the score difference. Is the reference team is changed, then TDI changes to $1 - \mathrm{TDI}$.

To each match we assign its Excitement Trend Index, denoted by ETI_m, as a global index of excitement. The index is defined as the expected number of changes in monotonicity of the posterior distribution of d_m which is equivalent to the expected number of zero-crossings of the posterior distribution of d'_m. We hence define

$$\begin{aligned} \mathrm{ETI}_m \mid \soccerball_m &= \mathrm{E}\left[\#\left\{t \in \mathcal{I}_m : d'_m(t) = 0\right\} \mid \mathcal{D}_m, \soccerball_m\right] \\ &= \int_{\mathcal{I}_m} d\mathrm{ETI}_m(t \mid \soccerball_m)\mathrm{d}t \end{aligned} \tag{5}$$

where $d\mathrm{ETI}_m$ is the instantaneous posterior probability of a zero-crossing of d' at any time point $t \in \mathcal{I}_m$. Under the model described in Equations (1) and (2) it can be shown that $d\mathrm{ETI}$ is equal to

$$d\mathrm{ETI}_m(t \mid \soccerball_m) = \lambda_m(t)\phi\left(\frac{\mu_{d'_m}(t)}{\Sigma_{d'_m}(t,t)^{1/2}}\right)\left(2\phi\left(\zeta_m(t)\right) + \zeta_m(t)\,\mathrm{Erf}\left(\frac{\zeta_m(t)}{2^{1/2}}\right)\right)$$

where $\phi\colon\ x \mapsto 2^{-1/2}\pi^{-1/2}\exp(-\frac{1}{2}x^2)$ is the standard normal density function, and λ_m, ω_m and ζ_m are functions defined by

$$\lambda_m(t) = \frac{\Sigma_{d''_m}(t,t)^{1/2}}{\Sigma_{d'_m}(t,t)^{1/2}}\left(1-\omega_m(t)^2\right)^{1/2}, \quad \omega_m(t) = \frac{\Sigma_{d'_m d''_m}(t,t)}{\Sigma_{d'_m}(t,t)^{1/2}\Sigma_{d''_m}(t,t)^{1/2}}$$
$$\zeta_m(t) = \frac{\mu_{d'_m}(t)\Sigma_{d'_m}(t,t)^{1/2}\omega_m(t)\Sigma_{d'_m}(t,t)^{-1/2} - \mu_{d''_m}(t)}{\Sigma_{d''_m}(t,t)^{1/2}\left(1-\omega_m(t)^2\right)^{1/2}}$$

A derivation of these expressions can be found in the supplementary material to Jensen and Ekstrøm (2020b). [**??? give some explanation/intuition about the value ???**]. We note that ETI is invariant with respect to the reference team in the definition of the score differences as it is defined as the expected number of both up- and down-crossings at zero of the posterior trend.

Both the Trend Direction Index and the Excitement Trend Index in Equations (4) and (5) are random variables due to their dependence on the hyper-parameters $\boldsymbol{\Psi}_m$. In our Bayesian framework these are specified under an additional layer of prior distributions according to $H(\boldsymbol{\Psi}_m \mid \boldsymbol{\Psi}_m)$. By fitting the model using Markov-Chain Monte Carlo methods (MCMC) we obtain the posterior distribution $\widetilde{\boldsymbol{\Psi}}_m \sim P(\boldsymbol{\Psi}_m \mid \mathcal{D}_m, \boldsymbol{\Psi}_m, \mathbf{t}_m)$ given by

$$\widetilde{\boldsymbol{\Psi}}_m \sim \frac{H(\boldsymbol{\Psi}_m \mid \boldsymbol{\Psi}_m)\int P(\mathbf{D}_m \mid d_m(\mathbf{t}_m), \boldsymbol{\Psi}_m, \boldsymbol{\Psi}_m, \mathbf{t}_m) dP(d_m(\mathbf{t}_m) \mid \boldsymbol{\Psi}_m, \boldsymbol{\Psi}_m, \mathbf{t}_m)}{\iint P(\mathbf{D}_m \mid d_m(\mathbf{t}_m), \boldsymbol{\Psi}_m, \boldsymbol{\Psi}_m, \mathbf{t}_m) dP(d_m(\mathbf{t}_m) \mid \boldsymbol{\Psi}_m, \boldsymbol{\Psi}_m, \mathbf{t}_m) dH(\boldsymbol{\Psi}_m \mid \boldsymbol{\Psi}_m)}$$

and the posterior estimates of TDI and ETI are therefore the random variables $\mathrm{TDI}_m(t) \mid \widetilde{\boldsymbol{\Psi}}_m$ and $\mathrm{ETI}_m \mid \widetilde{\boldsymbol{\Psi}}_m$. [**??? check notation ???**]

To summarize the posterior distributions of TDI and ETI to one can either use the average value calculated by integrating over the distribution of $\widetilde{\boldsymbol{\Psi}}_m$ or

$$\overline{\mathrm{TDI}}_m(t) = \int \widetilde{\boldsymbol{\Psi}}_m \mathrm{TDI}_m(t \mid \widetilde{\boldsymbol{\Psi}}_m) dP(\widetilde{\boldsymbol{\Psi}}_m)$$
$$\overline{\mathrm{TDI}^p}_m(t) = \sup\left\{\mathrm{TDI}_m(t \mid \widetilde{\boldsymbol{\Psi}}_m) : \mathrm{TDI}_m(t \mid \widetilde{\boldsymbol{\Psi}}_m) < p\right\}$$

[**??? look this through ???**]

Estimation

To complete the specification of the model in Equation (1) we need to specify prior mean and covariance functions for the latent process. The choice of these are application specific and can be based on prior knowledge of the game dynamics. We refer to the discussion in Jensen and Ekstrøm (2020b) for more information on such choices.

In our application we used a constant prior mean and the squared exponential covariance function given by

$$\mu_{\boldsymbol{\beta}_m}(t) = \beta_m, \quad C_{\boldsymbol{\theta}_m}(s,t) = \alpha_m^2 \exp\left(-\frac{(s-t)^2}{2\rho_m^2}\right)$$

with $\boldsymbol{\theta}_m = (\alpha_m, \rho_m) > 0$ and hence $\basketball_m = (\beta_m, \alpha_m, \rho_m, \sigma^2)$. [**??? Note: Infinitely differetiable sample paths. Sample path derivatives are well-defined**] For the hyper-parameters of $\basketball_m$ we used independent, heavy-tailed distribution with a moderate variance centered at the marginal maximum likelihood estimates of the form

$$H(\basketball_m \mid \boldsymbol{\Psi}_m) = H(\beta_m \mid \Psi_{\beta_m}) H(\alpha_m \mid \Psi_{\alpha_m}) H(\rho_m \mid \Psi_{\rho_m}) H(\sigma_m \mid \Psi_{\sigma_m})$$

with

$$\beta_m \sim T_4\left(\widehat{\beta_m^{\mathrm{ML}}}, 5\right), \quad \alpha_m \sim T_4^+\left(\widehat{\alpha_m^{\mathrm{ML}}}, 5\right), \quad \rho_m \sim T_4^+\left(\widehat{\rho_m^{\mathrm{ML}}}, 5\right), \quad \sigma_m \sim T_4^+\left(\widehat{\sigma_m^{\mathrm{ML}}}, 5\right)$$

where T_{df} denotes a location-scale T distribution with df degrees of freedom, and T_{df}^+ is the distribution truncated to the positive real line.

We have implemented the estimation procedure in the probabilistic programming language Stan (Carpenter et al. 2017) using the Hamiltonian MCMC (Markov Chain Monte Carlo) algorithm to sample from the posterior distributions $\mathrm{TDI}_m(t \mid \widetilde{\basketball}_m)$ and $d\mathrm{ETI}_m(t \mid \widetilde{\basketball}_m)$ on a equidistant grid of 200 time points in $\mathcal{I}_m$. For each match we ran four independent chains for 25,000 iterations each with half of the iterations used for warm-up. The posterior distribution of $\mathrm{ETI}_m \mid \widetilde{\basketball}_m$ was calculated by numerical integration of $d\mathrm{ETI}_m(t \mid \widetilde{\basketball}_m)$ using the trapezoidal rule on the grid of time points.

4 Application: The 2019–2020 NBA basketball season

To show the applicability of our proposed model we will apply it to data from all regular games from the 2019–2020 NBA basketball season (obtained from Sports Reference LLC (2020) and provided in Jensen and Ekstrøm (2020a)). A lot of points are scored during a basketball match so it is easy to see the development of the score difference in a single match.

The 2019–2020 NBA season was suspended mid-March because of Covid-19 but it was resumed again in July 2020. There were a total of 1059 regular season matches. The subsequent playoffs comprised 84 matches including the final for a grand total of 1143 matches. For ease of comparison we are only considering the first 48 regular minutes of each match — any part of a match that goes into overtime will be disregarded, and we hence let $\mathcal{I}_m = [0; 48]$ minutes.

We wish to use the trend analysis proposed for three purposes: 1) to show how the Trend Direction Index can be used to infer real-time and post-game evaluation of the trends in a match. 2) To evaluate the Excitement Trend Index for all 1143 matches in the 2019–2020 season to provide background and reference information about the matches, and 3) to summarize ETI at the team level in order to identify group of teams more/less likely to give an exciting game.

For our first purpose of analyzing trends in individuals matches we consider the final match of the 2019-2020 season. The raw data for the running score difference between LA Lakers and Miami Heat was shown in Figure 1. Figure 2 shows the results from the post-game analysis.

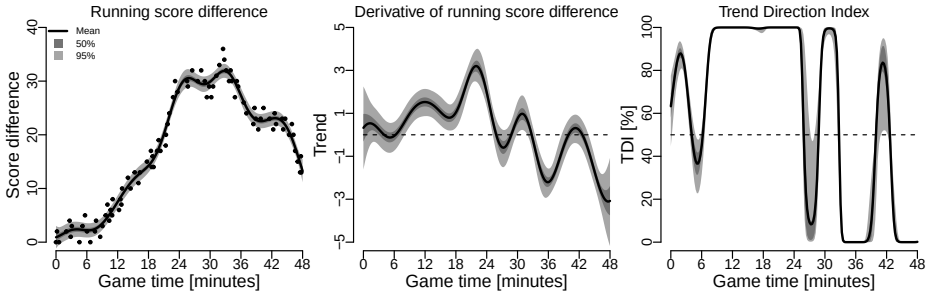

Fig. 2: Results from fitting the latent Gaussian Process to the final match between LA Lakers and Miami Heat in the 2019–2020 NBA season. Larger values in first and last panels reflect the situation where LA Lakers are doing better. The first panel shows the posterior distribution of the the latent process with the posterior means in bold. The gray areas show point-wise credibility intervals for the posterior distribution. The middle panel provides similar information for the derivative of the latent process, i.e., the posterior trend. The final panel shows the Trend Direction Index and can be used to read off probability statements about the trends in the running score difference.

Evaluating the game trends from the TDI in Figure 2 shows that LA Lakers had control of most of the match since the posterior probability of a positive trend was high throughout most of the match. Only towards the end of the third quarter did Miami Heat gain the upper hand and had a period where they fought back. In the 4th quarter from around 38 minutes to 43 minutes we can see that mean TDI increases to over 80% but the probability interval is very wide reflecting that it is difficult to say whether the trend is increasing or if it might as well just be random fluctuations in scores. Similarly for the first half of the 3rd quarter. Teams wishing to evaluate the match should primarily concentrate on periods where the latent trend and its probability interval is either close to 50% or when the trend is disadvantageous for the team.

To evaluate the overall distribution of ETIs we fitted our model to each of the 1143 matches during the season and estimated the ETI for each. The results are summarized in the left panel of Figure 3 which shows the marginal distribution of the median posterior Excitement Trend Indices. The solid line shows the fit of a Gaussian mixture model with four components, where the number of components were determined by sequential bootstrapped likelihood ratio testing (Scrucca et al. (2016)). The marginal distribution of ETIs is right skewed (skewness = 0.53) with a range of $[0.23; 26.28]$, and an median of 10.09 (mean = 10.62, SD = 4.52). This implies that the time-varying score differences of the games in the season changes monotonicity approximately 10 times during a game on average but with a large variation between matches. For comparison, the final match between LA Lakers and Miami Heat shown in Figure 2 has an ETI of **XXX**.

We wished to examine if there was a calendar time effect on game excitement as the season progressed in order to investigate if we could find that games became more close as the teams fought to stay in the competition to enter the final playoffs or if we could detect a fatigue over the season. The right panel of Figure 3 shows the

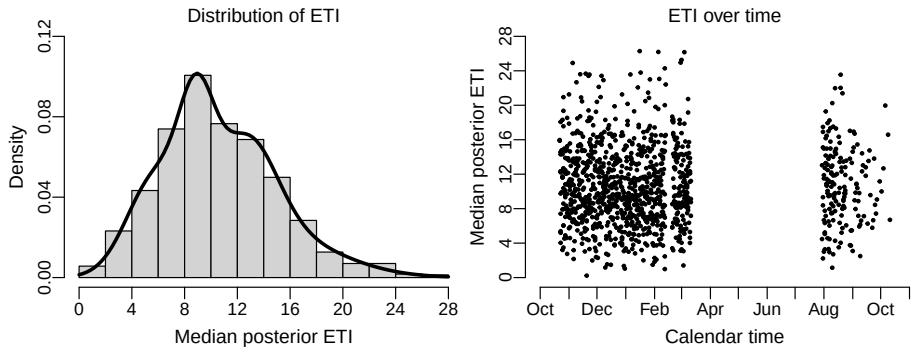

Fig. 3: Histogram and Gaussian mixture distribution estimate of the distribution of 1143 median posterior ETIs from the NBA 2019–2020 season (left panel) and the median posterior ETIs as a function of calendar time from October 22, 2019 to October 11, 2020 (right panel).

median posterior ETIs as a function of calendar time at which the matches were played. Besides illustrating the gap from the Covid-19 hiatus, the figure shows that the excitement indices are relatively evenly distributed throughout the season.

If we rank the matches from lowest to highest ETI we can extract the analyses for representative matches spanning the ETI range. Figure 4 shows the analyses of our proposed method for the minimum, 1st, 2nd, 3rd quantile, and maximum ETI matches. It is quite clear from the observed running score differences, posterior trends and the TDI that these five matches represent substantially different game experiences. The second row of Figure 4 for the Charlotte Hornets vs Chicago Bulls match indicates that while the Hornets were leading for almost the first three quarters then the game trend was rather flat since the two teams kept the pace with each other for most of the game. In contrast, the match between New Orleans Pelicans vs Utah Jazz (5th row in Figure 4) showed trends that varied direction frequently and where the TDI showed alternating periods of scoring bursts making it a very exciting and unpredictable game. **Okay ... lidt tykt smurt på her.**

Summarizing the posterior median ETIs at the team level across all matches during the season lead to comparable values for all 30 teams. Table 1 shows the summary statistics for all 30 teams ordered by their season average excitement. The New Orleans Pelicans had the highest median posterior ETI averaged over the season with an average median posterior ETI of 11.67 (SD = 4.91, IQR = $[3.09; 23.57]$), while the Charlotte Hornets had the lowest average median posterior ETI with a value of 9.29 (SD = 3.74, IQR = $[1.96; 15.98]$). This small fluctuation of averages suggests that the teams are comparable in terms of excitement when averaging across the season, and that the major source of variation in excitement during the season (as seen in Figure 3) is governed by the specific matches.

Although the team averages in Table 1 show limited variability it is of interest to estimate a number of subgroups among the teams that exhibited similar degree of excitement on average during the season – effectively clustering the teams. This would enable fans, promoters, and sponsors to infer which teams were more likely to

partake in an exciting game. The problem is mathematically equivalent to looking at the relationship between the median ETIs as the outcome in a linear regression model where the explanatory categorical variable ranges over the set of all partitions of 30 elements, and as the objective we seek the smallest number of partitions that best explains the observed ETIs by comparing all possible splits of the ranked teams for a given number of partitions. This will then define subgroups of teams.

As a optimization criterion we used root mean squared error of predicting based on leave-one-out cross-validation, denoted $\text{RMSEP}_{\text{LOO-CV}}$. [**??? finish this ???**].

Our sequential optimization procedure showed that the leave-one-out cross-validated root mean squared error of prediction stabilized at four subgroups ($\text{RMSEP}^{C=2}_{\text{LOO-CV}} = 4.500$, $\text{RMSEP}^{C=3}_{\text{LOO-CV}} = 4.497$, $\text{RMSEP}^{C=4}_{\text{LOO-CV}} = 4.496$, and subsequently for $C = 5, \ldots, 8$ it remained at the same value). The labels of these groups are shown in the rightmost column in Table 1. The result is thus an identification of three estimated change-points in the ranking of the teams according to the average median posterior ETI. The most noticeable result is that Charlotte Hornets constitute a singleton since that team has substantial lower ETI than the team with the second lowest ETI. To have a higher probability of seeing an exciting game it would be wise to avoid the Hornets.

5 Discussion

Some summarization goes on here. We have introduced etc...

We could define a weighted Excitement Trend Index, WETI_m, so that changes in monotonicity of the score differences are i) weighted higher towards the end of the game and ii) weighted lower if one team is already far away of the other team as measured by the absolute value of the posterior mean μ_{d_m}. This motivates a modification of the definition of the ETI in Equation (5) to the following weighted form

$$\text{WETI}_m \mid \text{⚽}_m = \int_{\mathcal{I}_m} d\text{ETI}_m(t \mid \text{⚽}_m) w(t, |\mu_{d_m}(t)|) \mathrm{d}t$$

where $w \colon \mathcal{I}_m \times \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ is a weight function being increasing in its first variable and decreasing in its second variable. Such weight functions could be constructed as a product of two kernel functions defined on their individual domains and with bandwidths based on studies of psychological perception.

A different approach would be to define team-specific excitement index nested with a match. Here we would only look at the **up**-crossings at zero of df_m and we would get two excitement indices for each match $(\text{ETI}_{am}, \text{ETI}_{bm})$. for teams a and b. This would somehow reflect how exciting each team were in match m with respect to chancing the sign of the score differences in their favor.

Bibliography

Baboota, Rahul, and Harleen Kaur. 2018. "Predictive Analysis and Modelling Football Results Using Machine Learning Approach for English Premier League." *Interna-*

tional Journal of Forecasting 35 (March). https://doi.org/10.1016/j.ijforecast.2018.01.003.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Cattelan, Manuela, Cristiano Varin, and David Firth. 2013. "Dynamic Bradley–Terry Modelling of Sports Tournaments." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62 (1): 135–50. https://doi.org/https://doi.org/10.1111/j.1467-9876.2012.01046.x.

Chen, Tao, and Qingliang Fan. 2018. "A Functional Data Approach to Model Score Difference Process in Professional Basketball Games." *Journal of Applied Statistics* 45 (1): 112–27.

Chen, Yaqing, Matthew Dawson, and Hans-Georg Müller. 2020. "Rank Dynamics for Functional Data." *Computational Statistics & Data Analysis*, 106963.

Cramer, Harald, and M. R. Leadbetter. 1967. *Stationary and Related Stochastic Processes – Sample Function Properties and Their Applications.* John Wiley & Sons, Inc.

Ekstrøm, Claus Thorn, Hans Van Eetvelde, Christophe Ley, and Ulf Brefeld. 2020. "Evaluating One-Shot Tournament Predictions." *Journal of Sports Analytics* Preprint (Preprint): 1–10. https://doi.org/10.3233/JSA-200454.

Gabel, Alan, and Sidney Redner. 2012. "Random Walk Picture of Basketball Scoring." *Journal of Quantitative Analysis in Sports* 8 (1).

Groll, Andreas, Christophe Ley, Gunther Schauberger, and Hans Van Eetvelde. 2019. "A hybrid random forest to predict soccer matches in international tournaments." *Journal of Quantitative Analysis in Sports* 15: 271–88.

Gu, Wei, and Thomas L. Saaty. 2019. "Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments." *Journal of Systems Science and Systems Engineering* 28 (3): 317–43. https://doi.org/10.1007/s11518-018-5395-3.

Jensen, Andreas Kryger, and Claus Thorn Ekstrøm. 2020a. "GitHub Repository for Having a Ball." 2020. https://github.com/aejensen/Having-a-Ball.

———. 2020b. "Quantifying the Trendiness of Trends." *Journal of the Royal Statistical Society: Series C.*

Karlis, Dimitris, and Ioannis Ntzoufras. 2003. "Analysis of Sports Data by Using Bivariate Poisson Models." *Journal of the Royal Statistical Society: Series D (The Statistician)* 52 (3): 381–93. https://doi.org/10.1111/1467-9884.00366.

Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian Processes in Machine Learning.* MIT Press.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal* 8 (1): 289–317. https://doi.org/10.32614/RJ-2016-021.

Sports Reference LLC. 2020. “Basketball Reference.” 2020. https://www.basketball-reference.com/.

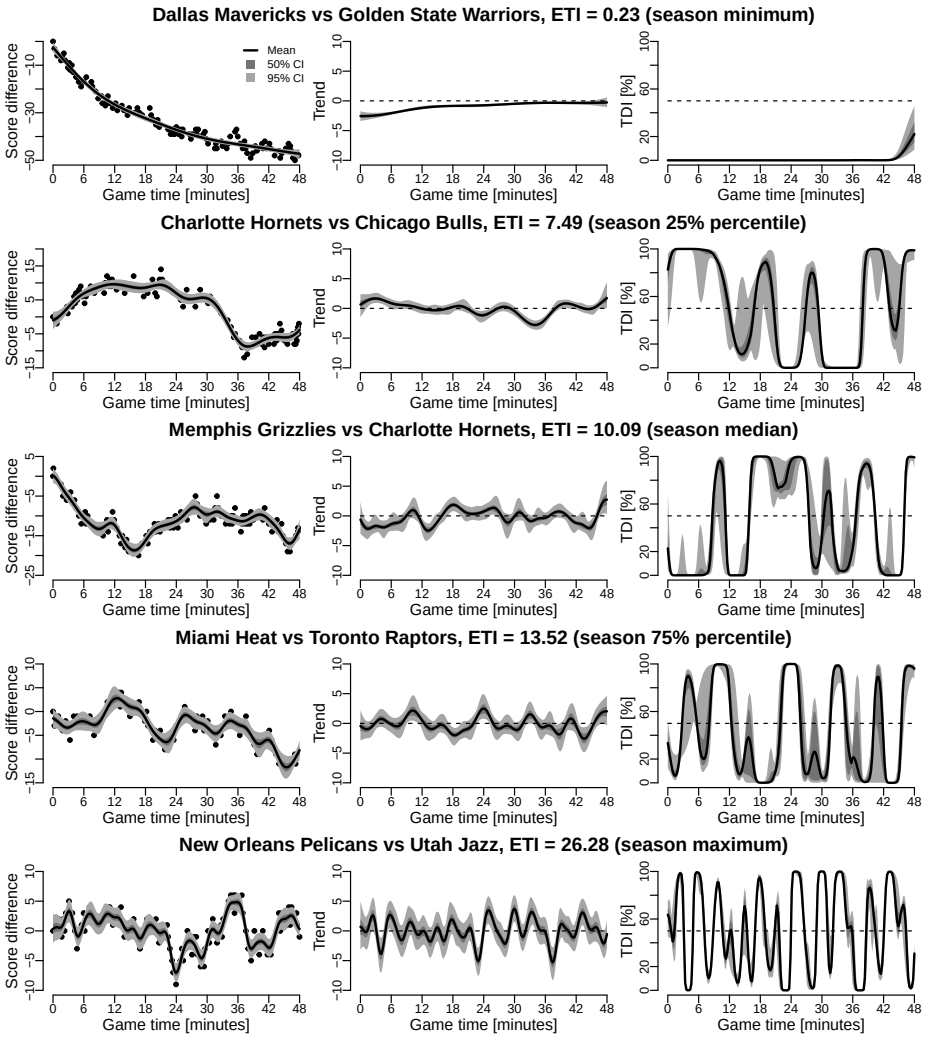

Fig. 4: Observed score differences with posteriors of the latent processes (left panels), posterior trends (middle panels) and posterior Trend Direction Indices (right panels) for five games from the NBA season 2019–2020 with median posterior Excitement Trend Indices corresponding to the 0%, 25%, 50%, 75%, and 100% percentiles of the distribution of all games in the season. Gray regions depict 50% and 95% point-wise credible intervals.

Table 1: Team specific median posterior Excitement Trend Indices summarized across all their matches in the 2019–2020 NBA season ordered by the season average. Additionally, standard deviations and 2.5%, 50%, and 97.5% percentiles are shown. The group column shows the clustering induced by the sequential optimization procedure.

	Average	SD	2.5%	50%	97.5%	Group
New Orleans Pelicans	11.67	4.91	3.09	11.60	23.57	A
Washington Wizards	11.39	4.47	2.98	11.64	18.45	A
San Antonio Spurs	11.33	5.29	3.52	9.88	23.41	A
Oklahoma City Thunder	11.31	5.23	2.87	11.33	23.73	A
Portland Trail Blazers	11.21	4.46	3.31	10.82	18.47	A
Los Angeles Lakers	11.15	4.42	3.61	10.53	19.81	A
Minnesota Timberwolves	11.08	4.28	4.01	10.69	19.43	A
Philadelphia 76ers	11.08	4.74	2.73	10.35	21.56	A
Memphis Grizzlies	10.95	4.55	3.60	10.38	20.84	B
Cleveland Cavaliers	10.89	4.50	4.90	10.42	22.40	B
Utah Jazz	10.80	5.53	2.98	9.55	23.56	B
Houston Rockets	10.68	4.12	3.42	10.74	19.17	B
Chicago Bulls	10.66	3.98	2.96	10.63	17.45	B
Toronto Raptors	10.64	4.53	2.54	10.32	21.83	B
Boston Celtics	10.61	4.66	2.81	10.20	20.94	B
Milwaukee Bucks	10.57	4.29	2.81	9.59	18.14	B
Orlando Magic	10.56	4.90	2.72	10.40	21.04	B
Los Angeles Clippers	10.54	4.14	4.04	10.73	20.58	B
Golden State Warriors	10.35	4.55	3.35	10.23	18.73	C
Miami Heat	10.33	4.37	2.67	9.83	18.56	C
Indiana Pacers	10.29	4.11	3.16	10.22	19.41	C
Atlanta Hawks	10.28	4.30	2.92	9.77	18.30	C
Dallas Mavericks	10.28	4.98	3.00	9.04	22.07	C
Phoenix Suns	10.25	4.32	3.78	9.29	19.85	C
Brooklyn Nets	10.15	4.71	2.93	10.06	18.95	C
New York Knicks	10.12	4.06	2.84	9.87	18.50	C
Denver Nuggets	10.12	4.39	3.42	9.56	19.61	C
Sacramento Kings	10.00	4.09	3.97	9.29	18.91	C
Detroit Pistons	9.86	4.01	3.73	9.43	17.08	C
Charlotte Hornets	9.29	3.74	1.96	9.41	15.98	D