

Pairwise Comparisons with Flexible Time-Dynamics

Lucas Maystre*
Spotify
lucasm@spotify.com

Victor Kristof
EPFL
victor.kristof@epfl.ch

Matthias Grossglauser
EPFL
matthias.grossglauser@epfl.ch

ABSTRACT

Inspired by applications in sports where the skill of players or teams competing against each other varies over time, we propose a probabilistic model of pairwise-comparison outcomes that can capture a wide range of time dynamics. We achieve this by replacing the static parameters of a class of popular pairwise-comparison models by continuous-time Gaussian processes; the covariance function of these processes enables expressive dynamics. We develop an efficient inference algorithm that computes an approximate Bayesian posterior distribution. Despite the flexibility of our model, our inference algorithm requires only a few linear-time iterations over the data and can take advantage of modern multiprocessor computer architectures. We apply our model to several historical databases of sports outcomes and find that our approach *a)* outperforms competing approaches in terms of predictive performance, *b)* scales to millions of observations, and *c)* generates compelling visualizations that help in understanding and interpreting the data.

KEYWORDS

Pairwise comparisons; ranking; time series; Kalman filter; Bayesian inference; sports; games.

ACM Reference Format:

Lucas Maystre, Victor Kristof, and Matthias Grossglauser. 2019. Pairwise Comparisons with Flexible Time-Dynamics. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19)*, August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3292500.3330831>

1 INTRODUCTION

In many competitive sports and games (such as tennis, basketball, chess and electronic sports), the most useful definition of a competitor's skill is the propensity of that competitor to win against an opponent. It is often difficult to measure this skill *explicitly*: take basketball for example, a team's skill depends on the abilities of its players in terms of shooting accuracy, physical fitness, mental preparation, but also on the team's cohesion and coordination, on its strategy, on the enthusiasm of its fans, and a number of other intangible factors. However, it is easy to observe this skill *implicitly* through the outcomes of matches.

*This work was done while the author was at EPFL.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions.acm.org.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6201-6/19/08...\$15.00

<https://doi.org/10.1145/3292500.3330831>

In this setting, probabilistic models of pairwise-comparison outcomes provide an elegant and practical approach to quantifying skill and to predicting future match outcomes given past data. These models, pioneered by Zermelo [37] in the context of chess (and by Thurstone [35] in the context of psychophysics), have been studied for almost a century. They posit that each competitor i (i.e., a team or player) is characterized by a latent score $s_i \in \mathbf{R}$ and that the outcome probabilities of a match between i and j are a function of the difference $s_i - s_j$ between their scores. By estimating the scores $\{s_i\}$ from data, we obtain an interpretable proxy for skill that is predictive of future match outcomes. If a competitor's skill is expected to remain stable over time, these models are very effective. But what if it varies over time?

A number of methods have been proposed to adapt comparison models to the case where scores change over time. Perhaps the best known such method is the Elo rating system [9], used by the World Chess Federation for their official rankings. In this case, the time dynamics are captured essentially as a by-product of the learning rule (c.f. Section 5). Other approaches attempt to model these dynamics explicitly [e.g., 6, 7, 10, 13]. These methods greatly improve upon the static case when considering historical data, but they all assume the simplest model of time dynamics (that is, Brownian motion). Hence, they fail to capture more nuanced patterns such as variations at different timescales, linear trends, regression to the mean, discontinuities, and more.

In this work, we propose a new model of pairwise-comparison outcomes with expressive time-dynamics: it generalizes and extends previous approaches. We achieve this by treating the score of an opponent i as a time-varying Gaussian process $s_i(t)$ that can be endowed with flexible priors [26]. We also present an algorithm that, in spite of this increased flexibility, performs approximate Bayesian inference over the score processes in linear time in the number of observations so that our approach scales seamlessly to datasets with millions of observations. This inference algorithm addresses several shortcomings of previous methods: it can be parallelized effortlessly and accommodates different variational objectives. The highlights of our method are as follows.

Flexible Dynamics As scores are modeled by continuous-time Gaussian processes, complex (yet interpretable) dynamics can be expressed by composing covariance functions.

Generality The score of an opponent for a given match is expressed as a (sparse) linear combination of features. This enables, e.g., the representation of a home advantage or any other contextual effect. Furthermore, the model encompasses a variety of observation likelihoods beyond win / lose, based, e.g., on the number of points a competitor scores.

Bayesian Inference Our inference algorithm returns a posterior distribution over score processes. This leads to better predictive performance and enables a principled way to learn

the dynamics (and any other model hyperparameters) by optimizing the log-marginal likelihood of the data.

Ease of Interpretation By plotting the score processes $\{s_i(t)\}$ over time, it is easy to visualize the probability of any comparison outcome under the model. As the time dynamics are described through the composition of simple covariance functions, their interpretation is straightforward as well.

Concretely, our contributions are threefold. First, we develop a probabilistic model of pairwise-comparison outcomes with flexible time-dynamics (Section 2). The model covers a wide range of use cases, as it enables *a)* opponents to be represented by a sparse linear combination of features, and *b)* observations to follow various likelihood functions. In fact, it unifies and extends a large body of prior work. Second, we derive an efficient algorithm for approximate Bayesian inference (Section 3). This algorithm adapts to two different variational objectives; in conjunction with the “reverse-KL” objective, it provably converges to the optimal posterior approximation. It can be parallelized easily, and the most computationally intensive step can be offloaded to optimized off-the-shelf numerical software. Third, we apply our method on several sports datasets and show that it achieves state-of-the-art predictive performance (Section 4). Our results highlight that different sports are best modeled with different time-dynamics. We also demonstrate how domain-specific and contextual information can improve performance even further; in particular, we show that our model outperforms competing ones even when there are strong intransitivities in the data.

In addition to prediction tasks, our model can also be used to generate compelling visualizations of the temporal evolution of skills. All in all, we believe that our method will be useful to data-mining practitioners interested in understanding comparison time-series and in building predictive systems for games and sports.

A Note on Extensions. In this paper, we focus on *pairwise* comparisons for conciseness. However, the model and inference algorithm could be extended to multiway comparisons or partial rankings over small sets of opponents without any major conceptual change, similarly to Herbrich et al. [16]. Furthermore, and even though we develop our model in the context of sports, it is relevant to all applications of ranking from comparisons, e.g., to those where comparison outcomes reflect human preferences or opinions [21, 29, 35].

2 MODEL

In this section, we formally introduce our probabilistic model. For clarity, we take a clean-slate approach and develop the model from scratch. We discuss in more detail how it relates to prior work in Section 5.

The basic building blocks of our model are *features*¹. Let M be the number of features; each feature $m \in [M]$ is characterized by a latent, continuous-time Gaussian process

$$s_m(t) \sim \text{GP}[0, k_m(t, t')]. \quad (1)$$

We call $s_m(t)$ the *score process* of m , or simply its *score*. The *covariance function* of the process, $k_m(t, t') \doteq \mathbb{E}[s_m(t)s_m(t')]$, is used to encode time dynamics. A brief introduction to Gaussian processes

¹In the simplest case, there is a one-to-one mapping between competitors (e.g., teams) and features, but decoupling them offers increased modeling power.

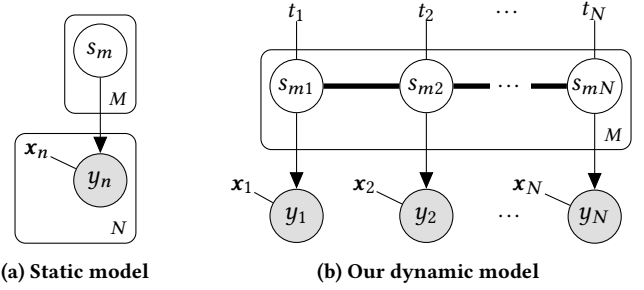


Figure 1: Graphical representation of a static model (left) and of the dynamic model presented in this paper (right). The observed variables are shaded. For conciseness, we let $x_n \doteq x_{n,i} - x_{n,j}$. Right: the latent score variables are mutually dependent across time, as indicated by the thick line.

as well as a discussion of useful covariance functions is given in Section 2.1. The M scores $s_1(t), \dots, s_M(t)$ are assumed to be (a priori) jointly independent, and we collect them into the *score vector*

$$s(t) = [s_1(t) \ \dots \ s_M(t)]^\top.$$

For a given match, each opponent i is described by a sparse linear combination of the features, with coefficients $\mathbf{x}_i \in \mathbb{R}^M$. That is, the score of an opponent i at time t^* is given by

$$s_i = \mathbf{x}_i^\top s(t^*). \quad (2)$$

In the case of a one-to-one mapping between competitors and features, \mathbf{x}_i is simply the one-hot encoding of opponent i . More complex setups are possible: For example, in the case of team sports and if the player lineup is available for each match, it could also be used to encode the players taking part in the match [20]. Note that \mathbf{x}_i can also depend contextually on the match. For instance, it can be used to encode the fact that a team plays at home [1].

Each observation consists of a tuple $(\mathbf{x}_i, \mathbf{x}_j, t^*, y)$, where $\mathbf{x}_i, \mathbf{x}_j$ are the opponents’ feature vectors, $t^* \in \mathbb{R}$ is the time, and $y \in \mathcal{Y}$ is the match outcome. We posit that this outcome is a random variable that depends on the opponents through their latent score difference:

$$y \mid \mathbf{x}_i, \mathbf{x}_j, t^* \sim p(y \mid s_i - s_j),$$

where p is a known probability density (or mass) function and s_i, s_j are given by (2). The idea of modeling outcome probabilities through score differences dates back to Thurstone [35] and Zermelo [37]. The likelihood p is chosen such that positive values of $s_i - s_j$ lead to successful outcomes for opponent i and vice-versa.

A graphical representation of the model is provided in Figure 1. For perspective, we also include the representation of a static model, such as that of Thurstone [35]. Our model can be interpreted as “conditionally parametric”: conditioned on a particular time, it falls back to a (static) pairwise-comparison model parametrized by real-valued scores.

Observation Models. Choosing an appropriate likelihood function $p(y \mid s_i - s_j)$ is an important modeling decision and depends on the information contained in the outcome y . The most widely applicable likelihoods require only *ordinal* observations, i.e., whether a match resulted in a win or a loss (or a tie, if applicable). In some cases, we might additionally observe points (e.g., in association football, the

Table 1: Examples of observation likelihoods. The score difference is denoted by $d \doteq s_i - s_j$ and the Gaussian cumulative density function is denoted by Φ .

Name	\mathcal{Y}	$p(y d)$	References
Probit	$\{\pm 1\}$	$\Phi(yd)$	[16, 35]
Logit	$\{\pm 1\}$	$[1 + \exp(-yd)]^{-1}$	[4, 37]
Ordinal probit	$\{\pm 1, 0\}$	$\Phi(yd - \alpha), \dots$	[11]
Poisson-exp	$\mathbb{N}_{\geq 0}$	$\exp(yd - e^d)/y!$	[19]
Gaussian	\mathbf{R}	$\propto \exp[(y - d)^2/(2\sigma^2)]$	[14]

number of goals scored by each team). To make use of this extra information, we can model *a*) the number of points of opponent i with a Poisson distribution whose rate is a function of $s_i - s_j$, or *b*) the points difference with a Gaussian distribution centered at $s_i - s_j$. A non-exhaustive list of likelihoods is given in Table 1.

2.1 Covariance Functions

A Gaussian process $s(t) \sim \text{GP}[0, k(t, t')]$ can be thought of as an infinite collection of random variables indexed by time, such that the joint distribution of any finite vector of N samples $\mathbf{s} = [s(t_1) \dots s(t_N)]$ is given by $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K} = [k(t_i, t_j)]$. That is, \mathbf{s} is jointly Gaussian with mean $\mathbf{0}$ and covariance matrix \mathbf{K} . We refer the reader to Rasmussen and Williams [26] for an excellent introduction to Gaussian processes.

Hence, by specifying the covariance function appropriately, we can express prior expectations about the time dynamics of a feature's score, such as smooth or non-smooth variations at different timescales, regression to the mean, discontinuities, linear trends and more. Here, we describe a few functions that we find useful in the context of modeling temporal variations. Figure 2 illustrates these functions through random realizations of the corresponding Gaussian processes.

Constant This covariance captures processes that remain constant over time. It is useful in composite covariances to model a constant offset (i.e., a mean score value).

Piecewise Constant Given a partition of \mathbf{R} into disjoint intervals, this covariance is constant inside a partition and zero between partitions. It can, for instance, capture discontinuities across seasons in professional sports leagues.

Wiener This covariance reflects Brownian motion dynamics (c.f. Section 5). It is non-stationary: the corresponding process drifts away from 0 as t grows.

Matérn This family of stationary covariance functions can represent smooth and non-smooth variations at various timescales. It is parametrized by a variance, a characteristic timescale and a smoothness parameter ν . When $\nu = 1/2$, it corresponds to a mean-reverting version of Brownian motion.

Linear This covariance captures linear dynamics.

Finally, note that composite functions can be created by adding or multiplying covariance functions together. For example, let k_a and k_b be constant and Matérn covariance functions, respectively. Then, the composite covariance $k(t, t') \doteq k_a(t, t') + k_b(t, t')$ captures dynamics that fluctuate around a (non-zero) mean value. Duvenaud

[8, Sec. 2.3] provides a good introduction to building expressive covariance functions by composing simple ones.

3 INFERENCE ALGORITHM

In this section, we derive an efficient inference algorithm for our model. For brevity, we focus on explaining the main ideas behind the algorithm. A reference software implementation, available online at <https://github.com/lucasmaystre/kickscore>, complements the description provided here.

We begin by introducing some notation. Let $\mathcal{D} = \{(\mathbf{x}_n, t_n, y_n) : n \in [N]\}$ be a dataset of N independent observations, where for conciseness we fold the two opponents $\mathbf{x}_{n,i}$ and $\mathbf{x}_{n,j}$ into $\mathbf{x}_n \doteq \mathbf{x}_{n,i} - \mathbf{x}_{n,j}$, for each observation². Let $\mathcal{D}_m \subseteq [N]$ be the subset of observations involving feature m , i.e., those observations for which $x_{nm} \neq 0$, and let $N_m = |\mathcal{D}_m|$. Finally, denote by $\mathbf{s}_m \in \mathbf{R}^{N_m}$ the samples of the latent score process at times corresponding to the observations in \mathcal{D}_m . The joint prior distribution of these samples is $p(\mathbf{s}_m) = \mathcal{N}(\mathbf{0}, \mathbf{K}_m)$, where \mathbf{K}_m is formed by evaluating the covariance function $k_m(t, t')$ at the relevant times.

We take a Bayesian approach and seek to compute the posterior distribution

$$p(\mathbf{s}_1, \dots, \mathbf{s}_M | \mathcal{D}) \propto \prod_{m=1}^M p(\mathbf{s}_m) \prod_{n=1}^N p[y_n | \mathbf{x}_n^\top \mathbf{s}(t_n)]. \quad (3)$$

As the scores are coupled through the observations, the posterior no longer factorizes over $\{\mathbf{s}_m\}$. Furthermore, computing the posterior is intractable if the likelihood is non-Gaussian.

To overcome these challenges, we consider a mean-field variational approximation [36]. In particular, we assume that the posterior can be well-approximated by a multivariate Gaussian distribution that factorizes over the features:

$$p(\mathbf{s}_1, \dots, \mathbf{s}_M | \mathcal{D}) \approx q(\mathbf{s}_1, \dots, \mathbf{s}_M) \doteq \prod_{m=1}^M \mathcal{N}(\mathbf{s}_m | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m). \quad (4)$$

Computing this approximate posterior amounts to finding the variational parameters $\{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ that best approximate the true posterior. More formally, the inference problem reduces to the optimization problem

$$\min_{\{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}} \text{div}[p(\mathbf{s}_1, \dots, \mathbf{s}_M | \mathcal{D}) \| q(\mathbf{s}_1, \dots, \mathbf{s}_M)], \quad (5)$$

for some divergence measure $\text{div}(p \| q) \geq 0$. We will consider two different such measures in Section 3.1.

A different viewpoint on the approximate posterior is as follows. For both of the variational objectives that we consider, it is possible to rewrite the optimal distribution $q(\mathbf{s}_m)$ as

$$q(\mathbf{s}_m) \propto p(\mathbf{s}_m) \prod_{n \in \mathcal{D}_m} \mathcal{N}[s_{mn} | \tilde{\mu}_{mn}, \tilde{\sigma}_{mn}^2].$$

Letting $\mathcal{X}_n \subseteq [M]$ be the subset of features such that $x_{nm} \neq 0$, we can now reinterpret the variational approximation as transforming every observation (\mathbf{x}_n, t_n, y_n) into several independent *pseudo-observations* with Gaussian likelihood, one for each feature $m \in \mathcal{X}_n$. Instead of optimizing directly $\{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ in (5), we can alternatively choose to optimize the parameters $\{\tilde{\mu}_{mn}, \tilde{\sigma}_{mn}^2\}$. For

²This enables us to write the score difference more compactly. Given an observation at time t^* and letting $\mathbf{x} \doteq \mathbf{x}_i - \mathbf{x}_j$, we have $s_i - s_j = \mathbf{x}_i^\top \mathbf{s}(t^*) - \mathbf{x}_j^\top \mathbf{s}(t^*) = \mathbf{x}^\top \mathbf{s}(t^*)$.

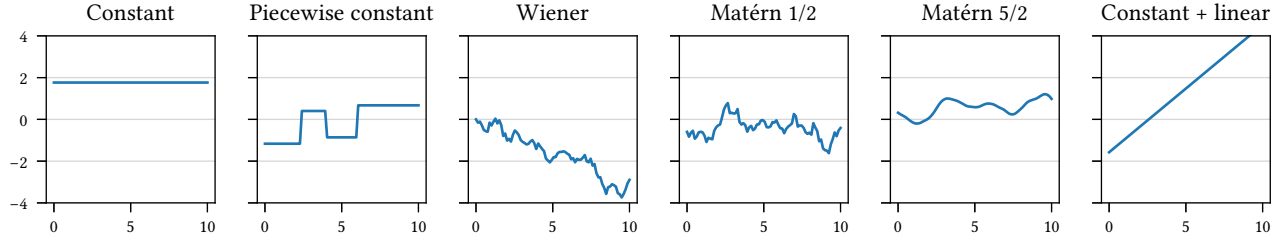


Figure 2: Random realizations of a zero-mean Gaussian process with six different covariance functions.

Algorithm 1 Model inference.

Require: $\mathcal{D} = \{(\mathbf{x}_n, t_n, y_n) : n \in [N]\}$

```

1:  $\tilde{\mu}_m, \tilde{\sigma}_m^2 \leftarrow 0, \infty \quad \forall m$ 
2:  $q(s_m) \leftarrow p(s_m) \quad \forall m$ 
3: repeat
4:   for  $n = 1, \dots, N$  do
5:      $\delta \leftarrow \text{DERIVATIVES}(\mathbf{x}_n, y_n)$ 
6:     for  $m \in \mathcal{X}_n$  do
7:        $\tilde{\mu}_{mn}, \tilde{\sigma}_{mn}^2 \leftarrow \text{UPDATEPARAMS}(\mathbf{x}_{nm}, \delta)$ 
8:   for  $m = 1, \dots, M$  do
9:      $q(s_m) \leftarrow \text{UPDATEPOSTERIOR}(\tilde{\mu}_m, \tilde{\sigma}_m^2)$ 
10: until convergence

```

any feature m , given the pseudo-observations' parameters $\tilde{\mu}_m$ and $\tilde{\sigma}_m^2$, computing $q(s_m)$ becomes tractable (c.f. Section 3.2).

An outline of our iterative inference procedure is given in Algorithm 1. Every iteration consists of two steps:

- (1) updating the pseudo-observations' parameters given the true observations and the current approximate posterior (lines 4–7), and
- (2) recomputing the approximate posterior given the current pseudo-observation (lines 8 and 9).

Convergence is declared when the difference between two successive iterates of $\{\tilde{\mu}_{mn}\}$ and $\{\tilde{\sigma}_{mn}^2\}$ falls below a threshold. Note that, as a by-product of the computations performed by the algorithm, we can also estimate the log-marginal likelihood of the data, $\log p(\mathcal{D})$.

Running Time. In Appendix A, we show that DERIVATIVES and UPDATEPARAMS run in constant time. In Section 3.2, we show that UPDATEPOSTERIOR runs in time $O(N_m)$. Therefore, if we assume that the vectors $\{\mathbf{x}_n\}$ are sparse, the total running time per iteration of Algorithm 1 is $O(N)$. Furthermore, each of the two outer *for* loops (lines 4 and 8) can be parallelized easily, leading in most cases to a linear acceleration with the number of available processors.

3.1 Updating the Pseudo-Observations

The exact computations performed during the first step of the inference algorithm—updating the pseudo-observations—depend on the specific variational method used. We consider two: expectation propagation [22], and reverse-KL variational inference [3]. The ability of Algorithm 1 to seamlessly adapt to either of the two methods

is valuable, as it enables practitioners to use the most advantageous method for a given likelihood function. Detailed formulae for DERIVATIVES and UPDATEPARAMS can be found in Appendix A.

3.1.1 Expectation Propagation. We begin by defining two distributions. The *cavity* distribution q_{-n} is the approximate posterior without the pseudo-observations associated with the n th datum, that is,

$$q_{-n}(s_1, \dots, s_M) \propto \frac{q(s_1, \dots, s_M)}{\prod_{m \in \mathcal{X}_n} \mathcal{N}[s_{mn} | \tilde{\mu}_{mn}, \tilde{\sigma}_{mn}^2]}.$$

The *hybrid* distribution \hat{q}_n is given by the cavity distribution multiplied by the n th likelihood factor, i.e.,

$$\hat{q}_n(s_1, \dots, s_M) \propto q_{-n}(s_1, \dots, s_M) p[y_n | \mathbf{x}_n^\top \mathbf{s}(t_n)].$$

Informally, the hybrid distribution \hat{q}_n is “closer” to the true distribution than q .

Expectation propagation (EP) works as follows. At each iteration and for each n , we update the parameters $\{\tilde{\mu}_{mn}, \tilde{\sigma}_{mn}^2 : m \in \mathcal{X}_n\}$ such that $\text{KL}(\hat{q}_n || q)$ is minimized. To this end, the function DERIVATIVES (on line 5 of Algorithm 1) computes the first and second derivatives of the log-partition function

$$\log \mathbb{E}_{q_{-n}} \{p[y_n | \mathbf{x}_n^\top \mathbf{s}(t_n)]\} \quad (6)$$

with respect to $\mu_{-n} \doteq \mathbb{E}_{q_{-n}}[\mathbf{x}_n^\top \mathbf{s}(t_n)]$. These computations can be done in closed form for the widely-used probit likelihood, and they involve one-dimensional numerical integration for most other likelihoods. EP has been reported to result in more accurate posterior approximations on certain classification tasks [23].

3.1.2 Reverse KL Divergence. This method (often referred to simply as *variational inference* in the literature) seeks to minimize $\text{KL}(q || p)$, i.e., the KL divergence from the approximate posterior q to the true posterior p .

To optimize this objective, we adopt the approach of Khan and Lin [18]. In this case, the function DERIVATIVES computes the first and second derivatives of the expected log-likelihood

$$\mathbb{E}_q \{ \log p[y_n | \mathbf{x}_n^\top \mathbf{s}(t_n)] \} \quad (7)$$

with respect to $\mu \doteq \mathbb{E}_q[\mathbf{x}_n^\top \mathbf{s}(t_n)]$. These computations involve numerically solving two one-dimensional integrals.

In comparison to EP, this method has two advantages. The first is theoretical: If the likelihood $p(y | d)$ is log-concave in d , then the variational objective has a unique global minimum, and we can guarantee that Algorithm 1 converges to this minimum [18]. The second is numerical: Excepted for the probit likelihood, computing (7) is numerically more stable than computing (6).

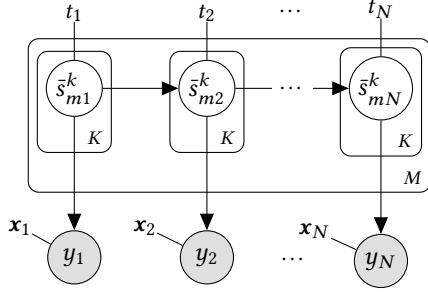


Figure 3: State-space reformulation of our model. With respect to the representation in Figure 1b, the number of latent variables has increased, but they now form a Markov chain.

3.2 Updating the Approximate Posterior

The second step of Algorithm 1 (lines 8 and 9) solves the following problem, for every feature m . Given Gaussian pseudo-observations $\{\tilde{\mu}_{mn}, \tilde{\sigma}_{mn} : n \in \mathcal{D}_m\}$ and a Gaussian prior $p(s_m) = \mathcal{N}(\mathbf{0}, K_m)$, compute the posterior

$$q(s_m) \propto p(s_m) \prod_{n \in \mathcal{D}_m} \mathcal{N}[s_{mn} | \tilde{\mu}_{mn}, \tilde{\sigma}_{mn}^2].$$

This computation can be done independently and in parallel for each feature $m \in [M]$.

A naive approach is to use the self-conjugacy properties of the Gaussian distribution directly. Collecting the parameters of the pseudo-observations into a vector $\tilde{\mu}_m$ and a diagonal matrix $\tilde{\Sigma}_m$, the parameters of the posterior $q(s_m)$ are given by

$$\Sigma_m = (K_m^{-1} + \tilde{\Sigma}_m^{-1})^{-1}, \quad \mu_m = \Sigma_m \tilde{\Sigma}_m^{-1} \tilde{\mu}_m. \quad (8)$$

Unfortunately, this computation runs in time $O(N_m^3)$, a cost that becomes prohibitive if some features appear in many observations.

Instead, we use an alternative approach that exploits a link between temporal Gaussian processes and state-space models [15, 27]. Without loss of generality, we now assume that the N observations are ordered chronologically, and, for conciseness, we drop the feature's index and consider a single process $s(t)$. The key idea is to augment $s(t)$ into a K -dimensional vector-valued Gauss-Markov process $\tilde{s}(t)$, such that

$$\tilde{s}(t_{n+1}) = A_n \tilde{s}(t_n) + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(\mathbf{0}, Q_n)$$

where $K \in \mathbb{N}_{>0}$ and $A_n, Q_n \in \mathbb{R}^{K \times K}$ depend on the time interval $|t_{n+1} - t_n|$ and on the covariance function $k(t, t')$ of the original process $s(t)$. The original (scalar-valued) and the augmented (vector-valued) processes are related through the equation

$$s(t) = \mathbf{h}^\top \tilde{s}(t),$$

where $\mathbf{h} \in \mathbb{R}^K$ is called the *measurement vector*.

Figure 3 illustrates our model from a state-space viewpoint. It is important to note that the mutual time dependencies of Figure 1b have been replaced by Markovian dependencies. In this state-space formulation, posterior inference can be done in time $O(K^3 N)$ by using the Rauch–Tung–Striebel smoother [34].

From Covariance Functions to State-Space Models. A method for converting a process $s(t) \sim \text{GP}[\mathbf{0}, k(t, t')]$ into an equivalent Gauss-Markov process $\tilde{s}(t)$ by explicit construction of \mathbf{h} , $\{A_n\}$ and $\{Q_n\}$ is given in Solin [31]. All the covariance functions described in Section 2.1 lead to exact state-space reformulations of order $K \leq 3$. The composition of covariance functions through addition or multiplication can also be treated exactly and automatically. Some other covariance functions, such as the squared-exponential function or periodic functions [26], cannot be transformed exactly but can be approximated effectively and to arbitrary accuracy [15, 32].

Finally, we stress that the state-space viewpoint is useful because it leads to a faster inference procedure; but defining the time dynamics of the score processes in terms of covariance functions is much more intuitive.

3.3 Predicting at a New Time

Given the approximate posterior $q(s_1, \dots, s_M)$, the probability of observing outcome y at a new time t^* given the feature vector \mathbf{x} is given by

$$p(y | \mathbf{x}, t^*) = \int_{\mathbb{R}} p(y | z) p(z) dz,$$

where $z = \mathbf{x}^\top \mathbf{s}(t^*)$ and the distribution of $s_m(t^*)$ is derived from the posterior $q(s_m)$. By using the state-space formulation of the model, the prediction can be done in constant time [28].

4 EXPERIMENTAL EVALUATION

In this section, we evaluate our model and inference algorithm on real data. Our experiments cover three aspects. First, in Section 4.1, we compare the predictive performance of our model against competing approaches, focusing on the impact of flexible time-dynamics. Second, in Section 4.2, we show that by carefully choosing features and observation likelihoods, predictive performance can be improved significantly. Finally, in Section 4.3, we study various facets of our inference algorithm. We measure the impact of the mean-field assumption and of the choice of variational objective, and we demonstrate the scalability of the algorithm.

Datasets. We consider six datasets of pairwise-comparison outcomes of various sports and games. Four of them contain time-stamped outcomes; they relate to tennis, basketball, association football and chess. Due to the large size of the chess dataset³, we also consider a subset of the data spanning 30 years. The two remaining datasets contain match outcomes of the StarCraft computer game and do not have timestamps. Table 2 provides summary statistics for all the datasets. Except for chess, all data are publicly available online⁴.

Performance Metrics. Let (\mathbf{x}, t^*, y) be an observation. We measure performance by using the logarithmic loss: $-\log p(y | \mathbf{x}, t^*)$ and the accuracy: $\mathbf{1}\{y = \arg \max_{y'} p(y' | \mathbf{x}, t^*)\}$. We report their average values on the test set.

³This dataset consists of all the match outcomes contained in *ChessBase Big Database 2018*, available at https://shop.chessbase.com/en/products/big_database_2018.

⁴Tennis: https://github.com/JeffSackmann/tennis_atp, basketball: https://projects.fivethirtyeight.com/nba-model/nba_elo.csv, football: <https://int.soccerway.com/>, StarCraft: https://github.com/csinpi/blade_chest.

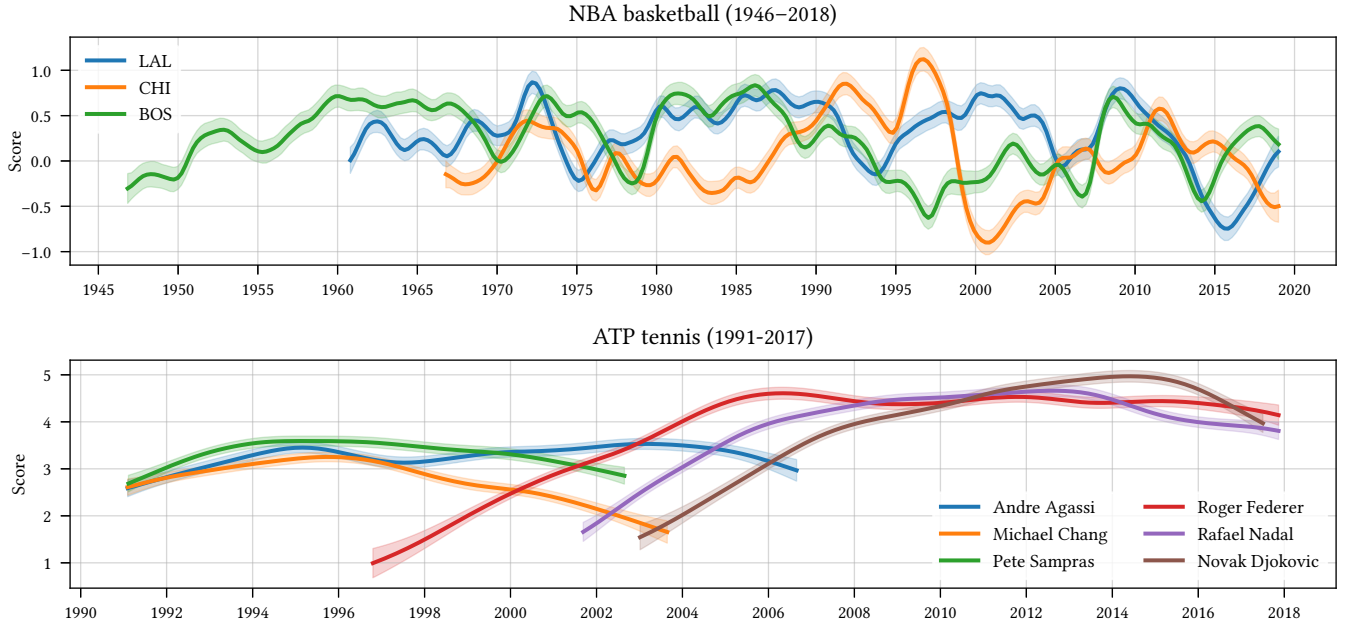


Figure 4: Temporal evolution of the score processes ($\mu \pm \sigma$) corresponding to selected basketball teams (top) and tennis players (bottom). The basketball teams are the Los Angeles Lakers (LAL), the Chicago Bulls (CHI) and the Boston Celtics (BOS).

Table 2: Summary statistics of the sports datasets.

Name	Ties	M	N	Time span
ATP tennis	No	20 046	618 934	1991–2017
NBA basketball	No	102	67 642	1946–2018
World football	Yes	235	19 158	1908–2018
ChessBase small	Yes	19 788	306 764	1950–1980
ChessBase full	Yes	343 668	7 169 202	1475–2017
StarCraft WoL	No	4381	61 657	—
StarCraft HotS	No	2287	28 582	—

Methodology. Unless specified otherwise, we partition every dataset into a training set containing the first 70% of the observations and a test set containing the remaining 30%, in chronological order. The various hyperparameters (such as covariance functions and their parameters, learning rates, etc.) are selected based on the training data only, by maximizing the log-marginal likelihood of Bayesian models and by minimizing the average leave-one-out log loss otherwise. The final hyperparameter configuration of all models can be found in Appendix B. In order to predict the outcome of an observation at time t^* , we use *all* the data (in both training and test sets) up to the day preceding t^* . This closely mimics the setting where a predictor must guess the outcome of an event in the near future based on all past data. Unless specified otherwise, we use Algorithm 1 with the EP variational objective, and we declare convergence when the improvement in log-marginal likelihood falls below 10^{-3} . Typically, the algorithm converges in less than a hundred iterations.

4.1 Flexible Time-Dynamics

In this experiment, we compare the predictive performance of our model against competing approaches on four timestamped datasets. In order to better isolate and understand the impact of accurately modeling *time dynamics* on predictive performance, we keep the remaining modeling choices simple: we treat all outcomes as ordinal-valued (i.e., *win*, *loss* and possibly *tie*) with a probit likelihood and use a one-to-one mapping between competitors and features. In Table 3, we report results for the following models:

- *Random*. This baseline assigns equal probability to every outcome.
- *Constant*. The model of Section 2 with a constant covariance function. This model assumes that the scores do not vary over time.
- *Elo*. The system used by the World Chess Federation [9]. Time dynamics are a by-product of the update rule (c.f. Section 5).
- *TrueSkill*. The Bayesian model of Herbrich et al. [16]. Time dynamics are assumed to follow Brownian motion (akin to our Wiener kernel) and inference is done in a single pass over the data.
- *Ours*. The model of Section 2. We try multiple covariance functions and report the one that maximizes the log-marginal likelihood.

Our model matches or outperforms other approaches in almost all cases, both in terms of log loss and in terms of accuracy. Interestingly, different datasets are best modeled by using different covariance functions, perhaps capturing underlying skill dynamics specific to each sport.

Table 3: Predictive performance of our model and of competing approaches on four datasets, in terms of average log loss and average accuracy. The best result is indicated in bold.

Dataset	Random		Constant		Elo		TrueSkill		Ours		
	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Covariance
ATP tennis	0.693	0.500	0.581	0.689	0.563	0.705	0.563	0.705	0.552	0.714	Affine + Wiener
NBA basketball	0.693	0.500	0.692	0.536	0.634	0.644	0.634	0.644	0.630	0.645	Constant + Matérn 1/2
World football	1.099	0.333	0.929	0.558	0.950	0.551	0.937	0.554	0.926	0.558	Constant + Matérn 1/2
ChessBase small	1.099	0.333	1.030	0.478	1.035	0.447	1.030	0.467	1.026	0.474	Constant + Wiener

Visualizing and Interpreting Scores. Figure 4 displays the temporal evolution of the score of selected basketball teams and tennis players. In the basketball case, we can recognize the dominance of the Boston Celtics in the early 1960’s and the Chicago Bulls’ strong 1995–96 season. In the tennis case, we can see the progression of a new generation of tennis champions at the turn of the 21st century. Plotting scores over time provides an effective way to compactly represent the history of a given sport. Analyzing the optimal hyperparameters (c.f. Table 8 in Appendix B) is also insightful: the characteristic timescale of the dynamic covariance component is 1.75 and 7.47 years for basketball and tennis, respectively. The score of basketball teams appears to be much more volatile.

4.2 Generality of the Model

In this section, we demonstrate how we can take advantage of additional modeling options to further improve predictive performance. In particular, we show that choosing an appropriate likelihood and parametrizing opponents with match-dependent combinations of features can bring substantial gains.

4.2.1 Observation Models. Basketball and football match outcomes actually consist of *points* (respectively, goals) scored by each team during the match. We can make use of this additional information to improve predictions [19]. For each of the basketball and football datasets, we compare the best model obtained in Section 4.1 to alternative models. These alternative models keep the same time dynamics but use either

- (1) a logit likelihood on the ordinal outcome,
- (2) a Gaussian likelihood on the points difference, or
- (3) a Poisson-exp likelihood on the points scored by each team.

The results are presented in Table 4. The logit likelihood performs similarly to the probit one [33], but likelihoods that take points into account can indeed lead to better predictions.

Table 4: Average predictive log loss of models with different observation likelihoods. The best result is indicated in bold.

Dataset	Probit	Logit	Gaussian	Poisson
NBA basketball	0.630	0.630	0.627	0.630
World football	0.926	0.926	0.927	0.922

4.2.2 Match-Dependent Parametrization. For a given match, we can represent opponents by using (non-trivial) linear combinations

of features. This enables, e.g., to represent context-specific information that might influence the outcome probabilities. In the case of football, for example, it is well-known that a team playing at home has an advantage. Similarly, in the case of chess, playing White results in a slight advantage. Table 5 displays the predictive performance achieved by our model when the score of the home team (respectively, that of the opponent playing White) is modeled by a linear combination of two features: the identity of the team or player and an *advantage* feature. Including this additional feature improves performance significantly, and we conclude that representing opponents in terms of match-dependent combinations of features can be very useful in practice.

Table 5: Predictive performance of models with a home or first-mover advantage in comparison to models without.

Dataset	Basic		Advantage	
	Loss	Acc.	Loss	Acc.
World football	0.926	0.558	0.900	0.579
ChessBase small	1.026	0.480	1.019	0.485

4.2.3 Capturing Intransitivity. Score-based models such as ours are sometimes believed to be unable to capture meaningful intransitivities, such as those that arise in the “rock-paper-scissors” game [5]. This is incorrect: if an opponent’s score can be modeled by using match-dependent features, we can simply add an *interaction* feature for every pair of opponents. In the next experiment, we model the score difference between two opponents i, j as $d \doteq s_i - s_j + s_{ij}$. Informally, the model learns to explain the transitive effects through the usual player scores s_i and s_j and the remaining intransitive effects are captured by the interaction score s_{ij} . We compare this model to the Blade-Chest model of Chen and Joachims [5] on the two StarCraft datasets, known to contain strong intransitivities. The Blade-Chest model is specifically designed to handle intransitivities in comparison data. We also include two baselines, a simple Bradley–Terry model without the interaction features (*logit*) and a non-parametric estimator (*naive*) that estimates probabilities based on match outcomes between each pair—without attempting to capture transitive effects. As shown in Figure 5, our model outperforms all other approaches, including the Blade-Chest model. More details on this experiment can be found in Appendix B.2.

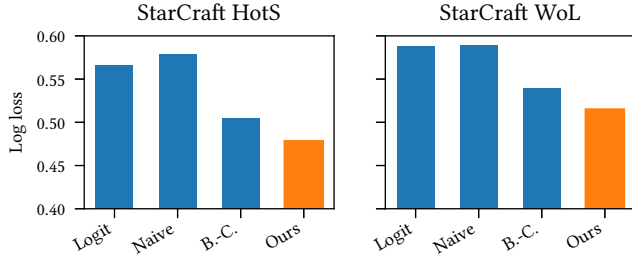


Figure 5: Average log loss of four models (Bradley-Terry, naive, blade-chest and ours) on the StarCraft datasets.

4.3 Inference Algorithm

We turn our attention to the inference algorithm and study the impact of several implementation choices. We start by quantifying the impact of the mean-field assumption (4) and of the choice of variational objective on predictive performance. Then, we demonstrate the scalability of the algorithm on the ChessBase dataset and measure the acceleration obtained by parallelizing the algorithm.

4.3.1 Mean-Field Approximation. In order to gain understanding on the impact of the factorization assumption in (4), we devise the following experiment. We consider a small subset of the basketball data containing all matches between 2000 and 2005 ($N = 6382$, $M = 32$). We evaluate the predictive performance on each week of the last season by using all the matches prior to the test week as training data. Our model uses a one-to-one mapping between teams and features, a constant + Matérn 1/2 covariance function, and a Gaussian likelihood on the points difference.

We compare the predictive performance resulting from two inference variants, *a*) mean-field approximate inference, i.e., Algorithm 1, and *b*) exact posterior inference⁵. Both approaches lead to an average log loss of 0.634 and an average accuracy of 0.664. Strikingly, both values are equal up to four decimal places, suggesting that the mean-field assumption is benign in practice [2].

4.3.2 Variational Objective. Next, we study the influence of the variational method. We re-run the experiments of Section 4.1, this time by using the reverse-KL objective instead of EP. The predictive performance in terms of average log loss and average accuracy is equal to the EP case (Table 3, last three columns) up to three decimal places, for all four datasets. Hence, the variational objective seems to have little practical impact on predictive performance. As such, we recommend using the reverse-KL objective for likelihoods whose log-partition function (6) cannot be computed in closed form, as the numerical integration of the expected log-likelihood (7) is generally more stable.

4.3.3 Scalability. Finally, we demonstrate the scalability of our inference algorithm by training a model on the full ChessBase dataset, containing over 7 million observations. We implement a multithreaded version of Algorithm 1 in the Go programming language⁶ and run the inference computation on a machine containing two 12-core Intel Xeon E5-2680 v3 (Haswell generation) processors

⁵This is possible for this particular choice of likelihood thanks to the self-conjugacy of the Gaussian distribution, but at a computational cost $O(N^3)$.

⁶The code is available at <https://github.com/lucasmaystre/gokick>.

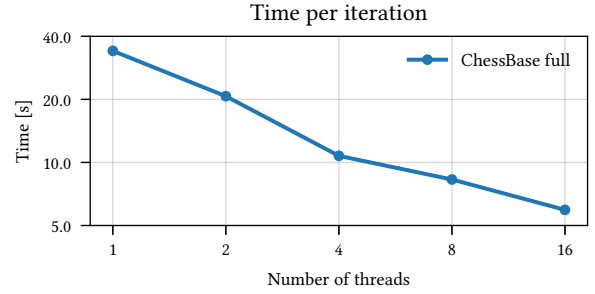


Figure 6: Running time per iteration of a multithreaded implementation of Algorithm 1 on the ChessBase full dataset, containing over 7 million observations.

clocked at 2.5 GHz. Figure 6 displays the running time per iteration as function of the number of worker threads. By using 16 threads, we need only slightly over 5 seconds per iteration.

5 RELATED WORK

Probabilistic models for pairwise comparisons have been studied for almost a century. Thurstone [35] proposed his seminal *law of comparative judgment* in the context of psychology. Almost concurrently, Zermelo [37] developed a method to rank chess players from match outcomes. Both rely on the same idea: objects are characterized by a latent score (e.g., the intrinsic quality of a perceptual variable, or a chess player’s skill) and the outcomes of comparisons between objects depend on the difference between the corresponding latent scores. Zermelo’s model was later rediscovered by Bradley and Terry [4] and is currently usually referred to as the Bradley-Terry model. Stern [33] provides a unifying framework and shows that, in practice, Thurstone’s and Zermelo’s models result in similar fits to the data. In the context of sports, some authors suggest going beyond ordinal outcomes and investigate pairwise-comparison models with Gaussian [14], Poisson [14, 19], or Skellam [17] likelihoods.

In many applications of practical interest, comparison outcomes tend to vary over time. In chess, for example, this is due to the skill of players changing over time. The World Chess Federation, which uses a variant of the Bradley-Terry model to rank players, updates player scores after each match by using a stochastic gradient update:

$$s_i \leftarrow s_i + \lambda \frac{\partial}{\partial s_i} \log p(y | s_i - s_j),$$

where $\lambda \in \mathbb{R}$ is a learning rate. It is interesting that this simple online update scheme (known as the Elo rating system [9]) enables a basic form of “tracking”: the sequence of scores gives an indication of a player’s evolution over time. Whereas, in this case, score dynamics occur as a by-product of the learning rule, several attempts have been made to model time dynamics explicitly. Usually, these models assume a variant of Brownian motion:

$$s(t_{n+1}) = s(t_n) + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2 | t_{n+1} - t_n). \quad (9)$$

Glickman [12] and Fahrmeir and Tutz [10] are, to the best of our knowledge, the first to consider such a model. Glickman [13] derives a computationally-efficient Bayesian inference method by using closed-form approximations of intractable integrals. Herbrich et al.

[16] and Dangauthier et al. [7] propose a similar method based on Gaussian filtering and expectation propagation, respectively. Coulom [6] proposes a method based on the Laplace approximation. Our model strictly subsumes these approaches; Brownian motion is simply a special case of our model obtained by using the Wiener kernel. One of the key contributions of our work is to show that it is not necessary to restrict the dynamics to Brownian motion in order to get linear-time inference.

Finally, we briefly review literature on the link between Gaussian processes (GPs) with scalar inputs and state-space models (SSMs), as this forms a crucial component of our fast inference procedure. Excellent introductions to this link can be found in the theses of Saatçi [28] and Solin [31]. The connection is known since the seminal paper of O’Hagan [25], which introduced Gaussian processes as a method to tackle general regression problems. It was recently revisited by Hartikainen and Särkkä [15], who provide formulae for going back-and-forth between GP covariance and state-space forms. Extensions of this link to non-Gaussian likelihood models are discussed in Saatçi [28] and Nickisch et al. [24]. To the best of our knowledge, we are the first to describe how the link between GPs and SSMs can be used in the context of observation models that combine *multiple* processes, by using a mean-field variational approximation.

6 CONCLUSIONS

We have presented a probabilistic model of pairwise comparison outcomes that can capture a wide range of temporal dynamics. This model reaches state-of-the-art predictive performance on several sports datasets, and it enables generating visualizations that help in understanding comparison time-series. To fit our model, we have derived a computationally efficient approximate Bayesian inference algorithm. To the best of our knowledge, our algorithm is the first linear-time Bayesian inference algorithm for dynamic pairwise comparison models that minimizes the reverse-KL divergence.

One of the strengths of our approach is that it enables to discover the structure of the time dynamics by comparing the log-marginal likelihood of the data under various choices of covariance functions. In the future, we would like to fully automatize this discovery process, in the spirit of the *automatic statistician* [8]. Ideally, given only the comparison data, we should be able to systematically discover the time dynamics that best explain the data, and generate an interpretable description of the corresponding covariance functions.

ACKNOWLEDGMENTS

We thank Holly Cogliati-Bauereis, Ksenia Konyushkova, and the anonymous reviewers for careful proofreading and constructive feedback.

REFERENCES

- [1] A. Agresti. 2012. *Categorical Data Analysis* (third ed.). Wiley.
- [2] A. Birlutiu and T. Heskes. 2007. Expectation Propagation for Rating Players in Sports Competitions. In *Proceedings of PKDD 2007*. Warsaw, Poland.
- [3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. 2017. Variational Inference: A Review for Statisticians. *J. Amer. Statist. Assoc.* 112, 518 (2017), 859–877.
- [4] R. A. Bradley and M. E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345.
- [5] S. Chen and T. Joachims. 2016. Modeling Intransitivity in Matchup and Comparison Data. San Francisco, CA, USA.
- [6] R. Coulom. 2008. Whole-History Rating: A Bayesian Rating System for Players of Time-Varying Strength. In *Proceedings of CG 2008*. Beijing, China.
- [7] P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel. 2007. TrueSkill Through Time: Revisiting the History of Chess. In *Advances in Neural Information Processing Systems 20*. Vancouver, BC, Canada.
- [8] D. Duvenaud. 2014. *Automatic Model Construction with Gaussian Processes*. Ph.D. Dissertation. University of Cambridge.
- [9] A. Elo. 1978. *The Rating Of Chess Players, Past & Present*. Arco Publishing.
- [10] L. Fahrmeir and G. Tutz. 1994. Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison Systems. *J. Amer. Statist. Assoc.* 89, 428 (1994), 1438–1449.
- [11] W. A. Glenn and H. A. David. 1960. Ties in Paired-Comparison Experiments Using a Modified Thurstone–Mosteller Model. *Biometrics* 16, 1 (1960), 86–109.
- [12] M. E. Glickman. 1993. *Paired Comparison Models with Time-Varying Parameters*. Ph.D. Dissertation. Harvard University.
- [13] M. E. Glickman. 1999. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 48, 3 (1999), 377–394.
- [14] S. Guo, S. Sanner, T. Graepel, and W. Buntine. 2012. Score-Based Bayesian Skill Learning. In *Proceedings of ECML PKDD 2012*. Bristol, United Kingdom.
- [15] J. Hartikainen and S. Särkkä. 2010. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of MLSP 2010*. Kittilä, Finland.
- [16] R. Herbrich, T. Minka, and T. Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 19*. Vancouver, BC, Canada.
- [17] D. Karlis and I. Ntzoufras. 2009. Bayesian modelling of football outcomes: using the Skellam’s distribution for the goal difference. *IMA Journal of Management Mathematics* 20, 2 (2009), 133–145.
- [18] M. E. Khan and W. Lin. 2017. Conjugate-Computation Variational Inference: Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. In *Proceedings of AISTATS 2017*. Fort Lauderdale, FL, USA.
- [19] M. J. Maher. 1982. Modelling association football scores. *Statistica Neerlandica* 36, 3 (1982), 109–118.
- [20] L. Maystre, V. Kristof, A. J. González Ferrer, and M. Grossglauser. 2016. The Player Kernel: Learning Team Strengths Based on Implicit Player Contributions. (Sept. 2016). Preprint, arXiv:1609.01176 [cs.LG].
- [21] D. McFadden. 1973. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, P. Zarembka (Ed.). Academic Press, 105–142.
- [22] T. P. Minka. 2001. *A family of algorithms for approximate Bayesian inference*. Ph.D. Dissertation. Massachusetts Institute of Technology.
- [23] H. Nickisch and C. E. Rasmussen. 2008. Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research* 9, Oct (2008), 2035–2078.
- [24] H. Nickisch, A. Solin, and A. Grigorievskiy. 2018. State Space Gaussian Processes with Non-Gaussian Likelihood. In *Proceedings of ICMML 2018*. Long Beach, CA, USA.
- [25] A. O’Hagan. 1978. Curve Fitting and Optimal Design for Prediction. *Journal of the Royal Statistical Society, Series B* 40, 1 (1978), 1–42.
- [26] C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.
- [27] S. Reece and S. Roberts. 2010. An introduction to Gaussian processes for the Kalman filter expert. In *Proceedings of ICIF 2010*. Edinburgh, UK.
- [28] Y. Saatçi. 2012. *Scalable Inference for Structured Gaussian Process Models*. Ph.D. Dissertation. University of Cambridge.
- [29] M. J. Salganik and K. E. C. Levy. 2015. Wiki Surveys: Open and Quantifiable Social Data Collection. *PLoS ONE* 10, 5 (2015), 1–17.
- [30] M. Seeger, S. Gerwin, and M. Bethge. 2007. Bayesian Inference for Sparse Generalized Linear Models. In *Proceedings of ECML 2007*. Warsaw, Poland.
- [31] A. Solin. 2016. *Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression*. Ph.D. Dissertation. Aalto University.
- [32] A. Solin and S. Särkkä. 2014. Explicit Link Between Periodic Covariance Functions and State Space Models. In *Proceedings of AISTATS 2014*. Reykjavik, Iceland.
- [33] H. Stern. 1992. Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences* 23, 1 (1992), 103–117.
- [34] S. Särkkä. 2013. *Bayesian Filtering and Smoothing*. Cambridge University Press.
- [35] L. L. Thurstone. 1927. A Law of Comparative Judgment. *Psychological Review* 34, 4 (1927), 273–286.
- [36] M. J. Wainwright and M. I. Jordan. 2012. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* 1, 1–2 (2012), 1–305.
- [37] E. Zermelo. 1928. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 29, 1 (1928), 436–460.

A INFERENCE ALGORITHM

For conciseness, we drop the index n and consider a single observation $(\mathbf{x}, t^*, y) \in \mathcal{D}$. Let $\mathbf{s} \doteq \mathbf{s}(t^*)$ be the vector containing the score of all features at the time of the observation. Instead of optimizing the “standard” parameters $\tilde{\mu}_m, \tilde{\sigma}_m^2$, we will optimize the corresponding *natural* parameters $\tilde{\alpha}_m, \tilde{\beta}_m$. They are related through the following equations.

$$\tilde{\alpha}_m = \tilde{\mu}_m / \tilde{\sigma}_m^2, \quad \tilde{\beta}_m = 1 / \tilde{\sigma}_m^2$$

Expectation Propagation. Let $q_-(\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ be the cavity distribution. The log-partition function can be rewritten as a one-dimensional integral:

$$\begin{aligned} \log Z &\doteq \log \mathbb{E}_{q_-} [p(y | \mathbf{x}^\top \mathbf{s})] = \log \int_{\mathbf{s}} p(y | \mathbf{x}^\top \mathbf{s}) \mathcal{N}(\mathbf{s} | \boldsymbol{\mu}, \Sigma) d\mathbf{s} \\ &= \log \int_u p(y | u) \mathcal{N}(u | \mu, \sigma^2) du, \end{aligned}$$

where $\mu = \mathbf{x}^\top \boldsymbol{\mu}$ and $\sigma^2 = \mathbf{x}^\top \Sigma \mathbf{x}$. The function `DERIVATIVES` computes the first and second derivatives with respect to this mean, i.e.,

$$\delta_1 = \frac{\partial}{\partial \mu} \log Z, \quad \delta_2 = \frac{\partial^2}{\partial \mu^2} \log Z.$$

Given these quantities, the function `UPDATEPARAMS` updates the pseudo-observations’ parameters for each $m \in \mathcal{X}$:

$$\begin{aligned} \tilde{\alpha}_m &\leftarrow (1 - \lambda) \tilde{\alpha}_m + \lambda \left[\frac{x_m \delta_1 - \mu_m x_m^2 \delta_2}{1 + \sum_{mm} x_m^2 \delta_2} \right], \\ \tilde{\beta}_m &\leftarrow (1 - \lambda) \tilde{\beta}_m + \lambda \left[\frac{-x_m^2 \delta_2}{1 + \sum_{mm} x_m^2 \delta_2} \right], \end{aligned}$$

where $\lambda \in (0, 1]$ is a learning rate. A formal derivation of these update equations can be found in Minka [22], Rasmussen and Williams [26], Seeger et al. [30].

Reverse KL Divergence. Let $q(\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ be the current posterior. Similarly to the log-partition function, the expected log-likelihood can be rewritten as a one-dimensional integral:

$$\begin{aligned} L &\doteq \mathbb{E}_q [\log p(y | \mathbf{x}^\top \mathbf{s})] = \int_{\mathbf{s}} \log p(y | \mathbf{x}^\top \mathbf{s}) \mathcal{N}(\mathbf{s} | \boldsymbol{\mu}, \Sigma) d\mathbf{s} \\ &= \int_u \log p(y | u) \mathcal{N}(u | \mu, \sigma^2) du, \end{aligned}$$

where $\mu = \mathbf{x}^\top \boldsymbol{\mu}$ and $\sigma^2 = \mathbf{x}^\top \Sigma \mathbf{x}$. The function `DERIVATIVES` computes the first and second derivatives with respect to this mean, i.e.,

$$\delta_1 = \frac{\partial}{\partial \mu} L, \quad \delta_2 = \frac{\partial^2}{\partial \mu^2} L.$$

Given these quantities, the function `UPDATEPARAMS` updates the pseudo-observations’ parameters for each $m \in \mathcal{X}$:

$$\begin{aligned} \tilde{\alpha}_m &\leftarrow (1 - \lambda) \tilde{\alpha}_m + \lambda [x_m \delta_1 - \mu_m x_m^2 \delta_2], \\ \tilde{\beta}_m &\leftarrow (1 - \lambda) \tilde{\beta}_m + \lambda [-x_m^2 \delta_2], \end{aligned}$$

where $\lambda \in (0, 1]$ is the learning rate. A formal derivation of these update equations can be found in Khan and Lin [18].

B EXPERIMENTAL EVALUATION

The code used to produce the experiments presented in this paper is publicly available online. It consists of two software libraries.

- A library written in the Python programming language, available at <https://github.com/lucasmaystre/kickscore>. This library provides a reference implementation of Algorithm 1 with a user-friendly API.
- A library written in the Go programming language, available at <https://github.com/lucasmaystre/gokick>. This library provides a multithreaded implementation of Algorithm 1, focused on computational performance.

Additionally, the scripts and computational notebooks used to produce the experiments and figures presented in this paper are available at <https://github.com/lucasmaystre/kickscore-kdd19>.

B.1 Hyperparameters

Generally speaking, we choose hyperparameters based on a search over 1000 configurations sampled randomly in a range of sensible values (we always make sure that the best hyperparameters are not too close to the ranges’ boundaries). In the case of our models, we choose the configuration that maximizes the log-marginal likelihood of the training data. In the case of TrueSkill and Elo, we choose the configuration that minimizes the leave-one-out log loss on the training data.

A list of all the hyperparameters is provided in Table 6. A formal definition of the covariance functions we use is given in Table 7. Finally, Table 8 lists the hyperparameter values used in most of the experiments described in Section 4.

Table 6: Hyperparameters and their description.

Symbol	Description
λ	Learning rate
α	Draw margin
σ_n^2	Observation noise (Gaussian likelihood)
σ_{cst}^2	Variance (constant covariance)
σ_{lin}^2	Variance (linear covariance)
σ_{W}^2	Variance (Wiener covariance)
ν	Smoothness (Matérn covariance)
σ_{dyn}^2	Variance (Matérn covariance)
ℓ	Timescale, in years (Matérn covariance)

Table 7: Covariance functions.

Name	$k(t, t')$
Constant	σ_{cst}^2
Linear	$\sigma_{\text{lin}}^2 t t'$
Wiener	$\sigma_{\text{W}}^2 \min\{t, t'\}$
Matérn, $\nu = 1/2$	$\sigma_{\text{dyn}}^2 \exp(- t - t' /\ell)$
Matérn, $\nu = 3/2$	$\sigma_{\text{dyn}}^2 (1 + \sqrt{3} t - t' /\ell) \exp(-\sqrt{3} t - t' /\ell)$

Table 8: Hyperparameter values used for the models of Section 4.

Dataset	Model	Likelihood	λ	α	σ_n^2	σ_{cst}^2	σ_{lin}^2	σ_W^2	ν	σ_{dyn}^2	ℓ
ATP tennis	Constant	Probit	1.000	—	—	0.817	—	—	—	—	—
	Elo	Logit	0.262	—	—	—	—	—	—	—	—
	TrueSkill	Probit	—	—	—	0.137	—	0.007	—	—	—
	Ours	Probit	1.000	—	—	0.366	0.001	0.147	—	—	—
	Figure 4	Probit	1.000	—	—	0.034	—	—	3/2	0.912	7.469
NBA Basketball	Constant	Probit	1.000	—	—	0.060	—	—	—	—	—
	Elo	Logit	0.095	—	—	—	—	—	—	—	—
	TrueSkill	Probit	—	—	—	0.128	—	0.001	—	—	—
	Ours	Probit	1.000	—	—	0.003	—	—	1/2	0.152	3.324
	Figure 4	Probit	1.000	—	—	0.003	—	—	3/2	0.138	1.753
	Table 4	Logit	1.000	—	—	0.001	—	—	1/2	0.417	3.429
	Table 4	Gaussian	1.000	—	143.451	0.059	—	—	1/2	17.667	3.310
	Table 4	Poisson-exp	0.800	—	—	5.470	—	—	1/2	0.003	2.378
World Football	Constant	Probit	1.000	0.372	—	0.933	—	—	—	—	—
	Elo	Logit	0.196	0.578	—	—	—	—	—	—	—
	TrueSkill	Probit	—	0.381	—	1.420	—	0.001	—	—	—
	Ours	Probit	1.000	0.386	—	0.750	—	—	1/2	0.248	69.985
	Table 4	Logit	1.000	0.646	—	2.001	—	—	1/2	0.761	71.693
	Table 4	Gaussian	1.000	—	3.003	4.062	—	—	1/2	2.922	175.025
	Table 4	Poisson-exp	0.800	—	—	0.300	—	—	1/2	0.210	83.610
	Table 5	Probit	1.000	0.407	—	0.895	—	—	1/2	0.220	44.472
ChessBase small	Constant	Probit	1.000	0.554	—	0.364	—	—	—	—	—
	Elo	Logit	0.157	0.856	—	—	—	—	—	—	—
	TrueSkill	Probit	—	0.555	—	0.240	—	0.001	—	—	—
	Ours	Probit	1.000	0.558	—	0.307	—	0.010	—	—	—
	Table 5	Probit	1.000	0.568	—	0.188	—	—	1/2	0.188	35.132

Table 9: Hyperparameter values for the experiment of Section 4.2.3.

Dataset	Bradley–Terry	Ours	
	α	σ_{cst}^2	σ_{\times}^2
StarCraft WoL	0.077	4.821	3.734
StarCraft HotS	0.129	4.996	4.342

B.2 Capturing Intransitivity

We closely follow the experimental procedure of Chen and Joachims [5] for the experiment of Section 4.2.3. In particular, we randomly partition each dataset into three splits: a training set (50% of the data), a validation set (20%), and a test set (30%). We train the model on the training set, choose hyperparameters based on the log loss

measured on the validation set, and finally report the average log loss on the test set.

For the Blade-Chest model, we were not able to reproduce the exact results presented in Chen and Joachims [5] using the open-source implementation available at https://github.com/csinpi/blade_chest. Instead, we just report the best values in Figures 3 and 4 of the paper (*blade-chest inner* model, with $d = 50$).

For the Bradley–Terry model, we use the Python library *choix* and its function `opt_pairwise`. The only hyperparameter to set is the regularization strength α .

For our model, since there are no timestamps in the StarCraft datasets, we simply use constant covariance functions. The two hyperparameters are σ_{cst}^2 and σ_{\times}^2 , the variance of the player features and of the interaction features, respectively. The hyperparameter values that we used are given in Table 9.