PLOS ONE

RESEARCH ARTICLE

A Bayesian approach to time-varying latent strengths in pairwise comparisons

Blaž Krese[1]*, **Erik Štrumbelj**[2]

1 GEN-I, d.o.o., Ljubljana, Slovenia, **2** Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

* blaz.krese@gen-i.eu

Citation: Krese B, Štrumbelj E (2021) A Bayesian approach to time-varying latent strengths in pairwise comparisons. PLoS ONE 16(5): e0251945. https://doi.org/10.1371/journal.pone.0251945

Editor: Inés P. Mariño, Universidad Rey Juan Carlos, SPAIN

Received: January 29, 2021

Accepted: May 1, 2021

Published: May 20, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Funding: Blaž Krese is employed by GEN-I, d.o.o., Ljubljana, Slovenia. The funder provided support in the form of salaries for author BK, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section. Erik Štrumbelj acknowledges the financial support from the

Abstract

The famous Bradley-Terry model for pairwise comparisons is widely used for ranking objects and is often applied to sports data. In this paper we extend the Bradley-Terry model by allowing time-varying latent strengths of compared objects. The time component is modelled with barycentric rational interpolation and Gaussian processes. We also allow for the inclusion of additional information in the form of outcome probabilities. Our models are evaluated and compared on toy data set and real sports data from ATP tennis matches and NBA games. We demonstrated that using Gaussian processes is advantageous compared to barycentric rational interpolation as they are more flexible to model discontinuities and are less sensitive to initial parameters settings. However, all investigated models proved to be robust to over-fitting and perform well with situations of volatile and of constant latent strengths. When using barycentric rational interpolation it has turned out that applying Bayesian approach gives better results than by using MLE. Performance of the models is further improved by incorporating the outcome probabilities.

Introduction

Modelling pairwise comparisons is an important practical problem and well established in research literature [1, 2]. The foundations were built in the 1950s by Bradley and Terry [3] and Luce [4], though the first idea goes back to Thurstone [5]. The classical approach is the Bradley-Terry model [3]. The model links the pairwise comparison probabilities with the compared objects' latent strengths, which are in the model's most simple variant assumed to be constant.

The Bradley-Terry model has been extended in several ways: handling ties [6], ranking individual players in multi-player competitions [7, 8], and stochastic non-transitivity of comparisons [9]. It has also been shown that Bradley-Terry model can be seen as a special case of a more general model. A very recent example of such treatment demonstrates a pairwise comparison model where the Weibull distribution is applied [10]. Another common generalization is to allow for the latent strengths to vary with time and it is the focus of our work. The quintessential application domain for time-varying strength models is sports, where ranking is important both for seeding competitions and for fan engagement. However, a player's strength

Slovenian Research Agency (research core funding No. P5-0410).

Competing interests: Blaž Krese is employed by GEN-I, d.o.o., Ljubljana, Slovenia. The funder provided support in the form of salaries for author BK, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

changes with age, experience, fatigue, and injuries. And a team's strength changes with players joining or leaving a team.

The classical time-varying approach is the ELO rating, designed by Arpad Elo [11, 12]. It was adopted, for example, by the International Chess Federation (FIDE) [13] and UEFA [14]. The ELO rating uses a scaled version of the Bradley-Terry model. After each comparison the underlying latent strength is changed with accordance to the previous strength and the output of the comparison. Glickman developed a non-iterative Bayesian algorithm [15]. This model assumes a normal distribution of the latent strengths conditional on the strength at the previous comparison with the standard deviation dependent on the elapsed time between comparisons. Based on this algorithm the Glicko and Glicko-2 rating systems ware developed, where the latter improves on ability to capture sudden changes [16]. One downside of incremental algorithms is that covariance is not taken into account when approximating probability distributions of latent strengths. This was addresses by Coulom [17] who used a Wiener process for the prior of latent strengths and applied it to the Bradley-Terry model, using maximum a posteriori (MAP) inference with Newton's approximation method. This approach has proven to be better than ELO and Glicko when applied to the game of Go. More recently, Baker and McHale applied deterministic approach to time-varying latent strengths by using barycentric rational interpolation (BRI) [18]. This approach was applied to football where pairwise comparisons were based on the Poisson distribution of the number of goals scored. Baker and McHale also applied BRI to tennis [19], using a symmetric beta distribution for ranking, deduced as a special case of Stern's gamma model, which can also be reduced to the Bradley-Terry model or Thurstone model. They also showed that BRI outperformed spline interpolation. A model based on the number of goals scored was also used by Owen [20] and Koopman [21], who used an incremental approach to model time dependence of latent strengths with a focus on outcome forecasting rather than hindcasting as in the case of Baker. Cattelan et al. [22] also used an incremental approach to model team's ability by using an exponentially weighted moving average processes applied to the Bradley-Terry model. Inference was done via maximum likelihood estimation and they applied their model to basketball and football.

In this paper we extend the Bradley-Terry model to allow for time-varying strengths by combining it with barycentric rational interpolants (BRI) [23] or Gaussian processes (GP) [24]. We also extend the model to handle not only binary comparison outcome data but also outcome probabilities, if available to be derived, for example, from bookmakers' odds. Compared to the majority of related work which is motivated by forecasting, our approach addresses hindcasting. When the focus is on forecasting, the main goal is to minimize the short-term prediction error and for these purpose modelling is based on incremental approach. However, incremental methods are not suitable for hindcasting where it is vital to take into account the covariance between model's parameters. With hindcasting we are not interested in just the next game output, but rather in the underlying dynamics of latent strengths where a longer period needs to be considered. Research with focus on hindcasting is sparse—Baker and McHale [18, 19] and Coulom [17] who model time-varying latent strengths deterministically with interpolation and the Wiener process, respectively. Compared to Baker and McHale [18, 19] we combined barycentric rational interpolation (BRI) with the Bradley-Terry model and we use Bayesian inference. We also model time-varying strengths with Gaussian processes (GPs). This is similar to Coulom [17], but with two significant differences. First, using GPs is more general, because a Wiener process is a special case of GPs when the kernel function is given by $k(t, t') = \min(t, t')$ [25]. And second, we utilize Markov Chain Monte Carlo (MCMC) instead of structural approximation of the posterior and MAP estimation. Notably, our Bayesian models are implemented in Stan [26] and we utilize Markov Chain Monte Carlo for inference. We empirically evaluate and compare the models on toy data and

two real-world sports data sets: ATP (Association of Tennis Professionals) tennis and NBA (National Basketball Association) basketball.

Methodology

The Bradley-Terry model

Pairwise comparison data are a set of observations, where each observation is the outcome of a pairwise comparison between two objects, where one of the objects is deemed to be superior to the other. We will not consider ties in this paper.

The classical model for such data is the Bradley-Terry model [3] which assumes that the comparison outcome probabilities are governed by unobserved (latent) strengths of the objects. Given a comparison between objects a and b, we have

$$P(a \text{ is superior to } b) \triangleq \frac{\theta_a}{\theta_a + \theta_b}, \tag{1}$$

where θ_a and θ_b are the latent strengths of objects a and b, respectively. In its most basic variant, these strengths are assumed to be constant.

Introducing time-varying latent strengths. We will focus on the extensions of the Bradley-Terry model where the latent strengths vary with time. The pairwise comparisons observations are then 4-tuples (t_i, a_i, b_i, y_i), where $t_i \in \mathbb{R}$ is the time when the comparison was made, $a_i, b_i \in \{1, \ldots, K\}$ are the two objects being compared, from a set of K objects, and $y_i \in \{0, 1\}$ is the outcome of the comparison. If object a_i was deemed to be superior to object b_i, then $y_i = 1$, otherwise $y_i = 0$. Times t_i are not necessarily unique—two comparisons can be made at the same time.

The Bradley-Terry model is a non-deterministic model. The comparison outcome is modeled as a random variable Y_i with support $\{0, 1\}$. In general, the probability mass function of Y_i is

$$p(y_i|\boldsymbol{\theta}, a_i, b_i, t_i) = P(Y_i = y_i|\boldsymbol{\theta}, a_i, b_i, t_i), \tag{2}$$

but because Y_i is Bernoulli, we will use the shorthand notation

$$p_i \triangleq p(1|\boldsymbol{\theta}, a_i, b_i, t_i), \tag{3}$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}(t) = (\theta_1(t), \ldots, \theta_K(t))$ and $\theta_j(t)$ are the unknown time-dependent latent strengths of the objects.

We can now generalize Eq (1) to

$$p_i = \frac{\theta_{a_i}(t_i)}{\theta_{a_i}(t_i) + \theta_{b_i}(t_i)}. \tag{4}$$

In Eq (4) we explicitly write t_i to stress the latent strengths' dependency on time. To simplify the notation, we will from now on assume this time dependency and omit the times whenever possible.

In order for p_i to be probabilities, the latent strengths have to be positive. Because it is more convenient to work with real parameters θ, we typically rewrite Eq (4) as

$$p_i = \frac{e^{\theta_{a_i}}}{e^{\theta_{a_i}} + e^{\theta_{b_i}}} = \frac{1}{1 + e^{\theta_{b_i} - \theta_{a_i}}} = \text{logit}^{-1}(\theta_{a_i} - \theta_{b_i}), \tag{5}$$

where logit^{-1} is the cumulative distribution of the standard logistic distribution, also known as

the inverse logistic function or inverse logit:

$$\text{logit}^{-1}(x) \triangleq \frac{1}{1+e^{-x}}. \tag{6}$$

This Bradley-Terry model can be viewed as logistic regression with one input variable—the difference between the latent strengths of objects being compared.

Model identifiability. Since the outcome probabilities depend only on the difference in latent strengths they are invariant to translation. In order to be able to identify parameters θ, we have to set a reference. We set the latent strength of the K-th object to be 0 [22].

Covariates. In Eq (5) the outcome probability depends solely on the latent strengths of the two objects being compared. In practice, other factors might affect the outcome. For example, home team advantage or weather. We will account for these covariates with a linear term

$$p_i = \text{logit}^{-1}(\theta_{a_i} - \theta_{b_i} + \boldsymbol{\beta}^\top \mathbf{x}_i), \tag{7}$$

where $\mathbf{x}_i$ is a vector of covariates for the i-th observation and $\boldsymbol{\beta}$ is a vector of coefficients. Covariates are assumed to be known and measured without error and coefficients are parameters of the model.

Note that the purpose of this work is not to study the effect that different covariates might have in a particular domain. However, for NBA data we do include a covariate for home team advantage, which is known to have a strong effect on sports match outcome probabilities. The home team advantage covariate $x_{\text{hta},i}$ can be coded as + 1, −1, or 0 when team a is playing at home, team b is playing at home, or when the game is played in a neutral venue, respectively.

Baseline model (BASE)

Our baseline for comparison will be the Bradley-Terry model where we assume that an object's latent strength is constant $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ and we fit the parameters using maximum likelihood estimation. Given n observations, the likelihood is

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \tag{8}$$

where the $p_i = \text{logit}^{-1}(\theta_{a_i} - \theta_{b_i} + \boldsymbol{\beta}^\top \mathbf{x}_i)$ as in Eq (7). Then the log-likelihood is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i). \tag{9}$$

Finding the maximum likelihood estimates reduces to the optimization problem

$$\begin{aligned} (\widehat{\boldsymbol{\theta}, \boldsymbol{\beta}})_{\text{BASE}} &= \arg\max_{(\boldsymbol{\theta}\boldsymbol{\beta})} \ell(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}) \\ \theta_K &= 0, \end{aligned} \tag{10}$$

which we solved using L-BFGS optimization.

Barycentric rational interpolation model (BRI)

BRI is an alternative to splines. A detailed comparison between BRI and splines is discussed in [27]. BRI is infinitely differentiable, which is a drawback when modelling a process with sudden changes in values. Still, it has been shown that BRI has the same or slightly lower errors in curve fitting than splines. BRI was used to model the attack and defence ability of football

teams combined with comparisons of goals scored by the teams modelled with Poisson distribution [18]. A similar study was conducted for ranking tennis players [19].

We start by introducing m nodes in time $(t_k^*, \lambda_k),\ k = 1, 2, \ldots, m$, where λ_k represents the quantity of interest at time t_k^*. We use the t^* notation to make it explicit that these nodes need not correspond to the times of the observations in our data. In practice, we typically use fewer nodes than observations.

The purpose of BRI is to interpolate between these nodes in order to get the quantity of interest at any time. In our case the quantity of interest are unobserved—the latent strengths of objects. We will perform BRI for each object separately. We then write the evolution of the j-th object's latent strength over time in the general barycentric form by interpolation between coordinates [27]

$$\theta_j(t) = \frac{\sum_{k=1}^{m_j} w_{jk}\lambda_{jk}/(t - t_{jk}^*)}{\sum_{k=1}^{m_j} w_{jk}/(t - t_{jk}^*)}. \tag{11}$$

The number of nodes m_j does not have to be the same for every object, but for our applications we do not lose by assuming that it is. Selecting the number and location of the nodes is analogous to spline interpolation [27]. Domain knowledge can be used but automated optimal placement is infeasible and has to be dealt with heuristically. We positioned the nodes equally spaced in time and empirically selected the best m from a finite set of possibilities. As a consequence, the notation t_{jk}^* reduces to t_k^* and weights are given in a simpler form $w_{jk} = (-1)^k, \forall j$ [18].

The general form of the log-likelihood is similar to Eq (9) but $\boldsymbol{\lambda} = \{\lambda_{jk;}\}$ are now the parameters

$$\ell(\boldsymbol{\lambda}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i), \tag{12}$$

where $p_i = \text{logit}^{-1}(\theta_{a_i} - \theta_{b_i} + \boldsymbol{\beta}^\top \mathbf{x}_i)$ and

$$\theta_j(t) = \frac{\sum_{k=1}^{m} (-1)^k \lambda_{jk}/(t - t_k^*)}{\sum_{k=1}^{m} (-1)^k/(t - t_k^*)}, \forall j \neq K. \tag{13}$$

Finding the maximum likelihood estimates reduces to the optimization problem

$$\begin{aligned} \widehat{(\boldsymbol{\lambda}, \boldsymbol{\beta})}_{\text{BRI}} &= \arg\max_{(\boldsymbol{\lambda}, \boldsymbol{\beta})} \ell(\boldsymbol{\lambda}, \boldsymbol{\beta}; \mathbf{y}) \\ \theta_K(t) &= 0, \forall t, \end{aligned} \tag{14}$$

which we solved using L-BFGS optimization.

Bayesian barycentric rational interpolation model ($\text{BRI}_{\text{bayes}}$). We also inferred from the BRI model using the Bayesian framework, treating the $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ as random variables. The

model and prior distributions are

$$\begin{aligned} y_i|\boldsymbol{\lambda}, \boldsymbol{\beta}, t_i, a_i, b_i, \mathbf{x}_i, t^* &\sim \text{Bernoulli}(\text{logit}^{-1}(\theta_{a_i} - \theta_{b_i} + \boldsymbol{\beta}^\top \mathbf{x}_i)) \\ \theta_j(t) &= \frac{\sum_{k=1}^m (-1)^k \lambda_{jk}/(t - t_k^*)}{\sum_{k=1}^m (-1)^k/(t - t_k^*)}, \forall j \neq K \\ \theta_K(t) &= 0, \forall t, \\ \boldsymbol{\lambda}_j &\sim \mathcal{N}(\boldsymbol{\mu}_\lambda, \sigma_\lambda^2 \mathbf{I}), \forall j \neq K \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}). \end{aligned} \tag{15}$$

It is standard to assume that $\boldsymbol{\beta}$ coefficients are centered around 0. The prior constants σ_λ^2 and σ_β^2 are user-defined constants. If little or no prior information is available, they can be set to some relatively large value. In the case of $\boldsymbol{\beta}$ this value depends on the scale of the covariates. In the case of $\boldsymbol{\lambda}$ this value can be small, because even differences in the order of 10 result in near 1 (or 0) probabilities due to the inverse logit transformation. Note that this model could easily be extended to use regularization on the covariates by placing a hyper-prior on $\boldsymbol{\beta}$.

We implemented the model in the Stan probabilistic programming language and inferred from it using the built in variant the No-U-turn Sampler (NUTS), an extension of the Hamiltonian Monte Carlo sampling algorithm [26, 28, 29].

Gaussian process model (GP)

GPs are a well-studied field with a rich theory [24]. The shape of a GP is determined primarily by its kernel function which is very flexible. By applying different kernel functions we can get for instance a Wiener proces [25] or a certain spline [30]. GPs are also closely connected to some of the more well-known models such as neural networks or support vector machines, but are more intuitive and easy to interpret [24]. On the other hand applying GPs is time demanding due to the covariance matrix inversion which is $\mathcal{O}(n^3)$ where n is the number of covariate points [24].

Instead of using BRI we now place a GP prior on each object's latent strength

$$\theta_j(t) \sim \mathcal{GP}(m(t), k(t, t')), \forall j,$$

where $m(t)$ is the mean function and $k(t, t')$ is the covariance function [24]. The mean function is usually taken to be $m(t) = 0; \forall t$.

The likelihood of the model is the same as in Eq (8), so the posterior distribution is

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{y}) \propto L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}) \prod_{j=1}^K \mathcal{GP}(0, k(t, t'))) = \left(\prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1 - y_i} \right) \prod_{j=1}^K \mathcal{GP}(0, k(t, t')), \tag{16}$$

where $p_i = \text{logit}^{-1}(\theta_{a_i} - \theta_{b_i} + \boldsymbol{\beta}^\top \mathbf{x}_i)$ and we abuse the notation $\mathcal{GP}$ to denote the multivariate normal (MVN) probability density function of a GP.

To predict latent strengths $\boldsymbol{\theta}_* \triangleq \boldsymbol{\theta}(t_*)$ for times t_*, we have to compute the posterior predictive density [31]

$$p(\boldsymbol{\theta}_*, |\mathbf{y}) = \int p(\boldsymbol{\theta}_*|\boldsymbol{\theta}) \left(\int p(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{y}) \, \mathrm{d}\boldsymbol{\beta} \right) \mathrm{d}\boldsymbol{\theta}, \tag{17}$$

where $p(\boldsymbol{\theta}|\mathbf{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{y}) d\boldsymbol{\beta}$ is the marginal posterior obtained by integrating the posterior

density over $\boldsymbol{\beta}$ (and any kernel hyper-parameters). The conditional multivariate Gaussian distribution $p(\boldsymbol{\theta}_*|\boldsymbol{\theta})$ is given by

$$\boldsymbol{\theta}_*|\boldsymbol{\theta} \sim \mathcal{N}(K_{t_*,t}\, K_{t,t}^{-1}\, \boldsymbol{\theta}, K_{t_*,t_*} - K_{t_*,t}\, K_{t,t}^{-1}\, K_{t,t_*}^{\top}). \tag{18}$$

$K_{\cdot,\cdot}$ are covariance matrices obtained by evaluating kernel functions on different combinations of given times t and t_*.

Eq (17) is only tractable when the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is normal [31, 32], so no closed form solution exists for our model and we have to resort to numerical methods. One approach is to use structural approximation methods such as Laplace approximation or variational inference, see [24, 31] for a quick overview. For instance, Laplace approximation algorithm uses a quadratic approximation and by optimization locates the mode of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$. Variational inference minimizes the divergence between a Gaussian approximation and the posterior distribution, but the likelihood function has to be factored as $p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\theta_i)$ [31]. These methods can be quite accurate, especially when the posterior is uni-modal, but they can also give biased results when posterior distribution has a more complex shape. To overcome restrictions of structural approximations we use MCMC sampling algorithms. These methods are more computationally intensive but guarantee convergence in distribution to the posterior in the limit of long runs [31].

The model and prior distributions are governed by

$$\begin{aligned}
y_i|\boldsymbol{\theta}, \boldsymbol{\beta}, t_i, a_i, b_i, \mathbf{x}_i &\sim \mathsf{Bernoulli}(p_i) \\
p_i &= \mathrm{logit}^{-1}(\theta_{a_i} - \theta_{b_i} + \boldsymbol{\beta}^{\top}\mathbf{x}_i) \\
\theta_j|\ell, \sigma &\sim \mathsf{MVN}(0, K(t, t'|\sigma, \ell)),\ \forall j \neq K \\
\theta_K(t) &= 0,\ \forall t \\
\sigma &\sim \mathcal{N}(0,\ \sigma_\sigma^2),\ \sigma > 0 \\
\ell &\sim \mathsf{GIG}(a_{\mathrm{gig}}, b_{\mathrm{gig}}, q_{\mathrm{gig}}) \\
\beta &\sim \mathcal{N}(\mathbf{0},\ \sigma_{\boldsymbol{\beta}}^2\, \mathbf{I}).
\end{aligned} \tag{19}$$

The choice of prior distributions requires additional explanation. For the kernel function $k(t, t'|\sigma, \ell)$ we considered the most commonly used squared exponential kernel

$$k(r) = \sigma^2 \exp\left(\frac{-r^2}{2\ell^2}\right), \tag{20}$$

where $r = |t - t'|$, and three Matérn kernels

$$\begin{aligned} k_{\nu=\frac{1}{2}}(r) &= \sigma^2 \exp\left(-\frac{r}{\ell}\right), \\ k_{\nu=\frac{3}{2}}(r) &= \sigma^2 \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right), \\ k_{\nu=\frac{5}{2}}(r) &= \sigma^2 \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right). \end{aligned} \tag{21}$$

Note that $\lim_{\nu \to \infty} k_\nu(r) = k(r)$. Each kernel also has hyper-parameters that need to be properly chosen, that are deviation σ and length-scale ℓ. For σ we have set prior mean to 0, but only consider positive non-zero values. This choice is due to the fact that latent strength can either be close to constant corresponding to stagnation or very wavy when some significant changes occur.

We put a generalized inverse Gaussian (GIG) prior on the length-scale ℓ estimation. The GIG probability density function is given by

$$p(x \mid a, b, q) = \frac{\left(\frac{a}{b}\right)^{\frac{q}{2}}}{2K_q(\sqrt{ab})} x^{q-1} \exp\left(-\frac{1}{2}\left(ax + \frac{b}{x}\right)\right), \tag{22}$$

where $x, a, b \in \mathbb{R}^+$, $q \in \mathbb{Z}$ and K_q represents a modified Bessel function of second kind. We chose the GIG distribution, because it has a sharp left tail putting very little probability mass on close-to-zero length-scales. The right-hand side the GIG has a thin tail which allows us to keep out the very large length-scales. We set $q_{\text{gig}} = 1$ and determined a_{gig} and b_{gig} by optimization such that the mode of the GIG was equal to the distance between time nodes (see subsection Auxiliary nodes for more efficient computation). Fig 1 shows how the parameters a_{gig} and b_{gig} allow for enough flexibility for our purposes even when keeping q_{gig} fixed to 1.

Gaussian process model with outcome probabilities (GP_{prob}). Sometimes additional data are available in the form of probabilistic predictions $\hat{p}_i$, which estimate the unknown outcome probabilities p_i. For example, probabilities derived from odds in sports, which are known to be good estimates of outcome probabilities [33].

Probabilistic predictions, even if moderately biased, should provide more information than binary outcomes. We extend the model from Eq (19) to allow for the inclusion of such data:

$$\begin{aligned} \hat{p}_i|\tau &\sim \text{Beta}(p_i\,\tau, (1 - p_i)\,\tau) \\ \tau &\sim \text{Uniform}(0, \tau_{\max}). \end{aligned} \tag{23}$$

We assume that the probability estimates are beta-distributed with the mean equal to the unknown true probability. The hyper-parameter τ can be interpreted as the quality of the source of probability estimates—smaller values indicate better probabilities.

Auxiliary nodes for more efficient computation. In certain domains, for example, in most professional sports, the comparisons are few and far apart and a single comparison provides very little information about the latent strengths, so we need a relatively long period of time to get a good estimate of latent strength. In the context of GPs, we can deal with this by increasing the length-scale. However, a larger length-scale results in more correlation in the posterior and therefore less efficient exploration of the posterior via MCMC.

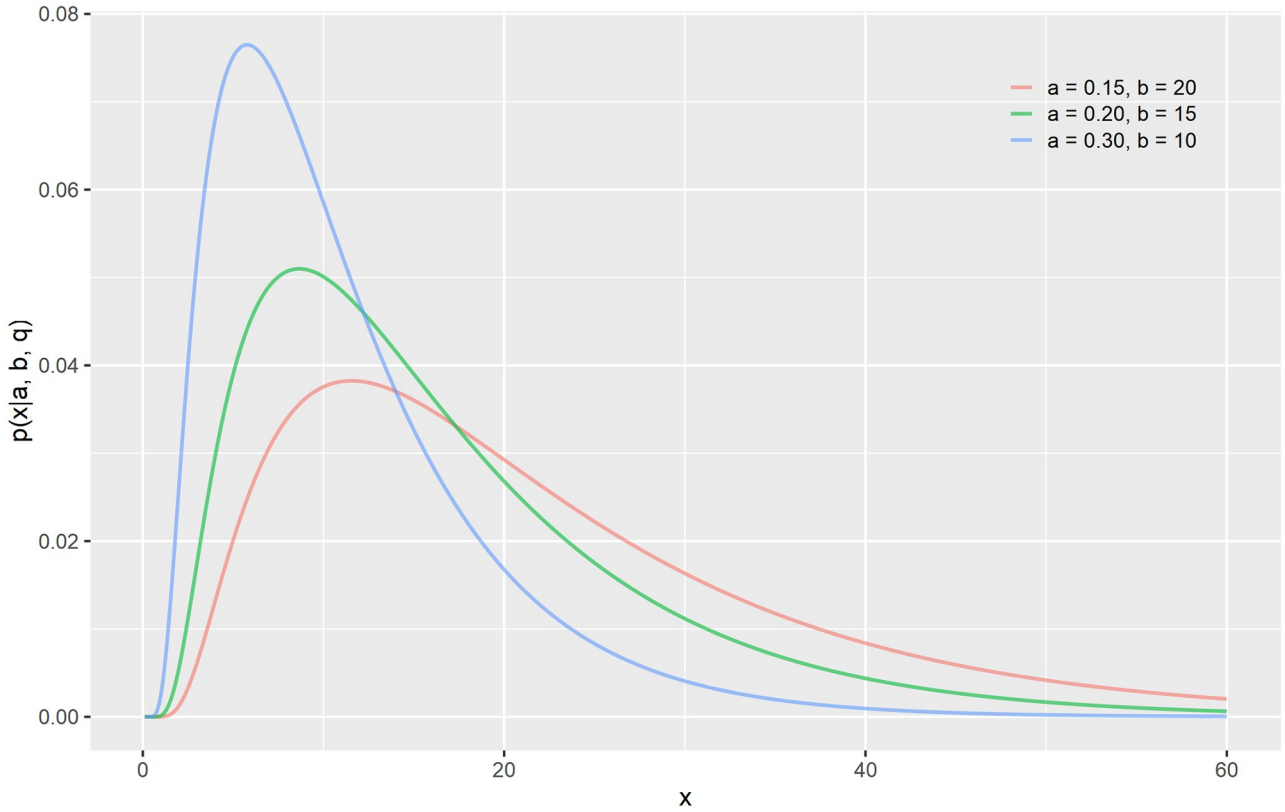

Fig 1. GIG probability density function. The GIG probability density function with $q = 1$ and different values of a and b.

https://doi.org/10.1371/journal.pone.0251945.g001

To allow for more efficient computation, we introduce auxiliary nodes (time points), similar to BRI. The likelihood is computed only at these nodes and each observation is assigned to the nearest auxiliary node. In the extreme case where an auxiliary node is placed at each observation, the method reduces to the initially described model.

Empirical evaluation

We empirically evaluated and compared the models on three data sets: a toy data set and two real world data sets: ATP (Association of Tennis Professionals) and NBA (National Basketball Association). We collected ATP data for the 20 players with the most games in the 5 seasons in the period from 2015 to 2019, for a total of 673 matches. We collected NBA game outcomes for 5904 regular season games in the 5 seasons period from 2013 to 2018. For the NBA data we also obtained bookmakers' wining odds for every match in the selected seasons period. The resources for data are the following:

- ATP: https://datahub.io/sports-data/atp-world-tour-tennis-data
- NBA: https://www.basketball-reference.com/
- NBA odds: https://www.betexplorer.com/

The raw data are available as supplementary material S1, S2, S3 and S4 Datasets.

Toy data

In the toy data set we compare 3 objects. The main feature of the data is a discontinuity in the latent strengths of the first and the second object. The latent strengths are:

$$\begin{aligned} \theta_1(t) &= -2\,\mathrm{H}(t-250)+1; \\ \theta_2(t) &= 2\,\mathrm{H}(t-167)-1; \\ \theta_3(t) &= 0, \end{aligned} \tag{24}$$

where $\mathrm{H}(\cdot)$ stands for the Heaviside function and $t \in \{0, 1, 2, \ldots, 499\}$.

The 3rd object's latent strength is held at constant value of 0. For the 1st object latent strength $\theta_1(t)$ is constant at value 1 for times $0 \leq t \leq 250$ and then jumps to value −1 for $250 < t < 500$. The shape for the 2nd object is complementary, i.e. $\theta_2(t)$ jumps from value −1 to 1 at time 167. The difference in latent strengths of value 1 corresponds to approximately a 73% chance of winning for the object with the higher latent strength.

In order to simulate comparison data we need to determine which objects are to be compared. Given three objects there are 3 possible combinations of pairwise comparisons. Each of the combinations was selected with a 50% probability for each time point $t_i \in t$. Win probabilities p_i are given with Eq (5) and the outputs of comparisons are determined with a sample from $y_i|p_i \sim \text{Bernoulli}(p_i)$.

Model evaluation and parameter tuning. We evaluated the models using the log-score and train-test (holdout) estimation repeated 10 times to account for train-test split variability. We approximated the standard error of the estimates using hierarchical bootstrap, accounting for inter-observation and inter-train-test split variability.

The models have several tunable parameters. For every experiment and every train-test split separately, their values were selected before training the model from a predetermined set of candidate values using internal train-test estimation on the training set, repeated 5 times.

A summary of experiments' settings for each data set is in Table 1. For the ATP and NBA data set we used half of the data for training. For the toy data set we used only 10% of data for the training—because these data are simulated, we could generate as many training observations as necessary to reduce the standard errors of the log-score estimates. For all three data sets we used a 90%-10% train-test split for internal selection of parameters.

Table 1. Experiments' parameters settings.

	Toy	ATP	NBA
Train data ratio [%]	10	50	50
Internal train data ratio [%]	90	90	90
#train-test splits	10	10	10
#internal train-test splits	5	5	5
Models	**BASE**, **BRI**, $\textbf{BRI}_{\text{bayes}}$, **GP**, $\textbf{GP}_{\text{prob}}$	**BASE**, **BRI**, $\textbf{BRI}_{\text{bayes}}$, **GP**,	**BASE**, **BRI**, $\textbf{BRI}_{\text{bayes}}$, **GP**, $\textbf{GP}_{\text{prob}}$
#nodes	(1, 2, 3, 5, 10, 15, 20)	(1, 5, 10, 20, 30, 50)	(1, 5, 10, 20, 30, 50)
Kernels k_ν	$\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \infty\}$	$\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \infty\}$	$\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \infty\}$
Prior parameters $(\boldsymbol{\mu}_\lambda, \sigma^2_\lambda, \sigma^2_\beta, \sigma^2_\sigma)$	$(\mathbf{0}, 4, 1, 1)$	$(\mathbf{0}, 4, 1, 1)$	$(\mathbf{0}, 4, 1, 1)$

Experiment settings and candidate tunable parameter values.

https://doi.org/10.1371/journal.pone.0251945.t001

We did not use the $\mathbf{GP}_{\text{prob}}$ model on the ATP data, because the data do not include outcome probabilities. For toy data we used a different set of nodes than with ATP and NBA data due to different time spans.

In the priors we set $\boldsymbol{\mu}_\lambda = \mathbf{0}$ since the reference object with $\theta_K(t) = 0, \forall t$ was selected randomly with no prior knowledge on relation to other objects' latent strengths. The corresponding variance was set to $\sigma_\lambda^2 = 4$. This is based on the assumption that teams in a competition are homogeneous in strength. It roughly corresponds to that a bottom 25% team has at least a 10% chance to beat a top 25% team. The variance hyper-parameter σ_β^2 for the home advantage prior was set to 1, which corresponds to ≈27% of increase in win probability. The same value was set to σ_σ^2 for the kernels' hyper-parameter σ which gives our prior belief on the rate of variation of the latent strength. For the Bayesian models we used 200 warmup and 800 sampling iterations. Effective sample sizes and R-hat diagnostics did not indicate any issues with MCMC. For the $\mathbf{GP}_{\text{prob}}$ model the hyper-parameter τ_{max} was set to 1000.

Results

Tunable parameter values. The selected tunable parameters for each train-test split are shown in Tables 5–7 in S1 Appendix, for toy, ATP, and NBA data sets, respectively:

- **Toy**: The parameters vary a lot between train-test splits. This is expected since there are discontinuities in the latent strengths and only 10% of the data were used for training. The two BRI-based models are similar as are the two **GP** models—any differences are difficult to discern due to the high variability. For the $\mathbf{GP}_{\text{prob}}$ model the number of nodes is mostly larger than with other models. Additional information in the form of probabilities allows for a smaller length-scale and a more detailed curve.
- **ATP**: A single node is consistently selected for both BRI-based models with a single exception in case of $\mathbf{BRI}_{\text{bayes}}$. The number of selected nodes for the **GP** model varies more, but 1 and 5 nodes are the most common, also suggesting a larger length-scale and that the models do not find a lot of variability in players' latent strengths.
- **NBA**: The number of nodes for the BRI-based models varies from 1 to 5 and the number of nodes for the **GP** model varies from 5 to 20. This suggests that NBA data has more variability in latent strengths than ATP data. For the $\mathbf{GP}_{\text{prob}}$ model the maximum allowed number of nodes (50) is consistently selected with only one exception where 30 nodes is selected. Additional information in the form of probabilities allows for a smaller length-scale and a more detailed curve. This also suggests that our estimate of the model performance is biased (pessimistic)—allowing a larger number of nodes could lead to even better performance.

Model performance. We organized the model performance results into upper-triangular tables where each row and column correspond to one of the models. Above-diagonal elements are the mean log-score differences between the row and column models. These elements facilitate a direct comparison of the two models. Diagonal elements are the estimated log-scores for a particular model. The results on toy data set are in Table 2.

All the models outperform the benchmark model **BASE**. In increasing order of performance, the models are **BASE**, **BRI**, $\mathbf{BRI}_{\text{bayes}}$, **GP**, and $\mathbf{GP}_{\text{prob}}$. The latter was expected to outperform the other models, because it uses more information. **GP** is better than the BRI-based models at handling the discontinuity in the latent strength. Fig 2 shows an illustrative example.

Table 2. Model performance on toy data set.

	BASE	BRI	BRI_{bayes}	GP	GP_{prob}
BASE	−0.709±0.006	**− 0.040 ± 0.033**	**− 0.105 ± 0.017**	**− 0.122 ± 0.012**	**− 0.176 ± 0.011**
BRI		−0.669±0.032	**− 0.066 ± 0.029**	**− 0.082 ± 0.026**	**− 0.136 ± 0.032**
BRI_{bayes}			−0.603±0.014	**− 0.017 ± 0.010**	**− 0.071 ± 0.011**
GP				−0.587±0.009	**− 0.054 ± 0.009**
GP_{prob}					−0.533±0.008

Diagonal elements are the estimated log-scores. Above-diagonal elements are the estimated difference between the log-scores of the corresponding row and column models. Standard errors of the estimates are provided and differences greater than 1 standard error are in bold.

https://doi.org/10.1371/journal.pone.0251945.t002

We note that in this particular illustration 2 nodes were selected for the **BRI** model and thus a linear solution, while for $\mathbf{BRI}_{bayes}$ and **GP** models 3 nodes were selected resulting in solutions with a closer fit.

The results on ATP data are in Table 3. As the selected parameters already suggested, the models find no meaningful variability in latent strengths and none of the models outperform the baseline model **BASE**, which assumes constant latent strengths. This can either be due to the top players indeed being consistent throughout the observed period or due to lack of information. Additional information could be incorporated, such as matches with players outside the top players and court-type, which plays an important role. However, this example illustrates that the more flexible models are robust to over-fitting the data and do not perform

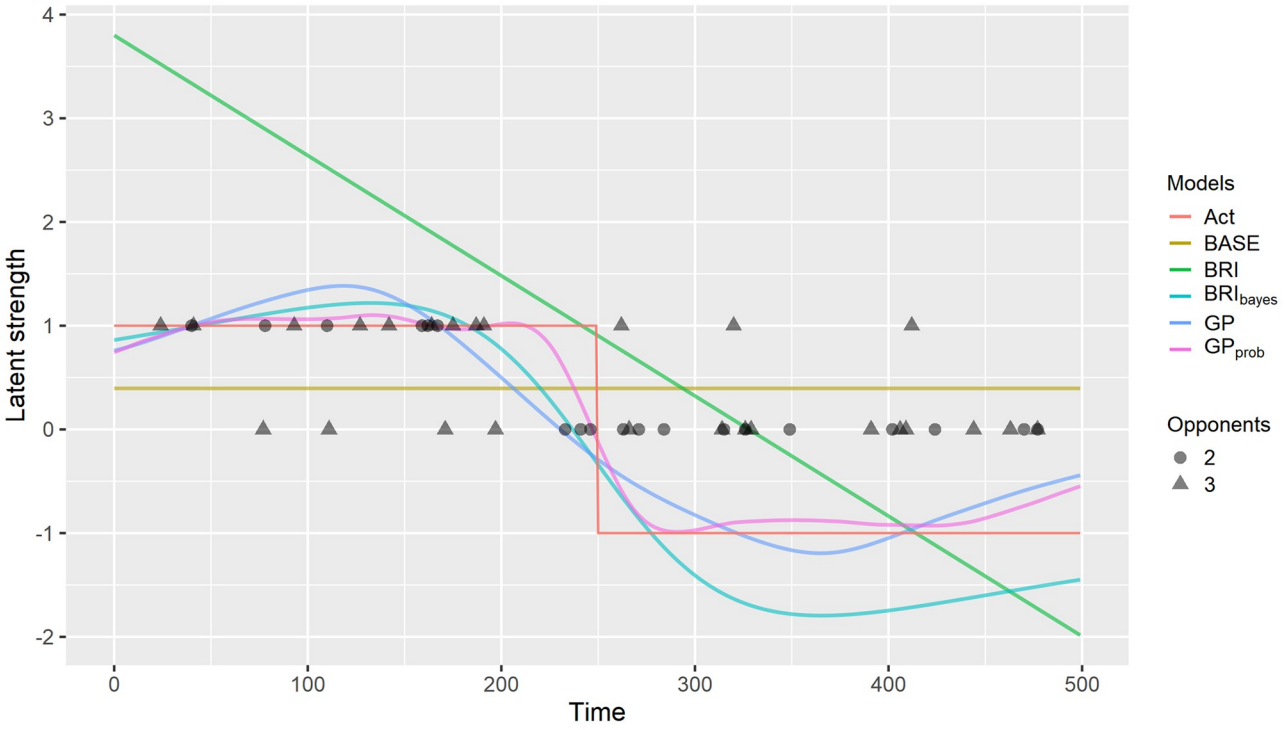

Fig 2. Model comparison of estimated latent strength for the 1st object in the toy data set. For models $\mathbf{BRI}_{bayes}$, **GP** and $\mathbf{GP}_{prob}$ we show the posterior mean. The red line represents the true latent strength. The points represent the training data. **GP** fits the true latent strength better than $\mathbf{BRI}_{bayes}$. $\mathbf{GP}_{prob}$, which uses additional probability data fits the true latent strength best.

https://doi.org/10.1371/journal.pone.0251945.g002

Table 3. Model performance on ATP data set.

	BASE	BRI	BRI_{bayes}	GP
BASE	−0.581±0.014	−0.000±0.000	−0.000±0.006	0.008±0.010
BRI		−0.581±0.014	−0.000±0.006	0.008±0.010
BRI_{bayes}			−0.580±0.013	**0.008 ± 0.006**
GP				−0.589±0.011

Diagonal elements are the estimated log-scores. Above-diagonal elements are the estimated difference between the log-scores of the corresponding row and column models. Standard errors of the estimates are provided and differences greater than 1 standard error are in bold.

https://doi.org/10.1371/journal.pone.0251945.t003

worse than a constant latent-strength model. We also note that in case of the **BRI** model only one node was chosen for all train-test splits giving the same result as the **BASE** model.

In Fig 3 we show latent strengths of top 5 tennis players obtained with the **BASE** model. These results show that from 2015 to 2019 Novak Djoković was the best player followed by Roger Federer, Andy Murray, Rafael Nadal, and Feliciano Lopez.

The results on NBA data are in Table 4. Unlike ATP data set, the selected tunable parameter values suggested that there is some variability in latent strengths to be modelled. Similar to toy data set the models are, in order of increasing performance, **BASE**, **BRI**, $\mathbf{BRI_{bayes}}$, **GP**, and

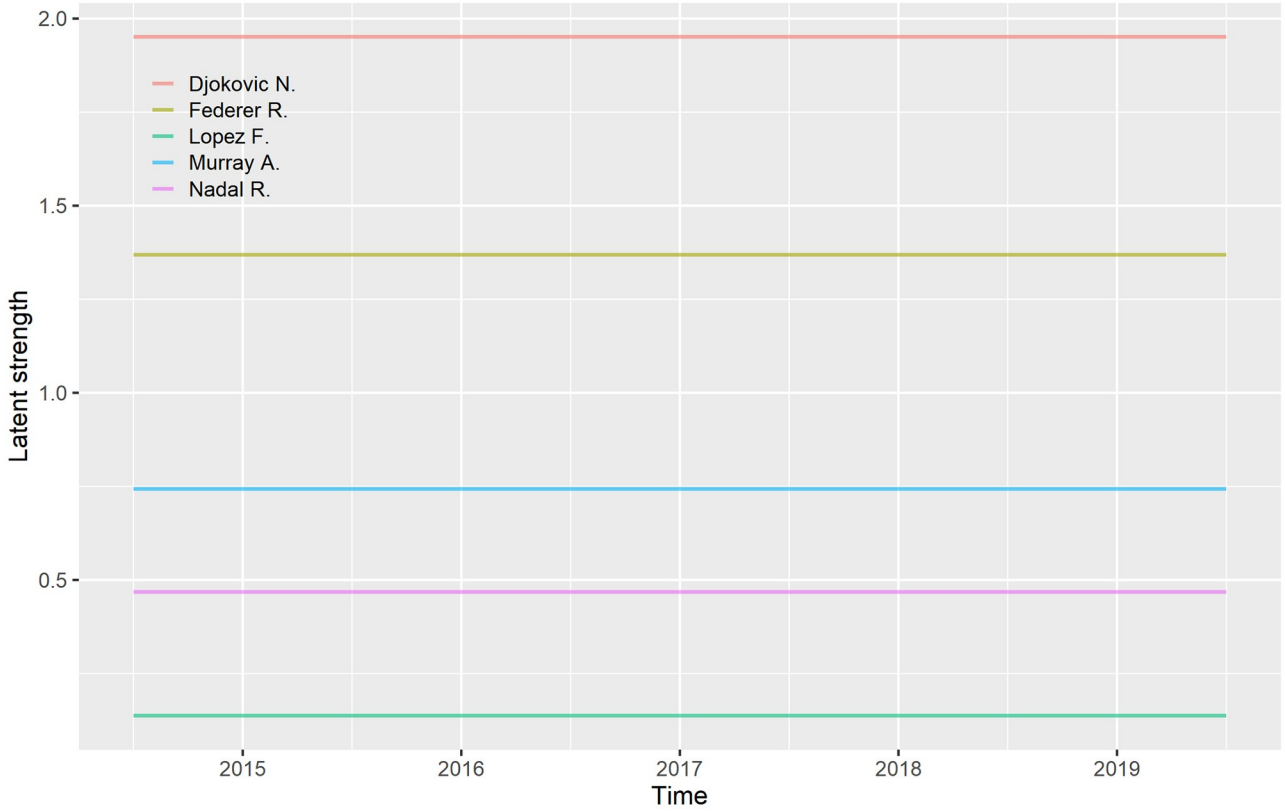

Fig 3. The five players with the highest latent strength according to the BASE model.

https://doi.org/10.1371/journal.pone.0251945.g003

Table 4. Model performance on NBA data set.

	BASE	BRI	BRI_{bayes}	GP	GP_{prob}	Odds
BASE	−0.627±0.002	**− 0.005 ± 0.002**	**− 0.009 ± 0.003**	**− 0.022 ± 0.002**	**− 0.047 ± 0.002**	**− 0.042 ± 0.002**
BRI		−0.622±0.003	**− 0.004 ± 0.002**	**− 0.017 ± 0.002**	**− 0.042 ± 0.003**	**− 0.037 ± 0.003**
BRI_{bayes}			−0.618±0.004	**− 0.013 ± 0.003**	**− 0.038 ± 0.003**	**− 0.033 ± 0.003**
GP				−0.605±0.002	**− 0.025 ± 0.002**	**− 0.019 ± 0.002**
GP_{prob}					−0.580±0.003	**0.005 ± 0.001**
Odds						−0.586±0.003

Diagonal elements are the estimated log-scores. Above-diagonal elements are the estimated difference between the log-scores of the corresponding row and column models. Standard errors of the estimates are provided and differences greater than 1 standard error are in bold.

https://doi.org/10.1371/journal.pone.0251945.t004

GP_{prob}. Again, the **GP_{prob}** model was expected to outperform the other models, because it uses more information and the **GP** model is better than the BRI-based models. As an additional benchmark we include a comparison with probabilities from bookmaker win odds (**Odds**). Our model when using these probabilities outperforms them. The other models give 3%—6% lower log-scores. The latent strengths of 5 selected NBA teams are shown in Fig 4.

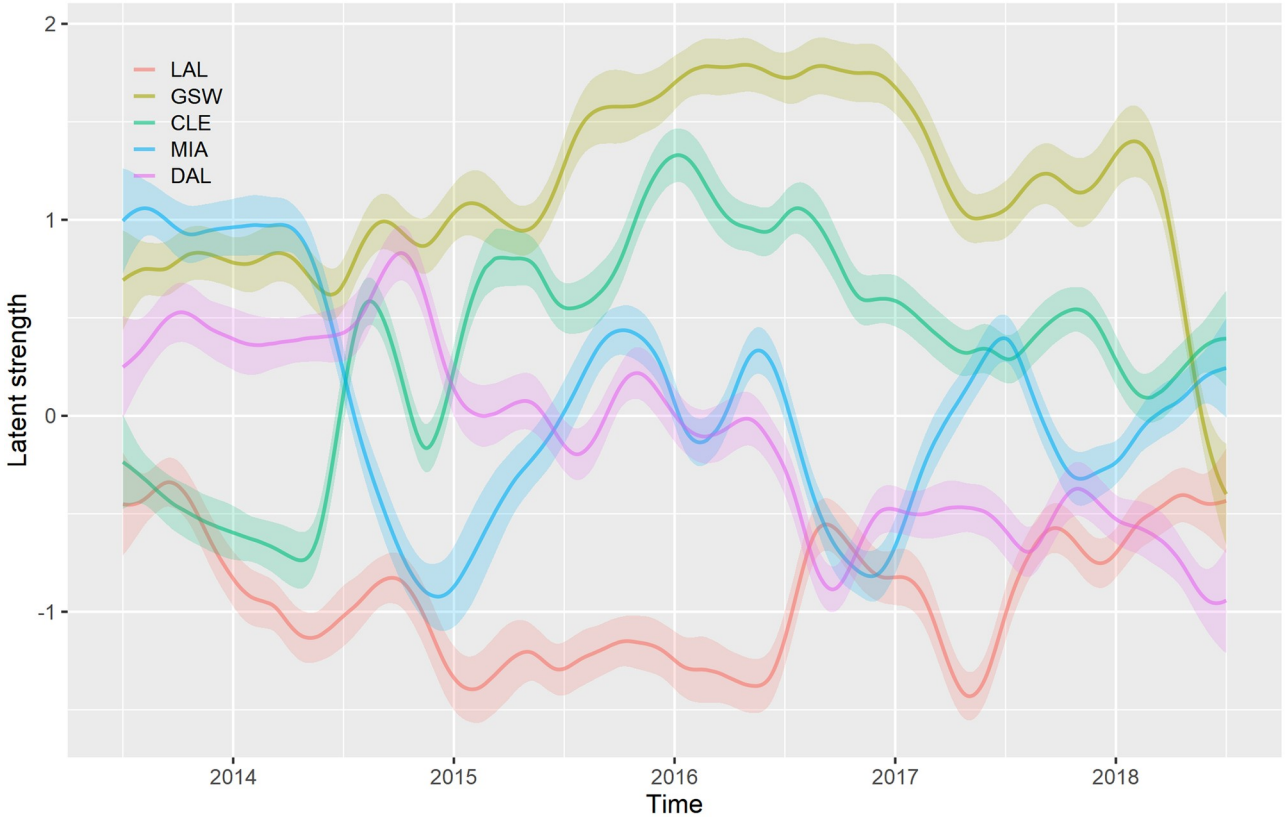

Fig 4. Comparison of latent strengths of selected five NBA teams using the GP_{prob} model. For each team a line and a ribbon are shown which represent a posterior mean and the corresponding standard deviation. The Golden States Warriors (GSW) were for most of the period the best out of these five teams. A drop can be seen in Miami Heat's (MIA) strength going from the 2014 to the 2015 season, while the Cleveland Cavaliers's (CLE) strength increases. These changes correspond with LeBron James leaving Miami Heat and returning to Cleveland Cavaliers.

https://doi.org/10.1371/journal.pone.0251945.g004

Conclusions

In this paper we extended the Bradley-Terry model using BRI and GPs to model latent strengths as the time-varying components of the model. In addition the model also allows for the inclusion of covariates and outcome probabilities. The use of outcome probabilities is overlooked in related work, although they are often available and substantially improve the model's performance as we demonstrated on toy and real data from NBA games. Even a biased estimate of the outcome probability provides more information than observing a single realization of the process.

We empirically demonstrated the advantages of GPs over BRI and the benefits of using a Bayesian approach to BRI instead of MLE. The BRI-based models are more sensitive to node selection than the GP-based models, the Bayesian BRI model less so than the MLE-based model. All the investigated models are robust to over-fitting and perform well even when the latent strengths are constant. As expected, BRI does not handle discontinuities as well as GPs. However, it is worth noting that this issue is not as pronounced when modelling latent strengths in a log-odds setting as it is when modelling observed data. Due to the exponential transformation, relatively sharp changes in observed performance can be modelled well by a smoother change in latent strength. This is an argument in favour of BRI as a useful alternative to splines and GPs when modelling latent strengths.

In our research we focused on hindcasting rather than forecasting. That is why we evaluated our models based on their performance on left-out games. If the goal was forecasting, we acknowledge that other approaches tailored to forecasting would give better results. Note, however, that our $\mathbf{GP}_{\text{prob}}$ model gives better results than log-scores calculated form bookmakers' odds. The down-side of our approach is the time complexity which comes with the MCMC methods and calculations of covariance matrix inverses. On the other hand our results are valuable to get a quantitative insight about the underlying strength dynamics of players or teams, which can be used for seeding competitions and matchmaking, scouting or visually engaging coaches and fans.

We could further improve our models in two ways. One direction is to use some other probability distribution function for modelling the comparison outcome which might be more suited to specific data. Another upgrade of the model would be to incorporate transitivity effect, which is often present in sports data.

Supporting information

S1 Appendix.
(PDF)

S1 Dataset. ATP data.
(CSV)

S2 Dataset. NBA games data.
(CSV)

S3 Dataset. NBA win odds data.
(CSV)

S4 Dataset. NBA teams data.
(CSV)

Acknowledgments

The authors would like to thank Gregor Pirš for technical support.

Author Contributions

Conceptualization: Blaž Krese, Erik Štrumbelj.

Data curation: Blaž Krese, Erik Štrumbelj.

Formal analysis: Blaž Krese.

Investigation: Blaž Krese, Erik Štrumbelj.

Methodology: Blaž Krese, Erik Štrumbelj.

Resources: Erik Štrumbelj.

Software: Blaž Krese.

Supervision: Erik Štrumbelj.

Validation: Blaž Krese, Erik Štrumbelj.

Visualization: Blaž Krese.

Writing – original draft: Blaž Krese, Erik Štrumbelj.

Writing – review & editing: Blaž Krese, Erik Štrumbelj.

References

1. David HA. The Method of Paired Comparisons. New York: Oxford University Press; 1988.
2. Cattelan M. Models for paired comparison data: A review with emphasis on dependent data. Statistical Science. 2012; 27(3):412–433. https://doi.org/10.1214/12-STS396
3. Bradley RA, Terry ME. The Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika. 1952; 39:324–345. https://doi.org/10.2307/2334029
4. Luce RD. Individual Choice Behavior: A Theoretical Analysis. New York, NY, USA: Wiley; 1959.
5. Thurstone LL. A Law of Comparative Judgement. Psychological Review. 1927; 34:278–286. https://doi.org/10.1037/h0070288
6. Rao PV, Kupper LL. Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. Journal of the American Statistical Association. 1967; 62(317):194–204. https://doi.org/10.1080/01621459.1967.10482901
7. Herbrich R, Minka T, Graepel T. TrueSkill(TM): A Bayesian Skill Rating System. In: Advances in Neural Information Processing Systems 20. MIT Press; 2007. p. 569–576.
8. Minka T, Cleven R, Zaykov Y. TrueSkill 2: An improved Bayesian skill rating system. Microsoft; 2018.
9. Makhijani R, Ugander J. Parametric Models for Intransitivity in Pairwise Rankings. In: The World Wide Web Conference; 2019. p. 3056–3062.
10. Ullah K, Aslam M, Sindhu TN. Bayesian analysis of the Weibull paired comparison model using informative prior. Alexandria Engineering Journal. 2020; 59(4):2371–2378. https://doi.org/10.1016/j.aej.2020.02.032
11. Elo AE. The rating of chessplayers, past and present. New York: Arco Pub.; 1978.
12. Aldous D. Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability? Statistical Science. 2017; 32(4):616–629. https://doi.org/10.1214/17-STS628
13. Glickman ME. A Comprehensive Guide to Chess Ratings. American Chess Journal. 1995; 3:59–102.
14. Chen C, Kok JN, Heiser W. Elo Rating System for UEFA Women's Euro 2017. The Predictive Power of Elo Ratings for the Performance of Teams and Players in the 2017 UEFA Women's Championship. Universiteit Leiden, The Netherlands; 2018.
15. Glickman ME. Parameter Estimation in Large Dynamic Paired Comparison Experiments. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1999; 48(3):377–394.

16. Glickman ME. Dynamic paired comparison models with stochastic variances. Journal of Applied Statistics. 2001; 28(6):673–689. https://doi.org/10.1080/02664760120059219

17. Coulom R. Whole-History Rating: A Bayesian Rating System for Players of Time-Varying Strength. In: Lecture Notes in Computer Science. vol. 5131; 2008. p. 113–124.

18. Baker RD, McHale IG. Time varying ratings in association football: the all-time greatest team is. . . Journal of the Royal Statistical Society: Series A (Statistics in Society). 2015; 178(2):481–492. https://doi.org/10.1111/rssa.12060

19. Baker RD, McHale IG. A dynamic paired comparisons model: Who is the greatest tennis player? European Journal of Operational Research. 2014; 236(2):677–684. https://doi.org/10.1016/j.ejor.2013.12.028

20. Owen A. Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. IMA Journal of Management Mathematics. 2011; 22(2):99–113. https://doi.org/10.1093/imaman/dpq018

21. Koopman SJ, Lit R. A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2015; 178(1):167–186. https://doi.org/10.1111/rssa.12042

22. Cattelan M, Varin C, Firth D. Dynamic Bradley–Terry modelling of sports tournaments. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2013; 62(1):135–150.

23. Floater M, Hormann K. Barycentric rational interpolation with no poles and high rates of approximation. Numerische Mathematik. 2007; 107:315–331. https://doi.org/10.1007/s00211-007-0093-y

24. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. MIT Press; 2006.

25. Shreve SE. Stochastic Calculus for Finance II: Continuous-Time Models. Springer; 2004.

26. Stan Development Team. Stan Modelling Language Users Guide and Reference Manual; 2019. Available from: https://mc-stan.org.

27. Baker RD, Jackson D. Statistical application of barycentric rational interpolants: an alternative to splines. Computational Statistics. 2014; 29:1065–1081. https://doi.org/10.1007/s00180-014-0480-7

28. Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014; 15(1):1593–1623.

29. Betancourt MJ. Generalizing the No-U-Turn Sampler to Riemannian Manifolds; 2013. Available from: https://arxiv.org/abs/1304.1920v1.

30. Kimeldorf GS, Wahba G. A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. Annals of Mathematical Statistics. 1970; 41(2):495–502. https://doi.org/10.1214/aoms/1177697089

31. Titsias M, Lawrence DN, Rattray M. Markov chain Monte Carlo algorithms for Gaussian processes. In: Inference and Estimation in Probabilistic Time-Series Models; 2008. p. 9.

32. Titsias M, Lawrence N, Rattray M. Efficient Sampling for Gaussian Process Inference using Control Variables. In: Advances in Neural Information Processing Systems. vol. 21; 2008. p. 1681–1688.

33. Štrumbelj E, Robnik Šikonja M. Online bookmakers' odds as forecasts: The case of European soccer leagues. International Journal of Forecasting. 2010; 26(3):482–488. https://doi.org/10.1016/j.ijforecast.2009.10.005