Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **00**, 1–21
https://doi.org/10.1093/jrsssc/qlae035

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series C Applied Statistics

Original Article

A state-space perspective on modelling and inference for online skill rating

Samuel Duffield[1], Samuel Power[2] and Lorenzo Rimella[3,4]

[1]Normal Computing, London, UK
[2]School of Mathematics, Fry Building, Woodland Road, University of Bristol, Bristol BS8 1UG, UK
[3]Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YR, UK
[4]ESOMAS, University of Torino and Collegio Carlo Alberto, Torino 10124, Italy

Address for correspondence: Samuel Power, School of Mathematics, Fry Building, Woodland Road, University of Bristol, Bristol BS8 1UG, UK. Email: sam.power@bristol.ac.uk

Abstract

We summarize popular methods used for skill rating in competitive sports, along with their inferential paradigms and introduce new approaches based on sequential Monte Carlo and discrete hidden Markov models. We advocate for a state-space model perspective, wherein players' skills are represented as time-varying, and match results serve as observed quantities. We explore the steps to construct the model and the three stages of inference: filtering, smoothing, and parameter estimation. We examine the challenges of scaling up to numerous players and matches, highlighting the main approximations and reductions which facilitate statistical and computational efficiency. We additionally compare approaches in a realistic experimental pipeline that can be easily reproduced and extended with our open-source Python package, abile.

Keywords: approximate inference, Bayesian inference, competitive sports, state-space models

1 Introduction

In the quantitative analysis of competitive sports, a fundamental task is to estimate the skills of the different agents ('players') involved in a given competition based on the outcome of comparisons ('matches') between said players, often in an online setting. Skill rating is of paramount importance as it serves as a foundational tool for assessing and comparing the abilities of players and how they vary over time. By accurately quantifying skill levels, rating systems inform strategic decision-making and enhance the overall sporting level. Common applications include chess (FIDE, 2023), online gaming (Herbrich *et al.*, 2006), education (Pelánek, 2016), tennis (Kovalchik, 2016), and team-based sports like football (Hvattum & Arntzen, 2010), basketball (Štrumbelj & Vračar, 2012), and many more (Ötting *et al.*, 2020; Stefani, 2011). Skill ratings not only facilitate player ranking but also serve as a basis for dynamic matchmaking and player seeding, ensuring that competitive matches are engaging and well-matched. Moreover, in professional sports, skill rating plays a pivotal role in talent scouting and gauging overall performance, providing data-driven insights for coaches, analysts, bettors, and fanatics. As technology and statistical methodologies continue to advance, skill rating systems are expected to evolve further, benefiting an ever-widening spectrum of sports and competitive domains.

There are several established approaches to the task of skill estimation, including among others the Bradley–Terry model (Bradley & Terry, 1952), the Elo rating system (Elo, 1978), the Glicko rating system (Glickman, 1999), and TrueSkill (Herbrich *et al.*, 2006) each with various levels of complexity and statistical foundation. In these works, statistical models are often left implicit or not even present, hindering the incorporation of additional structure into estimation routines.

Received: September 19, 2023. **Accepted:** July 6, 2024

Therefore, we propose a state-space modelling approach that separates model definition from inference and decomposes both aspects into their fundamental elements. This modularization empowers practitioners to easily experiment with different model structures (e.g. binary vs. non-binary match outcomes) and inferential paradigms (e.g. sequential Monte Carlo vs. extended Kalman filter), ultimately selecting the most suitable combination for their specific needs.

Our key contributions can be summarized as follows:

- We provide a unified state-space model framework for the skill rating, which encompasses and extends existing approaches by fostering the decoupling and modularization of model design and general purpose inference. The state-space model represents a customizable blueprint, adaptable to specific needs (e.g. non-Gaussian dynamics, Bivariate-Poisson likelihoods, and more). We detail the relevant inference objectives (filtering, smoothing, and parameter estimation) outlining their purpose, applications, and efficient implementation strategies.
- We rigorously formalize the *factorial approximation* as the necessary bias introduced in order to scale computations up to realistic settings with numerous players while retaining statistical fidelity.
- We draw from the state-space model literature a suite of inferential methods. Including sequential Monte Carlo and finite-space methods that are completely novel in the setting of skill ratings and suffer from fewer sources of bias than previous model-based approaches, see Table 1.
- We walk through a realistic workflow of fitting and deploying the aforementioned skill rating tools on real datasets. We emphasize the key steps of model design and inference and explain how different components can be incorporated to improve prediction, robustness, and interpretability.
- An open-source, lightweight, and extensible Python package is provided to ensure complete reproducibility of the results and facilitate the use of the proposed and reviewed methodologies (github.com/SamDuffield/abile).

The article will proceed as follows. In Section 2, we describe a high-level formalism for state-space formulations of the skill estimation problem. In Section 3, we outline the possible inference objectives for this problem and how they interact as well as the general structure of tractable inference, induced approximations and relevant complexity considerations. In Section 4, we review a variety of concrete procedures for the skill estimation problem, combining probabilistic models and

Table 1. Considered approaches and their features

Method	Skills	Filtering	Smoothing	Parameter estimation	Sources of error (beyond factorial)
Elo	Continuous	Location , $\mathcal{O}(1)$	N/A	N/A	Not model-based
Glicko	Continuous	Location and Spread, $\mathcal{O}(1)$	Location and Spread, $\mathcal{O}(1)$	N/A	Not model-based
Extended Kalman	Continuous	Location and Spread, $\mathcal{O}(1)$	Location and Spread, $\mathcal{O}(1)$	EM	Gaussian Approximation
TrueSkill2	Continuous	Location and Spread, $\mathcal{O}(1)$	Location and Spread, $\mathcal{O}(1)$	EM	Gaussian Approximation
SMC	General	Full Distribution, $\mathcal{O}(J)$	Full Distribution, $\mathcal{O}(J)$[1]	EM	Monte Carlo Variance
Discrete	Discrete	Full Distribution, $\mathcal{O}(S^2)$	Full Distribution, $\mathcal{O}(S^2)$[2]	(Gradient) EM	N/A

[1]The computational cost of SMC smoothing can be reduced from $\mathcal{O}(J^2)$ to $\mathcal{O}(J)$ using rejection sampling (Douc *et al.*, 2011) or MCMC (Dau & Chopin, 2023) strategies. [2]Full smoothing for a continuous-time fHMM has complexity $\mathcal{O}(S^3)$, however for factorized random walk dynamics, marginal smoothing and gradient EM can be applied at cost $\mathcal{O}(S^2)$, see A. 6.3 in the online supplementary material. *Note*. All approaches are linear in the number of players $\mathcal{O}(N)$ and the number of matches $\mathcal{O}(K)$.

inference algorithms unifying existing approaches and introducing new ones. In Section 5, we perform some numerical experiments, demonstrating and contrasting some of the varied approaches to real data. Finally, in Section 6, we conclude with a discussion on general recommendations and extensions.

2 Model specification

Throughout, we will use various terminology which is specific to the skill estimation problem. We will use 'sport' to denote a sport in the abstract (e.g. football), and will use 'match' to denote a specific instance of this sport being played (e.g. a match between two football teams). All matches will be contested between two competing 'players' (e.g. a football team is a 'player'), one of whom is designated as the 'home' player, and the other as the 'away' player (even for sports in which there is no notion of 'home ground' or 'home advantage'). A 'competition' refers to a specific sport, a collection of players of that sport, a set of matches between these players, and a set of results of these matches (e.g. the results of the English Premier League). Each match is also associated to a time $t \in [0, \infty)$ denoting when the match was played, which we call a 'matchtime'. Driven by the popularity and simplicity of pairwise comparison (Gorgi *et al.*, 2019; Wheatcroft, 2021), we will focus on the aforementioned scenario where a 'match' is played by only two 'players', although the framework is general and easily extends to multiplayer comparisons which we later discuss in Section 4.6.

Notationally, given an integer $N \in \mathbf{N}$ we use $[N]$ for the set of integers from 1 to N. The total number of players in the competition is denoted $N \in \mathbf{N}$, and $K \in \mathbf{N}$ denotes the total number of matches played in the competition. We order the matchtimes as $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_K$ (noting that matches may take place contemporaneously). Matches are then indexed in correspondence with this ordering, i.e. match 1 took place at time $t = t_1$, etc.; we also adopt the convention that $t_0 = 0$. We explicitly model the skills of all players over the full-time window $[0, T]$, even if individual players may enter the competition at a later time.

We now discuss modelling choices. Our key interest is to infer the skills of players in a fixed competition, given access to the outcomes of matches which are played in that competition. We model player skills as taking values in a totally ordered set $\mathcal{X}$, i.e. skills are treated as ordinal, with the convention that higher skill ratings are indicative of more favourable match outcomes for a player. The structure of the match outcomes (denoted y_k for the outcome of the kth match) depends on the application and model design. Most simply and commonly, they can be represented by a finite, discrete set $y_k \in \mathcal{Y}$, e.g. $\mathcal{Y} = \{\text{draw, home win, away win}\}$. Other structures are also readily accommodated, e.g. recording the number of points scored by each team as $y_k \in \mathcal{Y} = \mathbf{N}^2$.

Our generic model will consist of the players' skills and the match outcomes; we detail here the construction of the full joint likelihood. Players' skills are allowed to vary with time, and we write $x_t^i \in \mathcal{X}$ for the skill value of the ith player at time t and in an abuse of notation, we will also use x_k^i to denote the skill value of the ith player at observation time t_k. We will, where possible, index skills with letters t or k so that the continuous/discrete nature of the time index is clear from the context. For discrete-time indices, we also make use of the notation $x_{0:k}$ or $y_{1:k}$ to denote the joint vector of skills or observations at discrete times.

We assume that the initial skills of the players are all drawn mutually independently of one another, i.e. for $i \in [N]$, $x_0^i \sim m_0^i$ independently for some initial distribution m_0^i. We also assume that the evolution of each player's skill over time is independent of one another. We further assume that each of these evolutions is Markovian, i.e. for each $i \in [N]$, there is a semigroup of $\mathcal{X}$-valued Markov transition kernels $(M_{t,t'}^i : t \leqslant t')$ such that $x_{t'}^i \mid x_t^i \sim M_{t,t'}^i(x_t^i, \cdot)$ where $t \leqslant t'$ represent matchtimes. Again, we will often abuse notation to write $M_{k,k'}^i = M_{t_k,t_{k'}}^i$ with $t_k \leqslant t_{k'}$ and the nature of the index clear from context.

For the kth match, we write $h(k), a(k) \in [N]$ for the indices of the home and away players in that match, and write $y_k \in \mathcal{Y}$ for the outcome of the match. We assume that given the skills of players $h(k)$ and $a(k)$ at time t_k, the match's outcome y_k is conditionally independent of all other player skills and match outcomes (depicted in Figure 1) and is drawn according to some probability distribution $G_k(y_k \mid x_k^{h(k)}, x_k^{a(k)})$. We again emphasize that practitioners are completely free to choose the observation model that suits their application, whether sigmoidal, Poisson, or otherwise. The only structural condition imposed upon G_k is that the observation depends only on the few players that are involved in the match, i.e. $\{h(k), a(k)\}$.

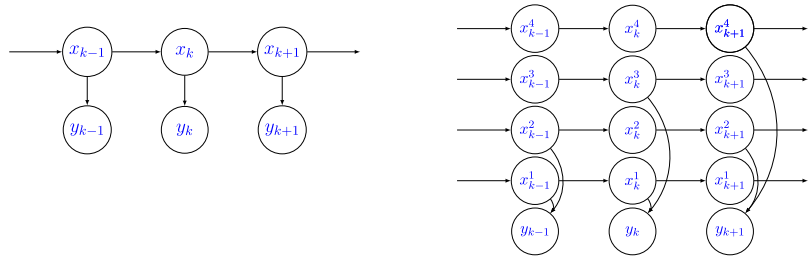

Figure 1. Left: conditional independence structure of an SSM. Right: conditional independence structure of an fSSM (1), with $N = 4$ and pairwise observation model.

Assembling these various assumptions, it follows that the joint law of all players' skills on all matchtimes, and of all match results, is thus given by the following

$$\mathbf{P}\left(x^{[N]}_{0:K}, y_{1:K}\right) = \prod_{i\in[N]}\left\{m^i_0\left(x^i_0\right)\cdot\prod_{k=1}^{K} M^i_{k-1,k}\left(x^i_{k-1}, x^i_k\right)\right\}\cdot\prod_{k=1}^{K} G_k\left(y_k \mid x^{b(k)}_k, x^{a(k)}_k\right). \tag{1}$$

Some simplifications which we will make in all subsequent examples are that (i) we will model the initial laws of all players' skill as being identical across players (i.e. m^i_0 will not depend on i), (ii) the dynamics of all players' skills will also be identical across players, i.e. for $t < t'$, we can write $M^i_{t,t'}(x, x') = M_{t,t'}(x, x')$, and (iii) the observation model G_k will not depend on k (although we will still use the notation G_k to emphasize dependence on y_k). These simplifications are made for ease of presentation, and deviations from each of these simplifications are typically straightforward to accommodate in the algorithms which we present. Such deviations are often relevant in practice, e.g. different match types (e.g. 3-set vs. 5-set tennis).

There are also a number of model features on which we insist, namely (i) we insist on modelling player skills as evolving in continuous time and (ii) we insist on the possibility of observations which occur at irregularly spaced intervals in time. We do this because these settings are of practical relevance, and because they pose particular computational challenges which deserve proper attention.

Terminologically, we use the term 'state-space model' (SSM) to denote a model, such as (1), in which there is an unobserved state x, taking values in a general space, which evolves in time according to a Markovian evolution, and is observed indirectly. We use the term 'hidden Markov model' (HMM) to refer to an SSM where the unobserved state x takes values on a finite state space. The term factorial state-space model (fSSM, or indeed fHMM) refers to a state-space model in which the state x is naturally partitioned into a collection of sub-states, each of which evolves independently of one another (Ghahramani & Jordan, 1995), see Figure 1.

3 Inference

In this section, we discuss the problem of inference in the skill rating problem, which features of the problem one might seek to understand, and how one might go about representing these features.

3.1 Inference objectives

Broadly speaking, inference in general state-space models tends to involve the solution of three related tasks, presented in (roughly) increasing order of complexity:

1. Filtering: inferring the current latent states, given the observations thus far, as well as prediction

$$\text{Filter}_k\left(x_k\right) := \mathbf{P}\left(x_k \mid y_{1:k}\right), \quad \text{Predict}_{k+1|k}\left(x_{k+1}\right) = \mathbf{P}\left(x_{k+1} \mid y_{1:k}\right).$$

for $t_k < t_{k+1}$.

2. Smoothing: inferring past latent skills, given the observations thus far

$$\text{Smooth}_{0:K\mid K}(x_{0:K}) = \mathbf{P}\big(x_{0:K} \mid y_{1:K}\big).$$

with $\text{Smooth}_{k,k+1\mid K}(x_k, x_{k+1}) = \mathbf{P}(x_k, x_{k+1} \mid y_{1:K})$ and $\text{Smooth}_{k\mid K}(x_k) = \mathbf{P}(x_k \mid y_{1:K})$.

3. Parameter estimation: when the dynamical and/or observational structure of the model depends on unknown parameters θ, one can calibrate these models based on the observed data by e.g. maximum-likelihood estimation:

$$\underset{\theta\in\Theta}{\arg\max}\,\mathbf{P}\,(y_{1:K} \mid \theta),$$

where $\mathbf{P}(y_{1:K} \mid \theta) = \int \mathbf{P}(x^{[N]}_{0:K}, y_{1:K} \mid \theta)\,\mathrm{d}x^{[N]}_{0:K}$. Note that for filtering and smoothing, θ is treated as constant, hence its omission from the preceding descriptions.

Depending on the application in question, each of these tasks can be of more or less interest. In our setting, because we are interested in using our model to make real-time decisions, filtering is directly relevant towards informing those decisions. However, this does not mean that we can immediately ignore the other two tasks. Firstly, without an accurate estimate of the parameters θ which govern the dynamical and observation models for our process of interest, our estimate of the filtering distribution can be badly misspecified. As a result, without incorporating some elements of parameter estimation, inference of the latent states can be quite poor. Moreover, obtaining good estimates of these model parameters tends to require developing an understanding of the full trajectory of the latent process; indeed, many algorithms for parameter estimation in SSMs require some form of access to the smoothing distribution, which in turn requires the filtering distributions. As such, the three tasks are deeply interconnected.

3.2 Techniques for filtering

We first present some generalities on algorithmic approaches to the filtering and smoothing problem, to contextualize the forthcoming developments. For a general SSM on $\mathcal{X}$ with transitions $M_{t,t'}$ and observations G_k, the following abstract filtering recursions hold

$$\begin{aligned}\text{Predict}_{k+1\mid k}(x_{k+1}) &= \int \text{Filter}_k(x_k)\cdot M_{k,k+1}(x_k, x_{k+1})\,\mathrm{d}x_k,\\ \text{Filter}_{k+1}(x_{k+1}) &\propto \text{Predict}_{k+1\mid k}(x_{k+1})\cdot G_{k+1}(x_{k+1}),\end{aligned}$$

or more suggestively,

$$\begin{aligned}\text{Predict}_{k+1\mid k} &= \mathsf{Propagate}\big(\text{Filter}_k; M_{k,k+1}\big),\\ \text{Filter}_{k+1} &= \mathsf{Assimilate}\big(\text{Predict}_{k+1\mid k}; G_{k+1}\big),\end{aligned}$$

where we note that the operators Propagate and Assimilate act on probability measures. Note that the Propagate operator can also readily be applied to times which are not associated with matches; this can be useful for forecasting purposes.

The same recursions also allow for computation of the likelihood of all observations so far, using that $\mathbf{P}(y_{1:k+1}) = \mathbf{P}(y_{1:k})\cdot\mathbf{P}(y_{k+1} \mid y_{1:k})$, the last term carries the interpretation of a predictive likelihood or in the skill-rating setting, match outcome predictions.

3.3 Techniques for smoothing

Similarly to filtering, there are 'backward' recursions which characterize the smoothing laws: if no observations occur in the interval (t_k, t_{k+1}), then

$$\begin{aligned}\text{Smooth}_{k,k+1\mid K}(x_k, x_{k+1}) &= \frac{\text{Filter}_k(x_k)\cdot M_{k,k+1}(x_k, x_{k+1})\cdot \text{Smooth}_{k+1\mid K}(x_{k+1})}{\text{Predict}_{k+1\mid k}(x_{k+1})},\\ \text{Smooth}_{k\mid K}(x_k) &= \int \text{Smooth}_{k,k+1\mid K}(x_k, x_{k+1})\,\mathrm{d}x_{k+1},\end{aligned}$$

or

$$\text{Smooth}_{k,k+1\,|\,K} = \textsf{Bridge}\big(\text{Filter}_k, \text{Smooth}_{k+1\,|\,K}; M_{k,k+1}\big)$$
$$\text{Smooth}_{k\,|\,K} = \textsf{Marginalize}\big(\text{Smooth}_{k,k+1\,|\,K}; k\big).$$

Smoothing between observation times is possible using the same approach, i.e. for $t_k \leqslant t_{k'} \leqslant t_{k+1}$ where $t_{k'}$ is not associated with an observation, we have

$$\text{Smooth}_{k',k+1\,|\,K} = \textsf{Bridge}\big(\text{Predict}_{k'\,|\,k}, \text{Smooth}_{k+1\,|\,K}; M_{k',k+1}\big).$$

In general, exact implementation of any of {Propagate, Assimilate, Bridge, Marginalize} tends to only be possible in models with substantial conjugacy properties, due to the general difficulty of integration and representation of probability measures of even moderate complexity. If one insists on an exact implementation of all of these operations, then one tends to be restricted to working with linear Gaussian SSMs or HMMs of moderate size. As such, practical algorithms must often make approximations which restore some level of tractability to the model.

Observe also that, given the filtering distributions, the smoothing recursions require no further calls to the likelihood term G_k, which can be noteworthy in the case that the likelihood is computationally expensive or otherwise complex.

3.4 Standing approximations and reductions

We are interested in developing procedures for filtering, smoothing, and parameter estimation involving (i) many players, i.e. $N \to \infty$ and (ii) many matches, i.e. $K \to \infty$. We will therefore focus on procedures for which the computational cost scales at most *linearly* in each of N and K.

3.4.1 Decoupling approximation

In seeking procedures with stable behaviour as the number of players grows, there is one approximation which seems to be near-universal in the setting of pairwise comparisons, and beyond, namely that the filtering (and smoothing) distributions over all of the players' skills are well-approximated by a decoupled representation. Using superscripts to index player-specific distributions (e.g. Filter_k^i denoting the filtering distribution for the skill of player i at time t_k, and so on), this corresponds to the approximation

$$\text{Filter}_k = \mathbf{P}\Big(x_k^{[N]} \mid y_{1:k}\Big) \approx \prod_{i\in[N]} \mathbf{P}\big(x_k^i \mid y_{1:k}\big) = \prod_{i\in[N]} \text{Filter}_k^i$$
$$\text{Smooth}_{\circ|K} = \mathbf{P}\big(x_\circ^{[N]} \mid y_{1:K}\big) \approx \prod_{i\in[N]} \mathbf{P}\big(x_\circ^i \mid y_{1:K}\big) = \prod_{i\in[N]} \text{Smooth}_{\circ|K}^i,$$

where we have $\circ = \{k\}$, $\{k, k+1\}$ or $\{0:K\}$ for the various marginal and joint smoothing objectives. This approximation is motivated by the practical difficulty of representing large systems of correlated variables and is further supported by the standing assumption that players' skills evolve independently of one another a priori.

3.4.2 Match sparsity

Our general formulation of the joint model includes more information than is strictly necessary. Due to the conditional independence structure of the model, one sees that instead of monitoring the skills of all players during all matches, it is sufficient to keep track of only the skills of players at times when they are playing in matches. Reformulating a joint likelihood which reflects this simplicity requires the introduction of some additional notation, but dramatically reduces the cost of working with the model, and is computationally crucial.

To this end, for $i \in [N]$, write $L^i \subseteq \{0, \ldots, K\}$ for the ordered indices of matches in which player i has played, and for $\ell^i \in L^i$, write $\ell^{i,-}$ for the element of L^i immediately before ℓ^i, i.e. $\ell^{i,-} = \sup\{\ell \in L^i : \ell < \ell^i\}$. It then holds that

$$\mathbf{P}\left(\left\{x^i_{\ell^i} : \ell^i \in L^i,\, i \in [N]\right\}, y_{1:K}\right) = \prod_{i\in[N]} \left\{ m_0\left(x^i_0\right) \cdot \prod_{\ell^i \in L^i} M_{\ell^{i,-},\ell^i}\left(x^i_{\ell^{i,-}}, x^i_{\ell^i}\right) \right\} \cdot \prod_{k=1}^{K} G_k\left(y_k \mid x_k^{h(k)}, x_k^{a(k)}\right). \tag{2}$$

Whilst our original likelihood (1) contained $\mathcal{O}(N \cdot K)$ terms, this new representation involves only $\mathcal{O}(N + K)$ terms. Given that the competition consists of N players and K matches, we see that this representation is essentially minimal.

3.4.3 *Pairwise updates*

A consequence of these two features is that when carrying out filtering and smoothing computations, only sparse access to the skills of players is required. In particular, consider assimilating the result of the kth match into our beliefs of the players' skills. Since this match involves the players $h(k)$ and $a(k)$, which we refer to as h and a, we have $k \in L^h \cap L^a$ and the filtering update requires only the following steps:

1. Compute the last matchtime indices on which the two players played $k^{h,-} \in L^h$ and $k^{a,-} \in L^a$, and retrieve the filtering distributions of the two players' skills on these matchtimes, i.e. $\text{Filter}^h_{k^{h,-}}$ and $\text{Filter}^a_{k^{a,-}}$ respectively.
2. Compute the predictive distributions of the two players' skills prior to the current matchtime by propagating them through the dynamics, i.e. $M_{k^{h,-},k}$ for the home player and $M_{k^{a,-},k}$ for the away player, compute

$$\begin{pmatrix} \text{Predict}^h_{k|k^{h,-}} \\ \text{Predict}^a_{k|k^{a,-}} \end{pmatrix} = \begin{pmatrix} \mathsf{Propagate}\left(\text{Filter}^h_{k^{h,-}}; M_{k^{h,-},k}\right) \\ \mathsf{Propagate}\left(\text{Filter}^a_{k^{a,-}}; M_{k^{a,-},k}\right) \end{pmatrix}.$$

(3) Compute the current filtering distributions of the two players' skills immediately following the current matchtime by assimilating the new result

$$\text{Filter}^{h,a}_k = \mathsf{Assimilate}\left(\begin{pmatrix} \text{Predict}^h_{k|k^{h,-}} \\ \text{Predict}^a_{k|k^{a,-}} \end{pmatrix}; G_k\right),$$

where Assimilate denotes a Bayesian procedure that converts predictive distributions for a pair of players and a likelihood into a joint filtering distribution.

(4) Marginalize to regain the factorial approximation

$$\begin{pmatrix} \text{Filter}^h_k \\ \text{Filter}^a_k \end{pmatrix} = \begin{pmatrix} \mathsf{Marginalize}\left(\text{Filter}^{h,a}_k; h\right) \\ \mathsf{Marginalize}\left(\text{Filter}^{h,a}_k; a\right) \end{pmatrix},$$

where Marginalize is used for a marginalization in space and not in time.

Note briefly that if at a specific time, multiple matches are being played between a disjoint set of players, then this structure implies that the results of all of these matches can be assimilated independently and in parallel; for high-frequency competitions in which many matches are played simultaneously, this is an important simplification. A key takeaway from this observation is then that the cost of assimilating the result of a single match is independent of both N and K.

Similar benefits are also available for smoothing updates. Indeed, the benefits of sparsity are even more dramatic in this case, since the smoothing recursions (Section 3.3) decouple *entirely*

across players, implying that all smoothing distributions can be computed independently and in parallel. Given access to sufficient compute parallelism, this implies the possibility of computing these smoothing laws in time $\mathcal{O}(\max\{\mathbf{card}(L^i) : i \in [N]\})$, which is potentially much smaller than the serial complexity of $\mathcal{O}(K)$. For example, if each player is involved in the same number of matches ($\sim K/N$), then the real-time complexity is reduced by a factor of N.

For ease of exposition, in what follows, we will present all considered methods with notation corresponding to the $\mathcal{O}(N \cdot K)$ posterior (1), rather than the memory-efficient $\mathcal{O}(N + K)$ posterior (2) which should be applied in practice, e.g. we will describe the Propagate step as taking us from time $k-1$ to time k, and so on.

3.5 Techniques for parameter estimation

Recall the goal of parameter estimation is to infer the unknown static (i.e. non-time-varying) parameters θ of the state-space model. A complete discussion of parameter estimation in state-space models would be time-consuming. In this work, we simply focus on estimation following the principle of maximum likelihood, due to its generality and compatibility with typical approximation schemes. Careful discussion of other approaches (e.g. composite likelihoods, Bayesian estimation, etc.; see e.g. Andrieu *et al.*, 2010; Varin *et al.*, 2011; Varin & Vidoni, 2008) is omitted for reasons of space. Additionally, we limit our attention to offline parameter estimation, where we have access to historical match outcomes and look to find static parameters θ which model them well, with the goal of subsequently using these parameters in the online setting (i.e. filtering). Techniques for online parameter estimation (with a focus on particle methods) are reviewed in Kantas *et al.* (2015).

A popular and general option for parameter estimation is an expectation–maximization (EM) strategy for maximizing the likelihood $\mathbf{P}(y_{1:K} \mid \theta)$; see e.g. Neal and Hinton (1998) or Chapter 14 in Chopin and Papaspiliopoulos (2020) for overview. EM is an iterative approach, with each iteration consisting of two steps, known as the E-step and M-step, respectively. The usual expectation step ('E-step') consists of taking the current parameter estimate $\hat{\theta}$, using it to form the smoothing distribution $\mathbf{P}(x_{0:K}^{[N]} \mid y_{1:K}, \hat{\theta})$, and constructing the surrogate objective function

$$\mathbf{Q}(\theta \mid \hat{\theta}) := \int \mathbf{P}\left(x_{0:K}^{[N]} \mid y_{1:K}, \hat{\theta}\right) \cdot \log \mathbf{P}\left(x_{0:K}^{[N]}, y_{1:K} \mid \theta\right) \mathrm{d}x_{0:K}^{[N]}, \tag{3}$$

which is a lower bound on $\log \mathbf{P}(y_{1:K} \mid \theta)$. The maximization step ('M-step') then consists of deriving a new estimate by maximizing this surrogate objective. When these steps are carried out exactly, each iteration of the EM algorithm is then guaranteed to ascend the likelihood $\mathbf{P}(y_{1:K} \mid \theta)$, and will thus typically yield a local maximizer when iterated until convergence.

Depending on the complexity of the model at hand, it can be the case that either the E-step, the M-step, or both cannot be carried out exactly. For the models considered in this work, the intractability of the smoothing distribution means that the E-step cannot be carried out exactly. As such, we will simply approximate the E-step by treating our approximate smoothing distribution as exact (noting that this compromises EM's usual guarantee of ascending the likelihood function). Similarly, when the M-step cannot be carried out in closed form, one can often approximate the maximizer of $\theta \mapsto \mathbf{Q}(\theta \mid \hat{\theta})$ through the use of numerical optimization schemes.

In some cases, constructing the smoothing distribution in a way that provides a tractable M-step may be expensive, and it is cheaper to directly form the log-likelihood gradient. By Fisher's identity (see e.g. Del Moral *et al.*, 2010), this gradient takes the following form

$$\nabla_\theta \log \mathbf{P}(y_{1:K} \mid \theta) = \int \mathbf{P}\left(x_{0:K}^{[N]} \mid y_{1:K}, \theta\right) \cdot \nabla_\theta \log \mathbf{P}\left(x_{0:K}^{[N]}, y_{1:K} \mid \theta\right) \mathrm{d}x_{0:K}^{[N]}. \tag{4}$$

As such, when the smoothing distribution is directly available, this offers a route to implementing a gradient method for optimizing $\log \mathbf{P}(y_{1:K} \mid \theta)$. When only approximate smoothing distributions are available, one obtains an inexact gradient method, which may nevertheless be practically useful.

It is important to note that the logarithmic nature of the expectations in (3)–(4) means that for many models, the components of θ can be treated independently. For example, when the parameters which influence m_0, $M_{t,t'}$ and G_k are disjoint, the intermediate objective $\mathbf{Q}(\theta \mid \hat{\theta})$ will be

separable with respect to this structure, which can simplify implementations. Moreover, depending on the tractability of solving each sub-problem, one can seamlessly blend the analytic maximization of some parameters with gradient steps for others, as appropriate.

4 Methods

In this section, we turn to some concrete models for two-player competition as well as natural inference procedures. We here focus on the key components of the approaches, highlighting the probabilistic model used (for model-based methods) and the inference paradigm. For ease of presentation, our initial focus will be on pairwise comparison and sigmoidal observations with match outcomes $y \in \mathcal{Y} \subseteq \{\text{draw, home win, away win}\}$, but extensions to a multiplayer scenario and more general observation models are discussed in Section 4.6. Detailed recursions for each approach can be found in Section A of the online supplementary material. The presented methods alongside notable features are summarized in Table 1.

Potentially the simplest model for latent skills is a (static) Bradley–Terry model[1] (Bradley & Terry, 1952), wherein skills take values in $\mathcal{X} = \mathbf{R}$ and (binary) match outcomes are modelled with the likelihood

$$G^{\mathrm{BT}}\left(y \mid x^h, x^a\right) = \begin{cases} \mathbf{sigmoid}\left(x^h - x^a\right) & \text{if } y = \mathrm{h}, \\ 1 - \mathbf{sigmoid}\left(x^h - x^a\right) & \text{if } y = \mathrm{a}, \end{cases}$$

where $\mathbf{sigmoid} : \mathbf{R} \to [0, 1]$ is an increasing function which maps real values to normalized probabilities, such that $\mathbf{sigmoid}(x) + \mathbf{sigmoid}(-x) = 1$. The full Bradley–Terry model thus takes the form

$$\mathbf{P}\left(x^{[N]}, y_{1:K}\right) = \prod_{i \in [N]} m_0^i\left(x^i\right) \cdot \prod_{k=1}^{K} G^{\mathrm{BT}}\left(y_k \mid x^{h(k)}, x^{a(k)}\right),$$

with prior m_0^i. In practice, it is relatively common to neglect the prior and estimate the players' skills through pure maximum likelihood, see e.g. Kiraly and Qian (2017).

In contrast to the other models considered in this work, this Bradley–Terry model treats player skills as static in time. Therefore, as the 'career' of each player progresses, our uncertainty over their skill level generally collapse to a point mass. We take the viewpoint that in many practical scenarios, this phenomenon is unrealistic, and so we advocate for models which explicitly model skills as varying dynamically in time. This leads naturally to the state-space model framework.

4.1 Elo

The Elo rating system (Elo, 1978), is a simple and transparent system for updating a database of player skill ratings as they partake in two player matches. Elo implicitly represents player skills as real-valued i.e. $\mathcal{X} = \mathbf{R}$, though typical presentations of Elo tend to eschew an explicit model. Skill estimates are updated incrementally in time according to the rule (for binary match outcomes)

$$\begin{pmatrix} x_k^h \\ x_k^a \end{pmatrix} = \begin{pmatrix} x_{k-1}^h + \gamma \cdot \left(\mathbb{I}\left[y_k = \mathrm{h}\right] - \mathbf{sigmoid}\left(\frac{x_{k-1}^h - x_{k-1}^a}{s}\right)\right) \\ x_{k-1}^a + \gamma \cdot \left(\mathbb{I}\left[y_k = \mathrm{a}\right] - \mathbf{sigmoid}\left(\frac{x_{k-1}^a - x_{k-1}^h}{s}\right)\right) \end{pmatrix},$$

where the sigmoid function is usually taken to be the logistic $\mathbf{sigmoid}_{\mathrm{L}}(x) = (1 + \exp(-x))^{-1}$. Here s is a scaling parameter and γ is a learning rate parameter; these are typically set empirically for each competition, e.g. for Chess (FIDE, 2023), one takes $s = 400/\log(10)$ and $\gamma \in \{10, 20, 40\}$ depending on a player's level of experience. Note that rescaling (γ, s) by a common factor leads to an essentially equivalent algorithm; it is thus mathematically convenient to work with $s = 1$ for

[1] Note that Bradley–Terry models are often (and indeed originally) described on an exponential scale i.e. in terms of $z = \exp(x) \in \mathbf{R}_+$.

purposes of identifiability. In practice, the ratio γ/s is identifiable and carries an interpretation of the speed at which player skills vary on their intrinsic scale per unit of time. The Elo rating system can be generalized to give valid normalized prediction probabilities in the case of draws (i.e. ternary match outcomes) via the Elo–Davidson system (Davidson, 1970; Szczecinski & Djebbi, 2020), the recursions for which we outline in A.1 in the online supplementary material.

The popularity of Elo arguably stems from its simplicity, as it can be well understood without a statistical background. Interestingly, it has also been shown to provide surprisingly hard-to-beat predictions in many cases (Hvattum & Arntzen, 2010; Kovalchik, 2016). However, we emphasize that Elo is not explicitly model-based, and this can make it difficult to extend to more complex scenarios, or to critique the assumptions by which it is underpinned.

4.2 Glicko and extended Kalman

It was noted in Glickman (1999) that the Elo rating system is reminiscent of a Bradley–Terry model with a dynamic element. This observation led to the development of the Glicko rating system, which explicitly seeks to take into account time-varying uncertainty over each player's latent skill ratings, i.e. representing $\text{Filter}_k^i \approx \mathcal{N}(x_k^i \mid \mu_k^i, \sigma_k^{i2})$. This enriches the Elo approach by tracking both the location and spread of player skills at a given instant. The Propagate and Assimilate steps used in Glicko invite comparison to a (local, marginal) variant of the (Extended) Kalman filter (see e.g. Chapter 7 of Särkkä & Svensson, 2023), a connection which was made formal in Ingram (2021) and later Szczecinski and Tihon (2023). Further details on Glicko can be found in Glickman (1999).

In Glicko, sports which permit draws are only heuristically permitted by treating them as 'half-victories'; this implementation does not provide normalized prediction probabilities for sports with draws, which is undesirable. By adopting the Extended Kalman filter perspective, we can readily provide a principled approach to sports with draws by considering the following (non-linear) factorial state-space model (considered in Dangauthier *et al.*, 2008 and Minka *et al.*, 2018)

$$
\begin{aligned}
m_0(x_0^i) &= \mathcal{N}(x_0^i \mid \mu_0, \sigma_0^2), \quad M_{t,t'}(x_t^i, x_{t'}^i) = \mathcal{N}(x_{t'}^i \mid x_t^i, \tau^2 \cdot (t'-t)), \\
G_k(y_k \mid x^h, x^a) &= \begin{cases} \textbf{sigmoid}\left(\frac{x^h - x^a + \epsilon}{s}\right) - \textbf{sigmoid}\left(\frac{x^h - x^a - \epsilon}{s}\right) & \text{if } y_k = \text{draw}, \\ \textbf{sigmoid}\left(\frac{x^h - x^a - \epsilon}{s}\right) & \text{if } y_k = \text{h}, \\ 1 - \textbf{sigmoid}\left(\frac{x^h - x^a + \epsilon}{s}\right) & \text{if } y_k = \text{a}. \end{cases}
\end{aligned} \tag{5}
$$

As with Elo, the logistic sigmoid function is typically used. Similarly, the parameters μ_0, σ_0, τ, and s are not jointly identifiable, due to a translation and scaling equivariance. We break this symmetry by setting $\mu_0 = 0$, $s = 1$; note that in practice, implementations of Glicko will often use alternative numerical values in service of interpretability, comparability with Elo ratings, and so on.

The state-space model perspective elucidates the interpretation of the parameters σ_0, τ, and ϵ. The initial variance σ_0^2 controls the uncertainty over the skill rating of a new player entering the database, τ is a rate parameter that controls how quickly players' skill vary over time (note that $\tau = 0$ recovers a static Bradley–Terry model) and ϵ is a draw parameter that dictates how common draws are for the given sport (for sports without draws, $\epsilon = 0$).

The original presentation of Glicko in Glickman (1999) observed that smoothing can be easily applied using the standard backward Kalman smoother (Särkkä & Svensson, 2023), which they used in service of 'pure smoothing', rather than for parameter estimation. The parameters for Glicko (σ_0 and τ) can be inferred by numerically minimizing a cross-entropy-type loss function which contrasts predicted and realized match outcomes (see Glickman, 1999, Section 4); this works reasonably in its own context, but does not necessarily scale well to more complex models with a larger parameter set. Adopting the framework described in Section 3.5, we determine maximum-likelihood estimators for the static parameters which scale gracefully to complex models and parameter sets. In particular, for (5), the convenient Gaussian form of the smoothing approximation means the maximization step of EM can be carried out analytically for σ_0 and τ, and efficiently numerically for ϵ (Duffield, 2024).

4.3 TrueSkill and expectation propagation/moment-matching

It was noted in Herbrich *et al.* (2006) that if we instead choose the sigmoid function for a static Bradley–Terry model to be the inverse probit function $\mathbf{sigmoid}_{\mathrm{IP}}(x) = \Phi(x)$ (where Φ is the CDF of a standard Gaussian), then certain integrals of interest become analytically tractable. In particular, we can use the identity $\int \mathcal{N}(z \mid \mu, \sigma^2) \cdot \Phi(z)\, \mathrm{d}z = \Phi(\mu/\sqrt{1+\sigma^2})$ to analytically calculate the marginal filtering means and variances of the non-Gaussian joint filtering distribution $\mathrm{Filter}_k^{h,a}$. This naturally motivates the moment-matching approach of Herbrich *et al.* (2006), wherein an approximate factorial filtering distribution $\mathrm{Filter}_k^{h,a} \approx \mathrm{Filter}_k^{h} \cdot \mathrm{Filter}_k^{a}$ can be defined by simply extracting the marginal means and variances. This moment-matching strategy can be seen as a specific instance of *assumed density filtering* (see e.g. Chapter 1 of Minka (2001a) and references therein) or its more general cousin, Expectation Propagation (Minka, 2001b). We provide some details on this connection in Section C in the online supplementary material. One can also reasonably consider applying other approximate filtering strategies based on Gaussian principles (e.g. Unscented Kalman Filter Julier & Uhlmann, 2004, Ensemble Kalman Filter Evensen, 2009, etc.) to the same model class; we do not explore this further here.

In the original TrueSkill (Herbrich *et al.*, 2006), this procedure was applied as an approximate inference procedure in the static Bradley–Terry model. In the follow-up works TrueSkillThroughTime (Dangauthier *et al.*, 2008) and TrueSkill2 (Minka *et al.*, 2018), the model was extended to allow the latent skills to vary over time. The resulting procedure is a treatment of the state-space model in (5), where the filtering distributions are formed by (i) applying the decoupling approximation and (ii) assimilating observations with predictions through the aforementioned moment-matching procedure. Smoothing is handled analogously to the Glicko and Extended Kalman settings, i.e. by running the Kalman smoother backwards from the terminal time. TrueSkill2 (Minka *et al.*, 2018) applies a gradient-based version of the parameter estimation techniques presented in Section 3.5, although there it is not presented explicitly in the state-space model context.

We note that the above description is TrueSkill (Herbrich *et al.*, 2006; Minka *et al.*, 2018) in its most basic form. The TrueSkill approach has been successfully applied in more complex settings, notably for online multiplayer games (Minka *et al.*, 2018).

4.4 Sequential Monte Carlo

The preceding strategies all hinge on the availability of a suitable parametric family for approximating the relevant probability distributions, enabling explicit computations and general ease of construction. This is counter-balanced by the limited flexibility of parametric approximations, where even in the presence of an increased computational budget, it is not always clear how to obtain improved estimation performance. This can sometimes be ameliorated by the use of non-parametric approximations, in the form of particle methods like sequential Monte Carlo (SMC). SMC can be applied to any state-space model for which we can (i) simulate from both the initial distribution m_0 and the Markovian dynamics $M_{t,t'}$ and (ii) evaluate the likelihood G_k. Naturally, in this work, it is of particular interest to consider the application of SMC to the (factorial) state-space model in (5). It is out of the scope of this article to review available variants of SMC. We instead prioritize conciseness by concentrating on the most widely used instance of SMC, the bootstrap particle filter.

SMC filtering maintains a (potentially weighted) particle approximation to the filtering distributions which in our context has an additional factorial approximation, reminiscent of a 'local' or 'blocked' SMC approach (Rebeschini & van Handel, 2015), $\mathrm{Filter}_k^i(x_k^i) \approx \sum_{j \in [J]} w_k^{ij} \cdot \delta(x_k^i \mid x_k^{ij})$, where $\delta(x \mid y)$ is a Dirac measure in x at point y and $j \in [J]$ indexes particles.

The bootstrap particle filter then executes Propagate by simply simulating from the dynamics $M_{k,k+1}$ to provide a new particle approximation to $\mathrm{Predict}_{k+1\mid k}^i$. Before applying the Assimilate step, the distributions $\mathrm{Predict}_{k+1\mid k}^h$ and $\mathrm{Predict}_{k+1\mid k}^a$ are paired together to form a joint distribution $\mathrm{Predict}_{k+1\mid k}^{h,a}$. The Assimilate step then consists of a reweighting step and a resampling step (to carry forward only high-probability particles) resulting in a joint weighted particle approximation to $\mathrm{Filter}_k^{h,a}$. The factorial approximation can then be regained by a simple Marginalize operation which unpairs the joint particles. Note that the factorial approximation is nonstandard in an SMC

context and is adopted here as a natural means of avoiding the curse of dimensionality that affects SMC (Rebeschini & van Handel, 2015).

Smoothing can be applied using a similar iterative importance sampling approach (Godsill *et al.*, 2004) that sweeps backwards for $k = K-1, \ldots, 0$ recycling the filtering approximations Filter_k^i into joint smoothing approximations $\text{Smooth}_{0:K|K}^i$. This procedure is (embarrassingly) parallel in both the number of players N and the number of particles J, given the filtering approximations; see Finke and Singh (2017) for a similar scenario. Parameter estimation can also be achieved by applying general-purpose EM or gradient ascent techniques from Section 3.5, where the integrals (3)–(4) are approximated using particle approximations to the smoothing law. Full details can be found in Section A.5 of the online supplementary material and Chopin and Papaspiliopoulos (2020) provides a thorough review of the field.

4.5 Finite state-space

From a modelling perspective, it is conceptually simple to consider skills which take values in a finite state-space. Recalling the general model formulated in (1), by choosing $\mathcal{X} = [S]$ for some $S \in \mathbf{N}$, one obtains a factorial hidden Markov model.

In this finite state-space, it is natural to model the skills of player i as evolving according to a continuous time Markov jump process. In the time-homogeneous case, such processes can be specified in terms of their so-called 'generator matrix' Q_S, an $S \times S$ matrix which encodes the rates at which the player's skill level moves up and down, and is typically sparse. Given such a matrix, it is typically straightforward to construct the corresponding transition kernels $M_{t,t'}$ at a cost of $\mathcal{O}(S^3)$ by diagonalization and matrix exponentiation. In many settings, such as filtering, one is not interested in the transition matrix $M_{t,t'}$ itself but rather its action on probability vectors. In this case, given appropriate pre-computations, the cost can often be controlled at the much lower $\mathcal{O}(S^2)$. We provide further details in Section A.6 of the online supplementary material.

In the context of skill ratings with pairwise comparisons, we can then consider the following fHMM inspired by (5). Writing Q_S for the generator matrix of the continuous-time random walk with reflection on $[S]$ (see Section B of the online supplementary material) and using exp for the matrix exponential, we can define

$$m_0 = v \cdot M_{0,\sigma_d}, \quad M_{t,t'} = \exp\big(\tau_d \cdot (t' - t) \cdot Q_S\big),$$
$$G_k\big(y_k \mid x^h, x^a\big) = \begin{cases} \textbf{sigmoid}\left(\dfrac{x^h - x^a + \epsilon_d}{s_d}\right) - \textbf{sigmoid}\left(\dfrac{x^h - x^a - \epsilon_d}{s_d}\right) & \text{if } y_k = \text{draw}, \\ \textbf{sigmoid}\left(\dfrac{x^h - x^a - \epsilon_d}{s_d}\right) & \text{if } y_k = \text{h}, \\ 1 - \textbf{sigmoid}\left(\frac{x^h - x^a + \epsilon_d}{s_d}\right) & \text{if } y_k = \text{a}, \end{cases} \tag{6}$$

with subscript d used to emphasize the connection to the discrete model.

Here, v is a probability vector whose mass is concentrated on the median state(s) $\{\lfloor \frac{S}{2} \rfloor, \lceil \frac{S}{2} \rceil\}$, so that m_0 resembles a (discrete) centred Gaussian law with standard deviation of order σ_d (for $\sigma_d \ll \sqrt{S}$). As one takes $\sigma_d \to \infty$, this converges towards the uniform distribution on $[S]$. Similarly to (5), for the dynamical model, we have a rate parameter $\tau_d \in \mathbf{R}^+$ which controls how quickly skill ratings vary over time, and for the observation model, we have a scaling parameter $s_d \in \mathbf{R}^+$ and a draw propensity parameter $\epsilon_d \in \mathbf{R}^+$; this can be set as $\epsilon_d = 0$ for sports without draws.

Computationally efficient filtering and smoothing follow from observing that both the filtering distributions and the transition kernel factorize across players, as in Rimella and Whiteley (2022). That is, the equations in Section 3.3 can be implemented in closed form via simple linear algebra operations.

Parameter estimation can be performed through the EM algorithm. For parameters not associated with the dynamical model, the **Q** function can be formed using only the marginal smoothing laws $\text{Smooth}_{k|K}^i$, which are available at a cost of $\mathcal{O}(S^2)$. EM updates for the dynamical parameter τ_d instead require joint laws of the form $\text{Smooth}_{k,k+1|K}^i$, which are more costly to assemble; we thus opt to instead update τ_d by a cheaper gradient ascent step.

4.6 Model extensions

The generality and flexibility of the state-space approach permits the modelling of more sophisticated data-generating processes with minimal or no alteration to the inference procedures. In simple cases, this might involve simply refining the static parameters. For instance, we could learn distinct $s_{3-\text{set}} > s_{5-\text{set}}$ to model the additional randomness associated with the shorter tennis format. Similarly, we could separate $\tau_{\text{off-season}}$ and $\tau_{\text{in-season}}$ to consider different levels of skills fluctuation during in- and off-season periods. We could even consider players' specific parameters, for example, learning player-specific initial distribution parameters or τ^i to allow player skills to evolve differently.

We now describe some more involved model modifications including how popular existing models can be re-framed within the state-space model framework.

Stationary dynamics: Ornstein–Uhlenbeck

A natural modification of the dynamics in (5) (which we use following the works of Glicko (Glickman, 1999) and TrueSkill (Dangauthier *et al.*, 2008)) would be to replace the Brownian motion skill evolution with that of an Ornstein–Uhlenbeck process

$$M_{t,t'}\left(x_t^i, x_{t'}^i\right) = \mathcal{N}\left(x_{t'}^i \mid x_t^i \cdot e^{-\tau(t'-t)} + \mu_0 \cdot \left(1 - e^{-\tau(t'-t)}\right), \sigma_0^2 \cdot \left(1 - e^{-2\tau(t'-t)}\right)\right).$$

Now each player's skill is mean reverting towards μ_0; this behaviour may be more desirable depending on the sport at hand. Contrarily, we may believe that the average ability of players in a sport has some underlying trend over time which may be better captured by the Brownian dynamics. Note that stationary dynamics are already used in (6) as the finite-space dynamical process reverts to the uniform distribution.

Points-based match outcomes: bivariate Poisson

The popular Dixon and Coles (1997) model uses an inflated Poisson likelihood to model the home and away goals scored in a football (or similar points scoring) match, meaning $y_k = (y_k^h, y_k^a) \in \mathbf{N}^2$. Each player's skill is considered as bivariate with $x^i = (x^{\text{att},i}, x^{\text{def},i}) \in \mathbf{R}^2$ and where $x^{\text{att},i}, x^{\text{def},i}$ model the attack and defence strength of player i respectively. More generally, Karlis and Ntzoufras (2003) use a bivariate Poisson likelihood

$$G_k(y_k \mid x^h, x^a) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{y_k^h}}{y_k^h!} \frac{\lambda_2^{y_k^a}}{y_k^a!} \sum_{k=0}^{\min\{y_k^h, y_k^a\}} \binom{y_k^h}{k} \binom{y_k^a}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k, \tag{7}$$

where $\lambda_1 = \exp(\alpha^h + x^{\text{att},h} - x^{\text{def},a})$, $\lambda_2 = \exp(\alpha^a + x^{\text{att},a} - x^{\text{def},h})$, $\lambda_3 = \exp(\beta)$ and $\alpha^h, \alpha^a, \beta \in \mathbf{R}$ are all static parameters (to accompany the static parameters for the initialization and dynamics).

These likelihoods can be easily integrated within a state-space model to infer how different teams' attacking and defensive strengths vary over time. The factorial assumption remains and the players' multivariate skills evolve independently preserving the form of Propagate. On the other hand, Assimilate should be modified to update jointly $x^{\text{att},h}, x^{\text{def},h}, x^{\text{att},a}, x^{\text{def},a}$, which is then followed by Marginalize to regain the factorial approximation.

Such model extensions can lead to new inferences of interest where one might want to query the existence of covariates representing external factors (Karlis & Ntzoufras, 2003). As a concrete example, one could examine the presence of a 'home advantage' factor in football by testing the hypothesis $\alpha^h > \alpha^a$ against $\alpha^h = \alpha^a$.

Multiplayer extension: Plackett–Luce

We have thus far focused on pairwise comparisons, where matches are assumed to be between two players. We can extend our formulation to a collection of players which can be of broad use e.g. in multiplayer games Joshy (2024), movie rankings Guiver and Snelson (2009), and consumer behaviour Luce (1959). All steps in Section 3.4.3 remain conceptually the same, with the Assimilate and Marginalize steps now acting on an enlarged intermediate joint distribution.

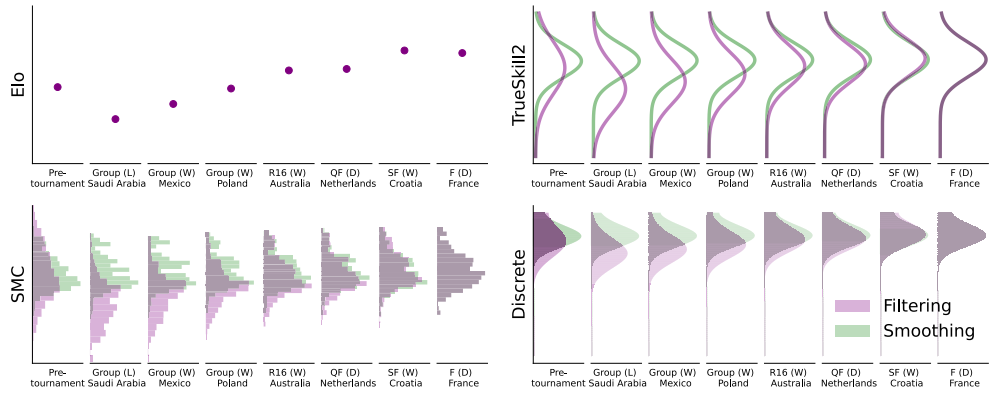

Figure 2. Visualization of the different skill representations for Argentina's 2023 FIFA World Cup triumph. Each *y*-axis represents the skill-rating scale for the different approaches.

Specifically, we can consider the model of Plackett (1975) and Luce (1959), where a *judge* ranks M players competing in match k. The observed data are $y_k = (y_k^1, \ldots, y_k^{\mathrm{M}})$, where each y_k^{m} uniquely indexes one of the M player indices, e.g. $y_k^3 = 26$ indicates the judge ranked player 26 in 3rd place. The likelihood (Guiver & Snelson, 2009) is

$$G_k\left(y_k^{1:M} \mid x^{[M]}\right) = \prod_{m=1}^{M} \frac{\exp(x^{y_k^m})}{\exp(x^{y_k^m}) + \ldots + \exp(x^{y_k^M})}.$$

As this likelihood applies to M players simultaneously (as opposed to just two previously) care must be taken with the dimensionality of the assimilation for SMC and discrete inference. In principle, the extended Kalman approach does not suffer in the same way, although the accuracy of the Taylor approximation may need to be validated.

5 Experiments

We now turn to the task of accurately quantifying sporting skills with some real-life data sets. We consider three sports; Women's Tennis Association (WTA) results (noting that all matches have the same 3-set format), football data for the English Premier League (EPL) (plus international data for Figure 2), and professional (classical format) chess matches. In all cases, the research question is to use the results of the matches (and potentially additional data, e.g. the margin of victory) to infer time-varying skill levels of all competitors alongside uncertainties in these skill estimates. We then use these skill levels to provide match predictions, whose accuracy we assess.

This section is structured to replicate a realistic workflow. We start with an exploratory analysis with some trial static parameters, testing against basic coherence checks on how we expect the latent skills to behave. We then turn to parameter estimation and learn the static parameters from historical data. Finally, we describe and analyse how filtering and smoothing can be utilized for online decision-making and historical evaluation, respectively. At each stage of the workflow, we compare and highlight similarities or differences across sports and modelling or inference approaches as appropriate (but not exhaustively).

Specific details and parameter specifications for the experiments can be found in Section D in the online supplementary material. A `python` package permitting easy application of the discussed techniques, and code to replicate all of the following simulations, can be found at github.com/SamDuffield/abile.

5.1 Exploratory analysis

The first step in any statistical procedure is to explore the model with some preliminary (perhaps arbitrarily chosen) static parameters. The goal here is not a thorough evaluation of skill ratings, but rather to assess our prior intuitions.

In Figure 2, we depict Argentina's 2023 football World Cup in terms of the evolution of their skill rating distribution with arbitrarily chosen static parameters (we run on international matches in 2020–2022, with only Argentina's post-match 2022 World Cup ratings displayed). We also take this opportunity to highlight the different approaches to encoding a probability distribution over skill ratings. We first see that Elo stores skill ratings as a point estimate without any uncertainty quantification and that it is a purely forward-based approach without any ability to update skill ratings based on future match results (i.e. smoothing). In contrast, the three model-based SSM approaches encode a (more informative) distribution over skill ratings. In the case of TrueSkill2 (Glicko and Extended Kalman would appear similar and are therefore omitted) a simple location and spread (Gaussian) distribution is used, whereas SMC and the discrete model can encode more complex distributions.

In terms of verifying our intuitions, we can see that Argentine victories increase their skill ratings, draws have little influence and defeats decrease the ratings. Particularly poignant is Argentina's defeat to Saudi Arabia (a low-ranked side according to both FIFA's official rankings and our computed filters) which in all approaches resulted in a sharp decrease in Argentina's estimated skill rating.

We can also draw insights from the smoothing distributions. We first note that at the final match in the dataset, the filtering and smoothing distributions match exactly, by definition. We also see that the smoothing distributions show less uncertainty than the filtering distributions; this matches our intuition since smoothing $\mathbf{P}(x_k \mid y_{1:K})$ has access to more data than filtering $\mathbf{P}(x_k \mid y_{1:k})$. We finally observe that the smoothing distributions are much less reactive to individual results and instead track a *smooth* trajectory of the team's skill rating over time.

5.2 Parameter estimation

Having verified informally that the algorithms are behaving sensibly, we turn to apply the techniques discussed in Section 3.5 to learn the static parameters from historical data in an offline setting. Our goal is to maximize the log-likelihood $\log \mathbf{P}(y_{1:K} \mid \theta)$.

We initially consider the WTA tennis dataset, which we train on the years 2019–2021 and leave 2022 as a test set for later. Draws do not occur in tennis, and therefore we can set $\epsilon = \epsilon_d = 0$, leaving only two parameters to tune for each approach. This two-dimensional optimization landscape of the log-likelihood can readily be visualized, as in Figure 3. Indeed, as the static parameter is only two-dimensional the optimization could be applied using a grid search (as is indeed the most natural option for Elo and Glicko). Furthermore, a filtering sweep provides an estimate of the optimization objective $\log \mathbf{P}(y_{1:K} \mid \theta)$, and therefore the grid search can be applied directly without running a smoothing routine. For more complex models and datasets, higher dimensional static parameters are inevitable, and a grid search will quickly become prohibitive. We therefore apply iterative EM to the tennis data in Figure 3 to investigate some of the properties and differences between approaches, indicative of parameter estimation in more complex situations.

For the three approaches considered (we again omit the Extended Kalman approach due to its similarity with TrueSkill2 in all steps beyond filtering), we display 1,000 EM iterations starting from three different initializations.

The TrueSkill2 and SMC approaches share the same model and differ only through their respective Gaussian and particle-based approximations of the skill distributions. The Gaussian approximation induces a significant bias, whereas the particle approximation is asymptotically unbiased, but induces Monte Carlo variance. We can see that the bias from the Gaussian approximation contorts the optimization landscape for TrueSkill2 relative to SMC, whose fuzzy landscape is due to the stochastic nature of the algorithm. We see that the additional bias from the TrueSkill2 approach results in the EM trajectory evading the global optimum; by contrast, this optimum is successfully identified by both the SMC and discrete approaches, which do not exhibit any systematic bias beyond the factorial approximation.

5.3 Filtering (for online decision-making)

Now that we have a principled choice for our static parameters, we can apply the methods to online data. In Table 2, we use static parameters (trained using grid search for Elo and Glicko, and EM for the remaining, model-based approaches). Recall that in the case of sports with draws,

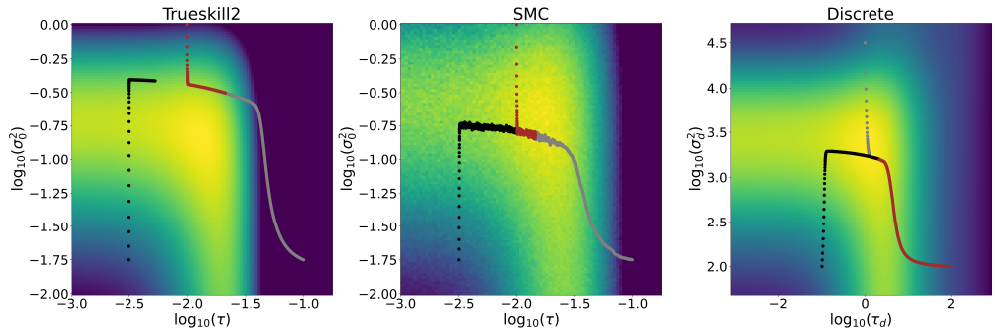

Figure 3. Log-likelihood grid and parameter estimation for WTA tennis data.

Glicko does not provide the normalized outcome predictions required for log-likelihood-based assessment.

For the tennis data, we notice that all models broadly perform quite similarly, with the exception of TrueSkill2; we suspect that this stems from the parameter estimation optimization issues discussed above. The tennis task without draws represents a simpler binary prediction problem, and it is therefore not surprising that (with principled parameter estimation) predictive performance saturates. For the more difficult tasks of Football and Chess (where draws do occur), we see that the model-based approaches equipped with uncertainty quantification significantly outperform Elo. Overall we can safely say that model-based approaches significantly outperform simple methods like Elo and Glicko and are preferable for the considered sports.

Here, we have assessed the accuracy of the approaches in predicting match outcomes, as this is a natural task in the context of online decision-making, and can be useful for a variety of purposes including seeding, scheduling, and indeed betting. Predictive distributions can also be used for more sophisticated decision-making such as those based on multiple future matches (competition outcomes, promotion/relegation results, etc.).

5.4 Smoothing (for historical evaluation)

Smoothing represents an integral subroutine for parameter estimation, though can also be of interest in its own right (Duffield & Singh, 2022; Glickman, 1999). In particular, when analysing the historical evolution of a player's skill over time, it is more appropriate to consider the smoothing distributions, rather than the filtering distributions which do not update in light of recent match results.

On the left of Figure 4, we display the historical evolution of Tottenham's EPL skill rating over time according to (5) and extended Kalman inference. When comparing filtering and smoothing, we immediately see that the smoothing distributions are less reactive and provide a more realistic trajectory of how a team's underlying skill is expected to evolve over time. Noting that the model permits a certain amount of randomness to occur in each match which can result in surprise results, we observe that the smoothing distributions do a much better job of handling this noise or uncertainty. Figure 4 only displays the extended Kalman method, but similar takeaways hold for all of the aforementioned model-based methods (as is highlighted in Figure 2).

Historical evaluation of skills can be particularly useful to analyse the impacts of various factors and how they impact the team's underlying skill level, relative to its competitors. In Figure 4, we highlight the different managers or head coaches that have served Tottenham during the time period, depicting their competitors' mean skills in the background. This can be particularly useful in evaluating the ability of the managers and their impact on the team. We observe from the smoothing output that Tottenham's skill rating was in ascendance under Villas-Boas and early Pochettino, before descending towards the end of the Pochettino era (although perhaps not as sharply as the filtering output would suggest). We note that there are likely many further factors influencing the underlying skill that are not highlighted in Figure 4, and also that (5) is relatively simple; it may of course be desirable to build a more complex model which can direct account for such additional factors or data.

Table 2. Average negative log-likelihood (low is good) for match outcomes across a variety of sports

Method	Tennis (WTA)		Football (EPL)		Chess	
	Train	Test	Train	Test	Train	Test
Elo–Davidson	0.640	0.636	1.000	0.973	0.802	1.001
Glicko	0.640	0.636	–	–	–	–
Extended Kalman	0.640	**0.635**	0.988	0.965	**0.801**	**0.972**
TrueSkill2	0.650	0.668	1.006	**0.961**	0.802	0.978
SMC	0.640	0.639	0.988	0.962	**0.801**	0.974
Discrete	**0.639**	0.636	**0.987**	**0.961**	**0.801**	0.976
Bivariate Poisson - Extended Kalman			**0.975**	0.954		
Bivariate Poisson - SMC			0.978	**0.950**		
Bivariate Poisson - Discrete			0.984	0.954		

Note. In each case, the training period was 3 years and the test period was the subsequent year. Note the draw percentages were 0% for tennis, 22% for football and 65% for chess. Note that although the bivariate Poisson likelihood models home and away goals, we here report the statistics for predicting draw, home win or away win, which are accessible as a byproduct of the bivariate likelihood.

In practice, one may want to update full smoothing trajectories in an online fashion so that historical evaluation is possible without having to run full $\mathcal{O}(K)$ backward sweeps. Here, a fixed-lag approximation is a natural option (Duffield & Singh, 2022).

5.5 Model refinement

Focusing on the case of football, simply modelling match outcomes as draw, home win or away win ignores valuable information recorded in the margin of victory and the number of goals scored.

As described in Section 4.6, we can intuitively model the number of goals scored in a match via a bivariate Poisson model (Karlis & Ntzoufras, 2003), where attack and defence strengths represent the underlying latent player skills. At the bottom of Table 2, we can see that our refined model provides better predictions for the match winner than the sigmoidal SSM (5), despite not being directly optimized for the purpose.

On the right of Figure 4, we visualize the extension of our model for attack and defensive skills, again for both filtering and smoothing. The more expressive model allows us to conclude that the key driver of Tottenham's decrease in overall skill in 2022 was due to a weakening of their defence, whilst their attack remained relatively proficient.

6 Discussion

In this work, we have advocated for a model-based approach to the skill rating problem. By taking this perspective, we can separate the tasks of modelling and inference. We have detailed several SSMs which are suitable for tackling the problem and are highly interpretable and extensible to suit more sophisticated data pipelines. We have also detailed different approximate inference schemes for analysing such models and discussed their relative strengths and shortcomings. We have conducted thorough case studies on how such methods apply to real data, highlighting a simple workflow, and the different roles of filtering, smoothing, and parameter estimation.

As we have demonstrated, a principled Bayesian approach offers both interpretable and predictive advantages over baseline methods such as Elo and Glicko. These advantages also apply to alternative algorithms that ignore the quality of the opposition, failing to provide predictions for the match outcomes and difficult to extend to more complex data settings or inferential questions. For example, it might be tempting to compare Figure 4 to a plot of a moving average of Tottenham's points-per-game over the considered period.

On the other hand, one might obtain good predictive performance via sophisticated machine learning or black-box methods (Menke & Martinez, 2008). However, one would sacrifice the

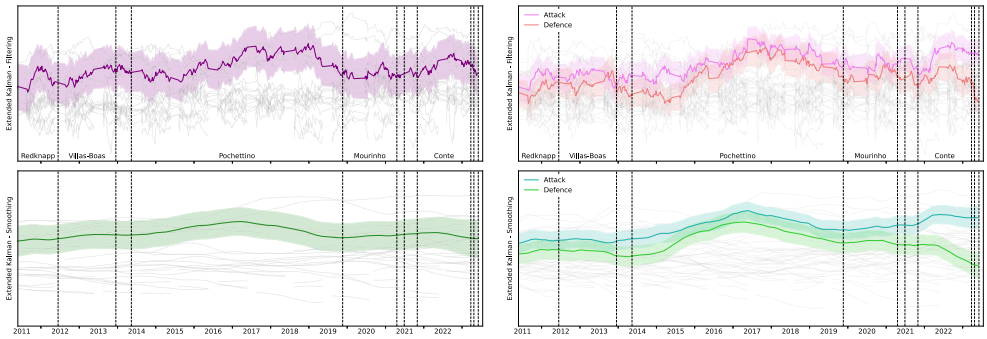

Figure 4. Extended Kalman filtering and smoothing with for Tottenham's EPL matches from 2011–2023. Top: filtering. Bottom: smoothing. Left: sigmoidal likelihood (5). Right: bivariate Poisson likelihood (7). Error bars represent one standard deviation with the other teams' mean skills in faded grey. Black dashed lines represent a change in Tottenham manager with long-serving ones named.

interpretability provided by the state-space model approach, resulting in a model that exhibits difficult-to-challenge assumptions that may fail in subtle ways (such as underfitting or overfitting). The state-space model framework is transparent and through Bayesian marginalization offers a principled route to uncertainty quantification and a broad range of predictive functionalities (such as the match result via the number of home and away goals in the bivariate Poisson model).

There is of course ample further potential to apply the same procedures to more complex models. Our techniques could even be easily applied point-by-point. For example, one could develop a model for cricket (or baseball) skills where the bowler and batter are the two 'players' with an asymmetric likelihood based on $\mathcal{Y} = \{\text{wicket, extras, 0, 1, 2, 3, 4, 6}\}$.

We further remark that point-by-point style state-space models have been considered in the 'hot-hand effect' literature (Mews & Ötting, 2023; Ötting *et al.*, 2020; Wetzels *et al.*, 2016). Here, players may enter a 'hot' state where they display exceptional performance and significantly affect the match outcome. However, the hot-hand literature primarily focuses on identifying the performance levels of individual players, neglecting cross-player interactions and their combined impact on the match (which may be a suitable assumption for sports such as darts). This focus leads to state-space models that are completely decoupled across players, resulting in a simplified scenario. In contrast, we consider state-space models for which the player skills jointly affect the observations, leading to additional inferential and computational challenges.

An appealing aspect of our framework is that the user can devote their time to carefully designing their model, and describing their data-generating process in state-space language. Having done this, the general-purpose inference schemes will typically apply directly, enabling the user to easily explore different model and parameter configurations, with the ability to refine the model in an iterative manner (Gelman *et al.*, 2020). The SSM framework also provides the tools (marginal likelihoods) required for model selection in a principled way via Bayesian model selection (Wasserman, 2000).

As a result of our specific interests in this problem, we tend to emphasize the role of filtering as a tool for online decision-making, and of smoothing for retrospective evaluation of policies (e.g. assessing coach efficacy, changes in conditions, etc.). For simplicity, we have given a comparatively lightweight treatment of parameter estimation; there are various extensions of our work in this direction which would be worthwhile to examine carefully, e.g. online parameter estimation (Cappé, 2011; Kantas *et al.*, 2015), Bayesian approaches (Andrieu *et al.*, 2010), and beyond.

With regard to practical recommendations, at a coarse resolution, we can offer that (i) when speed and scalability are of primary interest, Extended Kalman inference offers a good default, whereas (ii) when robustness and flexibility are a greater priority, the fHMM model with graph-based inference has many nice properties. We note that the task of evaluating the relative suitability of {Extended Kalman, SMC, HMM, …} approaches is not limited to the skill rating setting and is relevant across many fields and applications wherein state-space models are fundamental.

Conflicts of interest: None declared.

Funding

S. Power and L. Rimella were supported by EPSRC grant EP/R018561/1 (Bayes4Health).

Data availability

All data used are freely available online. Tennis data from tennis-data.co.uk. Football data from football-data.co.uk and github.com/martj42/international_results. Chess data from github.com/huffyhenry/forecasting-candidates. Reproducible code can be found at github.com/SamDuffield/abile.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Andrieu C., Doucet A., & Holenstein R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *72*(3), 269–342. https://doi.org/10.1111/j.1467-9868.2009.00736.x

Bradley R. A., & Terry M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. https://doi.org/10.2307/2334029

Cappé O. (2011). Online EM algorithm for hidden Markov models. *Journal of Computational and Graphical Statistics*, *20*(3), 728–749. https://doi.org/10.1198/jcgs.2011.09109

Chopin N., & Papaspiliopoulos O. (2020). *Introduction to sequential Monte Carlo*. Springer International Publishing.

Dangauthier P., Herbrich R., Minka T., & Graepel T. (2008). TrueSkill through time: Revisiting the history of chess. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems*. (Vol. 20). Curran Associates, Inc.

Dau H.-D., & Chopin N. (2023). On backward smoothing algorithms. *The Annals of Statistics*, *51*(5), 2145–2169. https://doi.org/10.1214/23-AOS2324

Davidson R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, *65*(329), 317–328. https://doi.org/10.1080/01621459.1970.10481082

Del Moral P., Doucet A., & Singh S. (2010). Forward smoothing using sequential Monte Carlo.

Dixon M. J., & Coles S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 265–280. https://doi.org/10.1111/1467-9876.00065

Douc R., Garivier A., Moulines E., & Olsson J. (2011). Sequential Monte Carlo smoothing for general state space hidden Markov models. *The Annals of Applied Probability*, *21*(6), 2109–2145. https://doi.org/10.1214/10-AAP735

Duffield S. (2024). ghq: Gauss-Hermite quadrature in JAX.

Duffield S., & Singh S. S. (2022). Online particle smoothing with application to map-matching. *IEEE Transactions on Signal Processing*, *70*, 497–508. https://doi.org/10.1109/TSP.2022.3141259

Elo A (1978). *The rating of chessplayers, past and present*. Ishi Press.

Evensen G. (2009). *Data assimilation: The ensemble Kalman filter*. (Vol. 2). Springer.

FIDE (2023). International chess federation.

Finke A., & Singh S. S. (2017). Approximate smoothing and parameter estimation in high-dimensional state-space models. *IEEE Transactions on Signal Processing*, *65*(22), 5982–5994. https://doi.org/10.1109/TSP.2017.2733504

Gelman A., Vehtari A., Simpson D., Margossian C. C., Carpenter B., Yao Y., Kennedy L., Gabry J., Bürkner P.-C., & Modrák M. (2020). Bayesian workflow.

Ghahramani Z., & Jordan M. (1995). Factorial hidden Markov models. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems*. (Vol. 8). MIT Press.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(3), 377–394. https://doi.org/10.1111/1467-9876.00159

Godsill S. J., Doucet A., & West M. (2004). Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, *99*(465), 156–168. https://doi.org/10.1198/016214504000000151

Gorgi P., Koopman S. J., & Lit R. (2019). The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(4), 1393–1409. https://doi.org/10.1111/rssa.12464
Guiver J., & Snelson E. (2009). Bayesian inference for Plackett-Luce ranking models. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 377–384). https://doi.org/10.1145/1553374.1553423
Herbrich R., Minka T., & Graepel T. (2006). TrueSkill™: A Bayesian skill rating system. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems*. (Vol. 19). MIT Press.
Hvattum L. M., & Arntzen H. (2010). Using Elo ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470. https://doi.org/10.1016/j.ijforecast.2009.10.002
Ingram M. (2021). How to extend Elo: A Bayesian perspective. *Journal of Quantitative Analysis in Sports*, *17*(3), 203–219. https://doi.org/10.1515/jqas-2020-0066
Joshy V. (2024). OpenSkill: A faster asymmetric multi-team, multiplayer rating system. *Journal of Open Source Software*, *9*(93), 5901. https://doi.org/10.21105/joss
Julier S. J., & Uhlmann J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, *92*(3), 401–422. https://doi.org/10.1109/JPROC.2003.823141
Kantas N., Doucet A., Singh S. S., Maciejowski J., & Chopin N. (2015). On particle methods for parameter estimation in state-space models. *Statistical Science*, *30*(3), 328–351. https://doi.org/10.1214/14-STS511
Karlis D., & Ntzoufras I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381–393. https://doi.org/10.1111/1467-9884.00366
Kiraly F. J., & Qian Z. (2017). Modelling competitive sports: Bradley-Terry-Elo models for supervised and online learning of paired competition outcomes.
Kovalchik S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*(3), 127–138. https://doi.org/10.1515/jqas-2015-0059
Luce R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.
Menke J. E., & Martinez T. R. (2008). A Bradley–Terry artificial neural network model for individual ratings in group competitions. *Neural Computing and Applications*, *17*(2), 175–186. https://doi.org/10.1007/s00521-006-0080-8
Mews S., & Ötting M. (2023). Continuous-time state-space modelling of the hot hand in basketball. *AStA Advances in Statistical Analysis*, *107*(1-2), 313–326. https://doi.org/10.1007/s10182-021-00410-y
Minka T., Cleven R., & Zaykov Y. (2018, March). *TrueSkill 2: An improved Bayesian skill rating system* (Technical Report MSR-TR-2018-8). Microsoft.
Minka T. P. (2001a). *A family of algorithms for approximate Bayesian inference* [Ph. D. thesis]. Massachusetts Institute of Technology.
Minka T. P. (2001b). Expectation propagation for approximate Bayesian inference. In J. S. Breese, & D. Koller (Eds.), *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001* (pp. 362–369). Morgan Kaufmann.
Neal R. M., & Hinton G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). Springer.
Ötting M., Langrock R., Deutscher C., & Leos-Barajas V. (2020). The hot hand in professional darts. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(2), 565–580. https://doi.org/10.1111/rssa.12527
Pelánek R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers and Education*, *98*, 169–179. https://doi.org/10.1016/j.compedu.2016.03.017
Plackett R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *24*(2), 193–202. https://doi.org/10.2307/2346567
Rebeschini P., & van Handel R. (2015). Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, *25*(5), 2809–2866. https://doi.org/10.1214/14-AAP1061
Rimella L., & Whiteley N. (2022). Exploiting locality in high-dimensional Factorial hidden Markov models. *Journal of Machine Learning Research*, *23*(4), 1–34. http://jmlr.org/papers/v23/19-267.html
Särkkä S., & Svensson L. (2023). *Bayesian filtering and smoothing*. (Vol. *17*). Cambridge University Press.
Stefani R. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, *7*(4). https://doi.org/10.2202/1559-0410.1347
Štrumbelj E., & Vračar P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, *28*(2), 532–542. https://doi.org/10.1016/j.ijforecast.2011.01.004
Szczecinski L., & Djebbi A. (2020). Understanding draws in Elo rating algorithm. *Journal of Quantitative Analysis in Sports*, *16*(3), 211–220. https://doi.org/10.1515/jqas-2019-0102
Szczecinski L., & Tihon R. (2023). Simplified Kalman filter for on-line rating: One-fits-all approach. *Journal of Quantitative Analysis in Sports*, *19*(4), 295. https://doi.org/10.1515/jqas-2021-0061

Varin C., Reid N., & Firth D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 5–42.

Varin, C., & Vidoni, P. (2008). Pairwise likelihood inference for general state space models. *Econometric Reviews*, *28*(1-3), 170–185. https://doi.org/10.1080/07474930802388009

Wasserman L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

Wetzels R., Tutschkow D., Dolan C., Van der Sluis S., Dutilh G., & Wagenmakers E.-J. (2016). A Bayesian test for the hot hand phenomenon. *Journal of Mathematical Psychology*, *72*, 200–209. https://doi.org/10.1016/j.jmp.2015.12.003

Wheatcroft E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics*, *7*(2), 77–97. https://doi.org/10.3233/JSA-200462