A Ranking Model Motivated by Nonnegative Matrix Factorization with Applications to Tennis Tournaments

Rui Xia[1], Vincent Y. F. Tan[1], Louis Filstroff[2], and Cédric Févotte[2]

[1] Department of Mathematics, National University of Singapore, SG
rui.xia@u.nus.edu, vtan@nus.edu.sg
[2] IRIT, Université de Toulouse, CNRS, France
{louis.filstroff, cedric.fevotte}@irit.fr

Abstract. We propose a novel ranking model that combines the Bradley-Terry-Luce probability model with a nonnegative matrix factorization framework to model and uncover the presence of latent variables that influence the performance of top tennis players. We derive an efficient, provably convergent, and numerically stable majorization-minimization-based algorithm to maximize the likelihood of datasets under the proposed statistical model. The model is tested on datasets involving the outcomes of matches between 20 top male and female tennis players over 14 major tournaments for men (including the Grand Slams and the ATP Masters 1000) and 16 major tournaments for women over the past 10 years. Our model automatically infers that the surface of the court (e.g., clay or hard court) is a key determinant of the performances of male players, but less so for females. Top players on various surfaces over this longitudinal period are also identified in an objective manner.

Keywords: BTL ranking model, Nonnegative matrix factorization, Majorization-minimization, Low-rank approximation, Sports analytics.

1 Introduction

The international rankings for both male and female tennis players are based on a rolling 52-week, cumulative system, where ranking points are earned from players' performances at tournaments. However, due to the limited observation window, such a ranking system is not sufficient if one would like to compare dominant players over a long period (say 10 years) as players peak at different times. The ranking points that players accrue depend only on the stage of the tournaments reached by him or her. Unlike the well-studied Elo rating system for chess [1], one opponent's ranking is not taken into account, i.e., one will not be awarded with bonus points by defeating a top player. Furthermore, the current ranking system does not take into account the players' performances under different conditions (e.g., surface type of courts). We propose a statistical model to ameliorate the above-mentioned shortcomings by (i) understanding the relative ranking of players over a longitudinal period and (ii) discovering the existence of any latent variables that influence players' performances.

The statistical model we propose is an amalgamation of two well-studied models in the ranking and dictionary learning literatures, namely, the *Bradley-Terry-Luce* (BTL) model [2,3] for ranking a population of items (in this case, tennis players) based on pairwise comparisons and *nonnegative matrix factorization* (NMF) [4,5]. The BTL model posits that given a pair of players (i, j) from a population of players $\{1, \dots, N\}$, the probability that the pairwise comparison "i beats j" is true is given by

$$\Pr(i \text{ beats } j) = \frac{\lambda_i}{\lambda_i + \lambda_j}. \tag{1}$$

Thus, $\lambda_i \in \mathbb{R}_+ := [0, \infty)$ can be interpreted as the *skill level* of player i. The row vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}_+^{1\times N}$ thus parametrizes the BTL model. Other

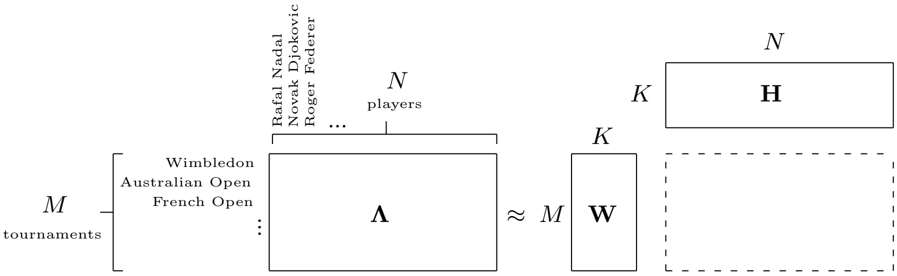


Fig. 1. The BTL-NMF Model

more general ranking models are discussed in [6] but the BTL model suffices as the outcomes of tennis matches are binary.

NMF consists in the following problem. Given a nonnegative matrix $\boldsymbol{\Lambda} \in \mathbb{R}_+^{M\times N}$, one would like to find two matrices $\mathbf{W} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ such that their product $\mathbf{WH}$ serves as a good low-rank approximation to $\boldsymbol{\Lambda}$. NMF is a linear dimensionality reduction technique that has seen a surge in popularity since the seminal papers by Lee and Seung [4, 7]. Due to the non-subtractive nature of the decomposition, constituent parts of objects can be extracted from complicated datasets. The matrix $\mathbf{W}$, known as the *dictionary matrix*, contains in its columns the parts, and the matrix $\mathbf{H}$, known as the *coefficient matrix*, contains in its rows activation coefficients that encode how much of each part is present in the columns of the data matrix $\boldsymbol{\Lambda}$. NMF has also been used successfully to uncover latent variables with specific interpretations in various applications, including audio signal processing [8], text mining analysis [9], and even analyzing soccer players' playing style [10]. We combine this framework with the BTL model to perform a *sports analytics* task on top tennis players.

1.1 Main Contributions

Model: In this paper, we amalgamate the aforementioned models to rank tennis players and uncover latent factors that influence their performances. We propose a hybrid *BTL-NMF* model (see Fig. 1) in which there are M different skill vectors $\boldsymbol{\lambda}_m, m \in \{1, \ldots, M\}$, each representing players' relative skill levels in various tournaments indexed by m. These row vectors are stacked into an $M \times N$ matrix $\boldsymbol{\Lambda}$ which is the given input matrix in an NMF model.

Algorithms and Theory: We develop computationally efficient and numerically stable majorization-minimization (MM)-based algorithms [11] to obtain a decomposition of $\boldsymbol{\Lambda}$ into $\mathbf{W}$ and $\mathbf{H}$ that maximizes the likelihood of the data. Furthermore, by using ideas from [12, 13], we prove that not only is the objective function monotonically non-decreasing along iterations, additionally, every limit point of the sequence of iterates of the dictionary and coefficient matrices is a *stationary point* of the objective function.

Experiments: We collected rich datasets of pairwise outcomes of $N = 20$ top male and female players and $M = 14$ (or $M = 16$) top tournaments over 10 years. Based on these datasets, our algorithm yielded factor matrices $\mathbf{W}$ and $\mathbf{H}$ that allowed us to draw interesting conclusions about the existence of latent variable(s) and relative rankings of dominant players over the past 10 years. In particular, we conclude that male players' performances are influenced, to a large extent, by the surface of the court. In other words, the surface turns out to be the pertinent latent variable for male players. This effect is, however, less pronounced for female players. Interestingly, we are also able to validate via our model, datasets, and algorithm that Nadal is undoubtedly the "King of Clay"; Federer, a precise and accurate server, is dominant on grass (a non-clay surface other than hard court) as evidenced by his winning of Wimbledon on multiple occasions; and Djokovic is a more "balanced" top player regardless of surface. Conditioned on playing on a clay court, the probability that Nadal beats Djokovic is larger than 1/2. Even though the results for the women are less pronounced, our model and longitudinal dataset confirms objectively that S. Williams, Sharapova, and Azarenka (in this order) are consistently the top three players over the past 10 years. Such results (e.g., that Sharapova is so consistent that she is second best over the past 10 years) are not directly deducible from official rankings because these rankings are essentially instantaneous as they are based on a rolling 52-week cumulative system.

1.2 Related Work

Most of the works that incorporate latent factors in statistical ranking models (e.g., the BTL model) make use of mixture models. See, for example, [14–16]. While such models are able to take into account the fact that subpopulations within a large population possess different skill sets, it is difficult to make sense of what the underlying latent variable is. In contrast, by merging the BTL model with the NMF framework—the latter encouraging the extraction of *parts* of complex objects—we are able to observe latent features in the learned dictionary matrix $\mathbf{W}$ (see Table 3) and hence to extract the semantic meaning of latent variables. In our particular application, it is the surface type of the court for male tennis players. See Sec. 4.5 where we also show that our solution is more stable and robust (in a sense to be made precise) than that of the mixture-BTL model.

The paper most closely related to the present one is [17] in which a topic modelling approach was used for ranking. However, unlike our work in which continuous-valued skill levels in $\mathbf{\Lambda}$ are inferred, *permutations* (i.e., discrete objects) and their corresponding mixture weights were learned. We opine that our model and results provide a more *nuanced* and *quantitative* view of the relative skill levels between players under different latent conditions.

1.3 Paper Outline

In Sec. 2, we discuss the problem setup, the statistical model, and its associated likelihood function. In Sec. 3, we derive efficient MM-based algorithms to maximize the likelihood. In Sec. 4, we discuss numerical results of extensive experiments on real-world tennis datasets. We conclude our discussion in Sec. 5.

2 Problem Setup, Statistical Model, and Likelihood

2.1 Problem Definition and Model

Given N players and M tournaments over a fixed number of years (in our case, this is 10), we consider a dataset $\mathcal{D} := \big\{b_{ij}^{(m)} \in \{0,1,2,\ldots\} : (i,j) \in \mathcal{P}_m\big\}_{m=1}^{M}$, where $\mathcal{P}_m$ denotes the set of games between pairs of players that have played at least once in tournament m, and $b_{ij}^{(m)}$ is the number of times that player i has beaten player j in tournament m over the fixed number of year.

To model the skill levels of each player, we consider a nonnegative matrix $\mathbf{\Lambda}$ of dimensions $M \times N$. The $(m,i)^{\text{th}}$ element $[\mathbf{\Lambda}]_{mi}$ represents the skill level of player i in tournament m. Our goal is to design an algorithm to find a factorization of $\mathbf{\Lambda}$ into two nonnegative matrices $\mathbf{W} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ such that the likelihood of $\mathcal{D}$ under the BTL model in (1) is maximized; this is the so-called *maximum likelihood framework*. Here $K \leq \min\{M, N\}$ is a small integer so the factorization is low-rank. In Sec. 3.3, we discuss different strategies to normalize $\mathbf{W}$ and $\mathbf{H}$ so that they are easily interpretable, e.g., as probabilities. Roughly speaking, the eventual interpretation of $\mathbf{W}$ and $\mathbf{H}$ is as follows. Each column of the dictionary matrix $\mathbf{W}$ encodes the "likelihood" that a certain tournament $m \in \{1,\ldots,M\}$ belongs to a certain latent class (e.g., type of surface). Each row of the coefficient matrix encodes the player's skill level in a tournament of a certain latent class. For example, referring to Fig. 1, if the latent classes indeed correspond to surface types, the $(1,1)$ entry of $\mathbf{W}$ could represent the likelihood that Wimbledon is a tournament that is played on clay. The $(1,1)$ entry of $\mathbf{H}$ could represent Nadal's skill level on clay.

2.2 Likelihood of the BTL-NMF Model

According to the BTL model and the notations above, the probability that player i beats player j in tournament m is

$$\Pr(i \text{ beats } j \text{ in tournament } m) = \frac{[\mathbf{\Lambda}]_{mi}}{[\mathbf{\Lambda}]_{mi} + [\mathbf{\Lambda}]_{mj}}.$$

We expect that $\mathbf{\Lambda}$ is close to a low-rank matrix as the number of latent factors governing players' skill levels is small. We would like to exploit the "mutual information" or "correlation" between tournaments of similar characteristics to find a factorization of $\mathbf{\Lambda}$. If $\mathbf{\Lambda}$ were unstructured, we could solve M independent, tournament-specific problems to learn $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M)$. We replace $\mathbf{\Lambda}$ by $\mathbf{WH}$ and the *likelihood* over all games in all tournaments (i.e., of the dataset $\mathcal{D}$), assuming conditional independence across tournaments and games, is

$$p(\mathcal{D}|\mathbf{W},\mathbf{H}) = \prod_{m=1}^{M} \prod_{(i,j)\in\mathcal{P}_m} \left(\frac{[\mathbf{WH}]_{mi}}{[\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj}} \right)^{b_{ij}^{(m)}}.$$

It is often more tractable to minimize the *negative log-likelihood*. In the sequel, we regard this as our objective function which can be expressed as

$$\begin{aligned} f(\mathbf{W},\mathbf{H}) &:= -\log p(\mathcal{D}|\mathbf{W},\mathbf{H}) \\ &= \sum_{m=1}^{M} \sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \Big[-\log\big([\mathbf{WH}]_{mi}\big) + \log\big([\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj}\big)\Big]. \end{aligned} \quad (2)$$

3 Algorithms and Theoretical Guarantees

In this section, we describe the algorithm to optimize (2), together with accompanying theoretical guarantees. We also discuss how we ameliorate numerical problems while maintaining the desirable guarantees of the algorithm.

3.1 Majorization-Minimization (MM) Algorithm

We now describe how we use an MM algorithm [11] to optimize (2). The MM framework iteratively solves the problem of minimizing a certain function $f(x)$, but its utility is most evident when the direct of optimization of $f(x)$ is difficult. One proposes an *auxiliary function* or *majorizer* $u(x, x')$ that satisfies the following two properties: (i) $f(x) = u(x, x), \forall\, x$ and (ii) $f(x) \le u(x, x'), \forall\, x, x'$ (majorization). In addition for a fixed value of x', the minimization of $u(\cdot, x')$ is assumed to be tractable (e.g., there exists a closed-form solution for $x^* = \arg\min_x u(x, x')$). Then, one adopts an iterative approach to find a sequence $\{x^{(l)}\}_{l=1}^{\infty}$. One observes that if $x^{(l+1)} = \arg\min_x u(x, x^{(l)})$ is a minimizer at iteration $l + 1$, then

$$f(x^{(l+1)}) \overset{\text{(ii)}}{\le} u(x^{(l+1)}, x^{(l)}) \le u(x^{(l)}, x^{(l)}) \overset{\text{(i)}}{=} f(x^{(l)}). \quad (3)$$

Hence, if such an auxiliary function $u(x, x')$ can be found, it is guaranteed that the sequence of iterates results in a sequence of non-increasing objective values.

Applying MM to our model is slightly more involved as we are trying to find *two* nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$. Borrowing ideas from using MM in NMFs problems (see for example the works [18,19]), the procedure first updates $\mathbf{W}$ by keeping $\mathbf{H}$ fixed, then updates $\mathbf{H}$ by keeping $\mathbf{W}$ fixed to its previously updated value. We will describe, in the following, how to optimize the original objective in (2) with respect to $\mathbf{W}$ with fixed $\mathbf{H}$ as the other optimization proceeds in an almost[3] symmetric fashion since $\mathbf{\Lambda}^T = \mathbf{H}^T\mathbf{W}^T$. As mentioned above, the MM algorithm requires us to construct an auxiliary function $u_1(\mathbf{W}, \tilde{\mathbf{W}}|\mathbf{H})$ that majorizes $-\log p(\mathcal{D}|\mathbf{W},\mathbf{H})$.

[3] The updates for $\mathbf{W}$ and $\mathbf{H}$ are not completely symmetric because the data is in the form of a 3-way tensor $\{b_{ij}^{(m)}\}$; this is also apparent in the objective in (2) and the updates in (4).

The difficulty in optimizing the original objective function in (2) is twofold. The first concerns the coupling of the two terms $[\mathbf{WH}]_{mi}$ and $[\mathbf{WH}]_{mj}$ inside the logarithm function. We resolve this using a technique introduced by Hunter in [20]. It is known that for any concave function f, its first-order Taylor approximation overestimates it, i.e., $f(y) \le f(x) + \nabla f(x)^T(y-x)$. Since the logarithm function is concave, we have the inequality $\log y \le \log x + \frac{1}{x}(y-x)$ which is an equality when $x=y$. These two properties mean that the following is a majorizer of the term $\log([\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj})$ in (2):

$$\log\big([\mathbf{W}^{(l)}\mathbf{H}]_{mi} + [\mathbf{W}^{(l)}\mathbf{H}]_{mj}\big) + \frac{[\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj}}{[\mathbf{W}^{(l)}\mathbf{H}]_{mi} + [\mathbf{W}^{(l)}\mathbf{H}]_{mj}} - 1.$$

The second difficulty in optimizing (2) concerns the term $\log([\mathbf{WH}]_{mi}) = \log(\sum_k w_{mk}h_{ki})$. By introducing the terms $\gamma_{mki}^{(l)} := w_{mk}^{(l)}h_{ki}/[\mathbf{W}^{(l)}\mathbf{H}]_{mi}$ for $k \in \{1,\dots,K\}$ (which have the property that $\sum_k \gamma_{mki}^{(l)} = 1$) to the sum in $\log(\sum_k w_{mk}h_{ki})$ as was done by Févotte and Idier in [18, Theorem 1], and using the convexity of $-\log x$ and Jensen's inequality, we obtain the following majorizer of the term $-\log([\mathbf{WH}]_{mi})$ in (2):

$$-\sum_k \frac{w_{mk}^{(l)}h_{ki}}{[\mathbf{W}^{(l)}\mathbf{H}]_{mi}} \log\left(\frac{w_{mk}}{w_{mk}^{(l)}}[\mathbf{W}^{(l)}\mathbf{H}]_{mi}\right).$$

The same procedure can be applied to find an auxiliary function $u_2(\mathbf{H}, \tilde{\mathbf{H}}|\mathbf{W})$ for the optimization for $\mathbf{H}$. Minimization of the two auxiliary functions with respect to $\mathbf{W}$ and $\mathbf{H}$ leads to the following MM updates:

$$\tilde{w}_{mk}^{(l+1)} \leftarrow \frac{\displaystyle\sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk}^{(l)}h_{ki}^{(l)}}{[\mathbf{W}^{(l)}\mathbf{H}^{(l)}]_{mi}}}{\displaystyle\sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{h_{ki}^{(l)} + h_{kj}^{(l)}}{[\mathbf{W}^{(l)}\mathbf{H}^{(l)}]_{mi} + [\mathbf{W}^{(l)}\mathbf{H}^{(l)}]_{mj}}}, \tag{4a}$$

$$\tilde{h}_{ki}^{(l+1)} \leftarrow \frac{\displaystyle\sum_m \sum_{j\neq i:(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk}^{(l+1)}h_{ki}^{(l)}}{[\mathbf{W}^{(l+1)}\mathbf{H}^{(l)}]_{mi}}}{\displaystyle\sum_m \sum_{j\neq i:(i,j)\in\mathcal{P}_m} \big(b_{ij}^{(m)} + b_{ji}^{(m)}\big) \frac{w_{mk}^{(l+1)}}{[\mathbf{W}^{(l+1)}\mathbf{H}^{(l)}]_{mi} + [\mathbf{W}^{(l+1)}\mathbf{H}^{(l)}]_{mj}}}. \tag{4b}$$

Note that since we first update $\mathbf{W}$, $\mathbf{H}$ is given and fixed which means that it is indexed by the previous iteration l; as for the update of $\mathbf{H}$, the newly calculated $\mathbf{W}$ at iteration $l+1$ will be used.

3.2 Resolution of Numerical Problems

While the above updates guarantee that the objective function does not decrease, numerical problems may arise in the implementation of (4). Indeed, it is possible that $[\mathbf{WH}]_{mi}$ becomes extremely close to zero for some (m,i). To prevent such numerical problems from arising, our strategy is to add a small number $\epsilon > 0$ to every element of $\mathbf{H}$ in (2). The intuitive explanation that justifies this is that we believe that each player has some default skill level in every type of tournament. By modifying $\mathbf{H}$ to $\mathbf{H} + \epsilon\mathbb{1}$, where $\mathbb{1}$ is the $K \times N$ all-ones matrix, we have the following new objective function:

$$f_\epsilon(\mathbf{W},\mathbf{H}) := \sum_{m=1}^{M} \sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \Big[-\log\big([\mathbf{W}(\mathbf{H}+\epsilon\mathbb{1})]_{mi}\big) + \log\big([\mathbf{W}(\mathbf{H}+\epsilon\mathbb{1})]_{mi} + [\mathbf{W}(\mathbf{H}+\epsilon\mathbb{1})]_{mj}\big)\Big]. \tag{5}$$

Note that $f_0(\mathbf{W},\mathbf{H}) = f(\mathbf{W},\mathbf{H})$, defined in (2). Using the same ideas involving MM to optimize $f(\mathbf{W},\mathbf{H})$ as in Sec. 3.1, we can find new auxiliary functions, denoted similarly as $u_1(\mathbf{W},\tilde{\mathbf{W}}|\mathbf{H})$ and $u_2(\mathbf{H},\tilde{\mathbf{H}}|\mathbf{W})$, leading to following updates

$$\tilde{w}_{mk}^{(l+1)} \leftarrow \frac{\displaystyle\sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk}^{(l)}(h_{ki}^{(l)}+\epsilon)}{[\mathbf{W}^{(l)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}}}{\displaystyle\sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{h_{ki}^{(l)}+h_{kj}^{(l)}+2\epsilon}{[\mathbf{W}^{(l)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}+[\mathbf{W}^{(l)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mj}}}, \tag{6a}$$

$$\tilde{h}_{ki}^{(l+1)} \leftarrow \frac{\displaystyle\sum_{m}\sum_{j\neq i:(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk}^{(l+1)}(h_{ki}^{(l)}+\epsilon)}{[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}}}{\displaystyle\sum_{m}\sum_{j\neq i:(i,j)\in\mathcal{P}_m} \frac{(b_{ij}^{(m)}+b_{ji}^{(m)})w_{mk}^{(l+1)}}{[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}+[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mj}}} - \epsilon. \tag{6b}$$

Notice that although this solution successfully prevents division by zero (or small numbers) during the iterative process, for the new update of $\mathbf{H}$, it is possible $h_{ki}^{(l+1)}$ becomes negative because of the subtraction by ϵ in (6b). To ensure h_{ki} is nonnegative as required by the nonnegativity of NMF, we set

$$\tilde{h}_{ki}^{(l+1)} \leftarrow \max\big\{\tilde{h}_{ki}^{(l+1)}, 0\big\}. \tag{7}$$

After this truncation operation, it is, however, unclear whether the likelihood function is non-decreasing, as we have altered the vanilla MM procedure.

We now prove that f_ϵ in (5) is non-increasing as the iteration count increases. Suppose for the $(l+1)^{\text{st}}$ iteration for $\tilde{\mathbf{H}}^{(l+1)}$, truncation to zero only occurs for the $(k,i)^{\text{th}}$ element and and all other elements stay unchanged, meaning $\tilde{h}_{ki}^{(l+1)}=0$ and $\tilde{h}_{k',i'}^{(l+1)}=\tilde{h}_{k',i'}^{(l)}$ for all $(k',i')\neq(k,i)$. We would like to show that $f_\epsilon(\mathbf{W},\tilde{\mathbf{H}}^{(l+1)}) \leq f_\epsilon(\mathbf{W},\tilde{\mathbf{H}}^{(l)})$. It suffices to show $u_2(\tilde{\mathbf{H}}^{(l+1)},\tilde{\mathbf{H}}^{(l)}|\mathbf{W}) \leq f_\epsilon(\mathbf{W},\tilde{\mathbf{H}}^{(l)})$, because if this is true, we have the following inequality

$$f_\epsilon(\mathbf{W},\tilde{\mathbf{H}}^{(l+1)}) \leq u_2(\tilde{\mathbf{H}}^{(l+1)},\tilde{\mathbf{H}}^{(l)}|\mathbf{W}) \leq f_\epsilon(\mathbf{W},\tilde{\mathbf{H}}^{(l)}), \tag{8}$$

where the first inequality holds as u_2 is an auxiliary function for $\mathbf{H}$. The truncation is invoked only when the update in (6b) becomes negative, i.e., when

$$\frac{\displaystyle\sum_{m}\sum_{j\neq i:(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk}^{(l+1)}(h_{ki}^{(l)}+\epsilon)}{[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}}}{\displaystyle\sum_{m}\sum_{j\neq i:(i,j)\in\mathcal{P}_m} \frac{(b_{ij}^{(m)}+b_{ji}^{(m)})w_{mk}^{(l+1)}}{[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}+[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mj}}} \leq \epsilon.$$

Using this inequality and performing some straightforward but tedious algebra as shown in Sec. S-1 in the supplementary material [21], we can justify the second inequality in (8) as follows

$$\begin{aligned}
&f_\epsilon(\mathbf{W},\tilde{\mathbf{H}}^{(l)}) - u_2(\tilde{\mathbf{H}}^{(l+1)},\tilde{\mathbf{H}}^{(l)}|\mathbf{W}) \\
&\geq \sum_{m}\sum_{j\neq i:(i,j)\in\mathcal{P}_m} \frac{(b_{ij}^{(m)}+b_{ji}^{(m)})w_{mk}}{[\mathbf{W}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}+[\mathbf{W}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mj}} \left[h_{ki}^{(l)} - \epsilon\log\Big(\frac{h_{ki}^{(l)}+\epsilon}{\epsilon}\Big)\right] \geq 0.
\end{aligned}$$

The last inequality follows because $b_{ij}^{(m)}$, $\mathbf{W}$ and $\mathbf{H}^{(l)}$ are nonnegative, and $h_{ki}^{(l)} - \epsilon\log(\frac{h_{ki}^{(l)}+\epsilon}{\epsilon}) \geq 0$ since $x \geq \log(x+1)$ for all $x \geq 0$ with equality at $x=0$. Hence, the likelihood is non-decreasing during the MM update even though we included an additional operation that truncates $\tilde{h}_{ki}^{(l+1)} < 0$ to zero in (7).

3.3 Normalization

It is well-known that NMF is not unique in the general case, and it is characterized by a scale and permutation indeterminacies [5]. For the problem at hand, for the learned $\mathbf{W}$ and $\mathbf{H}$ matrices to be interpretable as "skill levels" with respect to different latent variables, it is imperative we consider *normalizing* them appropriately after every MM iteration in (6). However, there are different ways to normalize the entries in the matrices and one has to ensure that after

Algorithm 1 MM Alg. for BTL-NMF model with column normalization of $\mathbf{W}$

```
**Input:** $M$ tournaments; $N$ players; number of times player $i$ beats player $j$ in tournament $m$ in dataset $\mathcal{D} = \left\{ b_{ij}^{(m)} : i, j \in \{1, ..., N\}, m \in \{1, ..., M\} \right\}$
**Init:** Fix $K \in \mathbb{N}$, $\epsilon > 0$, $\tau > 0$ and initialize $\mathbf{W}^{(0)} \in \mathbb{R}_{++}^{M \times K}, \mathbf{H}^{(0)} \in \mathbb{R}_{++}^{K \times N}$.
**while** diff $\geq \tau > 0$ **do**
(1)   **Update** $\forall m \in \{1, ..., M\}, \forall k \in \{1, ..., K\}, \forall i \in \{1, ..., N\}$
      $\tilde{w}_{mk}^{(l+1)} = \dfrac{\sum\limits_{i,j} b_{ij}^{(m)} \frac{w_{mk}^{(l)}(h_{ki}^{(l)}+\epsilon)}{[\mathbf{W}^{(l)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}}}{\sum\limits_{i,j} b_{ij}^{(m)} \frac{h_{ki}^{(l)}+h_{kj}^{(l)}+2\epsilon}{[\mathbf{W}^{(l)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}+[\mathbf{W}^{(l)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mj}}}$
      $\tilde{h}_{ki}^{(l+1)} = \max\left\{ \dfrac{\sum\limits_{m}\sum\limits_{j \neq i} b_{ij}^{(m)} \frac{w_{mk}^{(l+1)}(h_{ki}^{(l)}+\epsilon)}{[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}}}{\sum\limits_{m}\sum\limits_{j \neq i} \frac{(b_{ij}^{(m)}+b_{ji}^{(m)})w_{mk}^{(l+1)}}{[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mi}+[\mathbf{W}^{(l+1)}(\mathbf{H}^{(l)}+\epsilon\mathbb{1})]_{mj}}} - \epsilon, 0 \right\}$
(2)   **Normalize** $\forall\, m \in \{1, ..., M\}, \forall\, k \in \{1, ..., K\}, \forall\, i \in \{1, ..., N\}$
      $w_{mk}^{(l+1)} \leftarrow \frac{\tilde{w}_{mk}^{(l+1)}}{\sum\limits_{m} \tilde{w}_{mk}^{(l+1)}}$;  $\hat{h}_{ki}^{(l+1)} \leftarrow \tilde{h}_{ki}^{(l+1)} \sum\limits_{m} \tilde{w}_{mk}^{(l+1)} + \epsilon\Big( \sum\limits_{m} \tilde{w}_{mk}^{(l+1)} - 1 \Big)$
      Calculate $\beta = \frac{\sum_{k,i} \hat{h}_{ki}^{(l+1)}+KN\epsilon}{1+KN\epsilon}$, $h_{ki}^{(l+1)} \leftarrow \frac{\hat{h}_{ki}^{(l+1)}+(1-\beta)\epsilon}{\beta}$
(3)   diff $\leftarrow \max\left\{ \max\limits_{m,k} \left| w_{mk}^{(l+1)} - w_{mk}^{(l)} \right|, \max\limits_{k,i} \left| h_{ki}^{(l+1)} - h_{ki}^{(l)} \right| \right\}$
**end while**
**return** $(\mathbf{W}, \mathbf{H})$ that forms a local maximizer of the likelihood $p(\mathcal{D}|\mathbf{W}, \mathbf{H})$
```

normalization, the likelihood of the model stays unchanged. This is tantamount to keeping the ratio $\frac{[\mathbf{W}(\mathbf{H}+\epsilon\mathbb{1})]_{mi}}{[\mathbf{W}(\mathbf{H}+\epsilon\mathbb{1})]_{mi}+[\mathbf{W}(\mathbf{H}+\epsilon\mathbb{1})]_{mj}}$ unchanged for all (m, i, j). The key observations here are twofold: First, concerning $\mathbf{H}$, since terms indexed by (m, i) and (m, j) appear in the denominator but only (m, i) appears in the numerator, we can normalize over all elements of $\mathbf{H}$ to keep this fraction unchanged. Second, concerning $\mathbf{W}$, since only terms indexed by m term appear both in numerator and denominator, we can normalize either rows or columns as we will show in the following.

Row Normalization of W and Global Normalization of H

Define the row sums of $\mathbf{W}$ as $r_m := \sum_k \tilde{w}_{mk}$ and let $\alpha := \frac{\sum_{k,i} \tilde{h}_{ki}+KN\epsilon}{1+KN\epsilon}$. Now consider the following operations:

$$w_{mk} \leftarrow \frac{\tilde{w}_{mk}}{r_m}, \quad \text{and} \quad h_{ki} \leftarrow \frac{\tilde{h}_{ki} + (1-\alpha)\epsilon}{\alpha}.$$

The above update to obtain h_{ki} may result in it being negative; however, the truncation operation in (7) ensures that h_{ki} is eventually nonnegative.[4] See also the update to obtain $\tilde{h}_{ki}^{(l+1)}$ in Algorithm 1. The operations above keep the likelihood unchanged and achieve the desired row normalization of $\mathbf{W}$ since

$$\frac{\sum_k \tilde{w}_{mk}(\tilde{h}_{ki}+\epsilon)}{\sum_k \tilde{w}_{mk}(\tilde{h}_{ki}+\epsilon) + \sum_k \tilde{w}_{mk}(\tilde{h}_{kj}+\epsilon)} = \frac{\sum_k \frac{\tilde{w}_{mk}}{r_m}(\tilde{h}_{ki}+\epsilon)}{\sum_k \frac{\tilde{w}_{mk}}{r_m}(\tilde{h}_{ki}+\epsilon) + \sum_k \frac{\tilde{w}_{mk}}{r_m}(\tilde{h}_{kj}+\epsilon)}$$

$$= \frac{\sum_k w_{mk} \frac{(\tilde{h}_{ki}+\epsilon)}{\alpha}}{\sum_k w_{mk} \frac{(\tilde{h}_{ki}+\epsilon)}{\alpha} + \sum_k w_{mk} \frac{(\tilde{h}_{ki}+\epsilon)}{\alpha}} = \frac{\sum_k w_{mk}(h_{ki}+\epsilon)}{\sum_k w_{mk}(h_{ki}+\epsilon) + \sum_k w_{mk}(h_{kj}+\epsilon)}.$$

Column Normalization of W and Global Normalization of H

Define the column sums of $\mathbf{W}$ as $c_k := \sum_m \tilde{w}_{mk}$ and let $\beta := \frac{\sum_{k,i} \hat{h}_{ki}+KN\epsilon}{1+KN\epsilon}$. Now consider the following operations:

$$w_{mk} \leftarrow \frac{\tilde{w}_{mk}}{c_k}, \quad \hat{h}_{ki} \leftarrow \tilde{h}_{ki}c_k + \epsilon(c_k - 1), \quad \text{and} \quad h_{ki} \leftarrow \frac{\hat{h}_{ki} + (1-\beta)\epsilon}{\beta}.$$

[4] One might be tempted to normalize $\mathbf{H} + \epsilon\mathbb{1} \in \mathbb{R}_{+}^{K \times N}$. This, however, does not resolve numerical issues (analogous to division by zero in (4)) as some entries of $\mathbf{H} + \epsilon\mathbb{1}$ may be zero.

This would keep the likelihood unchanged and achieve the desired column normalization of $\mathbf{W}$ since

$$\frac{\sum_k \tilde{w}_{mk}(\tilde{h}_{ki}+\epsilon)}{\sum_k \tilde{w}_{mk}(\tilde{h}_{ki}+\epsilon)+\sum_k \tilde{w}_{mk}(\tilde{h}_{kj}+\epsilon)} = \frac{\sum_k \frac{\tilde{w}_{mk}}{c_k}(\tilde{h}_{ki}+\epsilon)c_k}{\sum_k \frac{\tilde{w}_{mk}}{c_k}(\tilde{h}_{ki}+\epsilon)c_k+\sum_k \frac{\tilde{w}_{mk}}{c_k}(\tilde{h}_{kj}+\epsilon)c_k}$$

$$= \frac{\sum_k w_{mk}\frac{(\hat{h}_{ki}+\epsilon)}{\beta}}{\sum_k w_{mk}\frac{(\hat{h}_{ki}+\epsilon)}{\beta}+\sum_k w_{mk}\frac{(\hat{h}_{ki}+\epsilon)}{\beta}} = \frac{\sum_k w_{mk}(h_{ki}+\epsilon)}{\sum_k w_{mk}(h_{ki}+\epsilon)+\sum_k w_{mk}(h_{kj}+\epsilon)}.$$

Using this normalization strategy, it is easy to verify that all the entries of $\mathbf{\Lambda} = \mathbf{W}\mathbf{H}$ sum to one.[5] This allows us to interpret the entries of $\mathbf{\Lambda}$ as "conditional probabilities".

3.4 Summary of Algorithm

Algorithm 1 presents pseudo-code for optimizing (5) with columns of $\mathbf{W}$ normalized. The algorithm when the rows of $\mathbf{W}$ are normalized is similar; we replace the normalization step with the procedure outlined in Sec. 3.3. In summary, we have proved that the sequence of iterates $\{(\mathbf{W}^{(l)}, \mathbf{H}^{(l)})\}_{l=1}^{\infty}$ results in the sequence of objective functions $\{f_\epsilon(\mathbf{W}^{(l)}, \mathbf{H}^{(l)})\}_{l=1}^{\infty}$ being non-increasing. Furthermore, if $\epsilon > 0$, numerical problems do not arise and with the normalization as described in Sec. 3.3, the entries in $\mathbf{\Lambda}$ can be interpreted as "conditional probabilities" as we will further illustrate in Sec. 4.3.

3.5 Convergence of Matrices $\{(\mathbf{W}^{(l)}, \mathbf{H}^{(l)})\}_{l=1}^{\infty}$ to Stationary Points

While we have proved that the sequence of objectives $\{f_\epsilon(\mathbf{W}^{(l)}, \mathbf{H}^{(l)})\}_{l=1}^{\infty}$ is non-increasing (and hence it converges because it is bounded), it is not clear as to whether the sequence of *iterates* generated by the algorithm $\{(\mathbf{W}^{(l)}, \mathbf{H}^{(l)})\}_{l=1}^{\infty}$ converges and if so to what. We define the *marginal functions* $f_{1,\epsilon}(\mathbf{W}|\overline{\mathbf{H}}) := f_\epsilon(\mathbf{W}, \overline{\mathbf{H}})$ and $f_{2,\epsilon}(\mathbf{H}|\overline{\mathbf{W}}) := f_\epsilon(\overline{\mathbf{W}}, \mathbf{H})$. For any function $g : \mathcal{D} \to \mathbb{R}$, we let $g'(x; d) := \liminf_{\lambda\downarrow 0}(g(x+\lambda d) - g(x))/\lambda$ be the *directional derivative* of g at point x in direction d. We say that $(\overline{\mathbf{W}}, \overline{\mathbf{H}})$ is a *stationary point* of the minimization problem

$$\min_{\mathbf{W}\in\mathbb{R}_+^{M\times K}, \mathbf{H}\in\mathbb{R}_+^{K\times N}} f_\epsilon(\mathbf{W}, \mathbf{H}) \tag{9}$$

if the following two conditions hold:

$$f'_{1,\epsilon}(\overline{\mathbf{W}}; \mathbf{W} - \overline{\mathbf{W}}|\overline{\mathbf{H}}) \geq 0, \qquad \forall\, \mathbf{W} \in \mathbb{R}_+^{M\times K},$$
$$f'_{2,\epsilon}(\overline{\mathbf{H}}; \mathbf{H} - \overline{\mathbf{H}}|\overline{\mathbf{W}}) \geq 0, \qquad \forall\; \mathbf{H} \in \mathbb{R}_+^{K\times N}.$$

This definition generalizes the usual notion of a stationary point when the function is differentiable and the domain is unconstrained (i.e., $\overline{x}$ is a stationary point if $\nabla f(\overline{x}) = 0$). However, in our NMF setting, the matrices are constrained to be nonnegative, hence the need for this generalized definition.

If the matrices are initialized to some $\mathbf{W}^{(0)}$ and $\mathbf{H}^{(0)}$ that are (strictly) positive and $\epsilon > 0$, then we have the following desirable property.

Theorem 1. *If* $\mathbf{W}$ *and* $\mathbf{H}$ *are initialized to have positive entries (i.e.,* $\mathbf{W}^{(0)} \in \mathbb{R}_{++}^{M\times K} = (0,\infty)^{M\times K}$ *and* $\mathbf{H}^{(0)} \in \mathbb{R}_{++}^{K\times N}$*) and* $\epsilon > 0$*, then every limit point of* $\{(\mathbf{W}^{(l)}, \mathbf{H}^{(l)})\}_{l=1}^{\infty}$ *generated by Algorithm 1 is a stationary point of* (9).

Thus, apart from ensuring that there are no numerical errors, another reason why we incorporate $\epsilon > 0$ in the modified objective function in (5) is because a stronger convergence guarantee can be ensured. The proof of this theorem, provided in Sec. S-2 of [21], follows along the lines of the main result in Zhao and Tan [12], which itself hinges on the convergence analysis of block successive minimization methods provided by Razaviyayn, Hong, and Luo [13]. In essence, we need to verify that $f_{1,\epsilon}$ and $f_{2,\epsilon}$ together with their auxiliary functions u_1 and u_2 satisfy the five regularity conditions in Definition 3 of [12]. However, there are some important differences vis-à-vis [12] (e.g., analysis of the normalization step in Algorithm 1) which we describe in detail in Remark 1 of [21].

[5] We have $\sum_m \sum_i [\mathbf{\Lambda}]_{mi} = \sum_m \sum_i \sum_k w_{mk}h_{ki} = \sum_i \sum_k h_{ki} \sum_m w_{mk} = \sum_{k,i} h_{ki} = 1$.

Table 1. Partial men's dataset for the French Open

Against	R. Nadal	N. Djokovic	R. Federer	A. Murray
R. Nadal	0	5	3	2
N. Djokovic	1	0	1	2
R. Federer	0	1	0	0
A. Murray	0	0	0	0

Table 2. Sparsity of datasets $\{b_{ij}^{(m)}\}$

	Male		**Female**	
Total Entries	$14 \times 20 \times 20 = 5600$		$16 \times 20 \times 20 = 6400$	
	Number	Percentage	Number	Percentage
Non-zero	1024	18.30%	788	12.31%
Zeros on the diagonal	280	5.00%	320	5.00%
Missing data	3478	62.10%	4598	71.84%
True zeros	818	14.60%	694	10.85%

4 Numerical Experiments and Discussion

In this section, we describe how the datasets are collected and provide interesting and insightful interpretations of the numerical results.

4.1 Details on the Datasets Collected

The Association of Tennis Professionals (ATP) is the main governing body for male tennis players. The official ATP website contains records of all matches played on tour. The tournaments of the ATP tour belong to different categories; these include the four Grand Slams, the ATP Masters 1000, etc. The points obtained by the players that ultimately determine their ATP rankings and qualification for entry and seeding in following tournaments depend on the categories of tournaments that they participate or win in. We selected the most important $M = 14$ tournaments for men's dataset, i.e., tournaments that yield the most ranking points which include the four Grand Slams, ATP World Tour Finals and nine ATP Masters 1000, listed in the first column of Table 3. After determining the tournaments, we selected $N = 20$ players. We wish to have as many matches as possible between each pair of players, so that the matrices $\{b_{ij}^{(m)}\}, m \in \{1, \ldots, M\}$ would not be too sparse and the algorithm would thus have more data to learn from. We chose players who both have the highest amount of participation in the $M = 14$ tournaments from 2008 to 2017 and also played the most number of matches played in the same period. These players are listed in the first column of Table 4.

For each tournament m, we collected an $N \times N$ matrix $\{b_{ij}^{(m)}\}$, where $b_{ij}^{(m)}$ denotes the number of times player i beat player j in tournament m. A submatrix consisting of the statistics of matches played between Nadal, Djokovic, Federer, and Murray at the French Open is shown in Table 1. We see that over the 10 years, Nadal beat Djokovic three times and Djokovic beat Nadal once at the French Open.

The governing body for women's tennis is the Women's Tennis Association (WTA) instead of the ATP. As such, we collected data from WTA website. The selection of tournaments and players is similar to that for the men. The tournaments selected include the four Grand Slams, WTA Finals, four WTA Premier Mandatory tournaments, and five Premier 5 tournaments. However, for the first "Premier 5" tournament of the season, the event is either held in Dubai or Doha, and the last tournament was held in Tokyo between 2009 and 2013; this has since been replaced by Wuhan. We decide to treat these two events as four distinct tournaments held in Dubai, Doha, Tokyo and Wuhan. Hence, the number of tournaments chosen for the women is $M = 16$.

After collecting the data, we checked the sparsity level of the dataset $\mathcal{D} = \{b_{ij}^{(m)}\}$. The zeros in $\mathcal{D}$ can be categorized into three different classes.

Table 3. Learned dictionary matrix **W** for the men's dataset

Tournaments	Row Normalization		Column Normalization	
Australian Open	5.77E-01	4.23E-01	1.15E-01	7.66E-02
Indian Wells Masters	6.52E-01	3.48E-01	1.34E-01	6.50E-02
Miami Open	5.27E-01	4.73E-01	4.95E-02	4.02E-02
Monte-Carlo Masters	1.68E-01	8.32E-01	2.24E-02	1.01E-01
Madrid Open	3.02E-01	6.98E-01	6.43E-02	1.34E-01
Italian Open	0.00E-00	1.00E-00	1.82E-104	1.36E-01
French Open	3.44E-01	6.56E-01	8.66E-02	1.50E-01
Wimbledon	6.43E-01	3.57E-01	6.73E-02	3.38E-02
Canadian Open	1.00E-00	0.00E-00	1.28E-01	1.78E-152
Cincinnati Masters	5.23E-01	4.77E-01	1.13E-01	9.36E-02
US Open	5.07E-01	4.93E-01	4.62E-02	4.06E-02
Shanghai Masters	7.16E-01	2.84E-01	1.13E-01	4.07E-02
Paris Masters	1.68E-01	8.32E-01	1.29E-02	5.76E-02
ATP World Tour Finals	5.72E-01	4.28E-01	4.59E-02	3.11E-02

1. (Zeros on the diagonal) By convention, $b_{ii}^{(m)} = 0$ for all (i, m);
2. (Missing data) By convention, if player i and j have never played with each other in tournament m, then $b_{ij}^{(m)} = b_{ij}^{(m)} = 0$;
3. (True zeros) If player i has played with player j in tournament m but lost every such match, then $b_{ij}^{(m)} = 0$ and $b_{ji}^{(m)} > 0$.

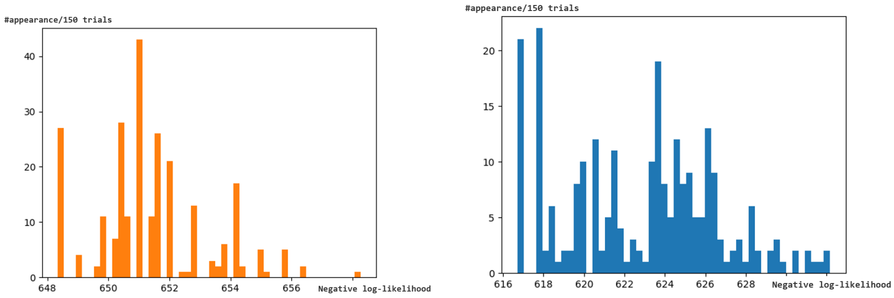
The distributions over the three types of zeros and non-zero entries for male and female players are presented in Table 2. We see that there is more missing data in the women's dataset. This is because there has been a small set of dominant male players (e.g., Nadal, Djokovic, Federer) over the past 10 years but the same is not true for women players. For the women, this means that the matches in the past ten years are played by a more diverse set of players, resulting in the number of matches between the top $N = 20$ players being smaller compared to the top $N = 20$ men, even though we have selected the same number of top players.

4.2 Running of the Algorithm

The number of latent variables is expected to be small and we set K to be 2 or 3. We only present results for $K = 2$ in the main paper; the results corresponding to Tables 3 to 6 for $K = 3$ are displayed in Tables S-1 to S-4 in the supplementary material [21]. We also set $\epsilon = 10^{-300}$ which is close to the smallest positive value in the Python environment. The algorithm terminates when the difference of every element of $\mathbf{W}$ and $\mathbf{H}$ between in the successive iterations is less than $\tau = 10^{-6}$. We checked that the ϵ-modified algorithm in Sec. 3.2 results in non-decreasing likelihoods. See Fig. S-1 in the supplementary material [21]. Since (5) is non-convex, the MM algorithm can be trapped in local minima. Hence, we considered 150 different random initializations for $\mathbf{W}^{(0)}$ and $\mathbf{H}^{(0)}$ and analyzed the result that gave the maximum likelihood among the 150 trials. Histograms of the negative log-likelihoods are shown in Figs. 2(a) and 2(b) for $K = 2$ and $K = 3$ respectively. We observe that the optimal value of the log-likelihood for $K = 3$ is higher than that of $K = 2$ since the former model is richer. We also observe that the $\mathbf{W}$'s and $\mathbf{H}$'s produced over the 150 runs are roughly the same up to permutation of rows and columns, i.e., our solution is *stable* and *robust* (cf. Theorem 1 and Sec. 4.5).

4.3 Results for Men Players

The learned dictionary matrix $\mathbf{W}$ is shown in Table 3. In the "Tournaments" column, those tournaments whose surface types are known to be clay are highlighted in gray. For ease of visualization, higher values are shaded darker. If the rows of $\mathbf{W}$ are normalized, we observe that for clay tournaments, the value in the second column is always larger than that in the first, and vice versa.



(a) $K = 2$

(b) $K = 3$

Fig. 2. Histogram of negative log-likelihood in the 150 trials

Table 4. Learned transpose $\mathbf{H}^T$ of the coefficient matrix for the men's dataset

Players	matrix $\mathbf{H}^T$		Total Matches
Novak Djokovic	1.20E-01	9.98E-02	283
Rafael Nadal	2.48E-02	1.55E-01	241
Roger Federer	1.15E-01	2.34E-02	229
Andy Murray	7.57E-02	8.43E-03	209
Tomas Berdych	0.00E-00	3.02E-02	154
David Ferrer	6.26E-40	3.27E-02	147
Stan Wawrinka	2.93E-55	4.08E-02	141
Jo-Wilfried Tsonga	3.36E-02	2.71E-03	121
Richard Gasquet	5.49E-03	1.41E-02	102
Juan Martin del Potro	2.90E-02	1.43E-02	101
Marin Cilic	2.12E-02	0.00E-00	100
Fernando Verdasco	1.36E-02	8.79E-03	96
Kei Nishikori	7.07E-03	2.54E-02	94
Gilles Simon	1.32E-02	4.59E-03	83
Milos Raonic	1.45E-02	7.25E-03	78
Philipp Kohlschreiber	2.18E-06	5.35E-03	76
John Isner	2.70E-03	1.43E-02	78
Feliciano Lopez	1.43E-02	3.31E-03	75
Gael Monfils	3.86E-21	1.33E-02	70
Nicolas Almagro	6.48E-03	6.33E-06	60

The only exception is the Paris Masters.[6] Since the row sums are equal to 1, we can interpret the values in the first and second columns of a fixed row as the probabilities that a particular tournament is being played on non-clay or clay surface respectively. If the columns of $\mathbf{W}$ are normalized, it is observed that the tournaments with highest value of the second column are exactly the four tournaments played on clay. From $\mathbf{W}$, we learn that surface type—in particular, whether or not a tournament is played on clay—is a germane latent variable that influences the performances of men players.

Table 4 displays the transpose of $\mathbf{H}$ whose elements sum to one. Thus, if the column $k \in \{1, 2\}$ represents the surface type, we can treat h_{ki} as the skill of player i conditioned on him playing on surface type k. We may regard the first and second columns of $\mathbf{H}^T$ as the skill levels of players on non-clay and clay respectively. We observe that Nadal, nicknamed the “King of Clay”, is the best player on clay among the $N = 20$ players, and as an individual, he is also much more skilful on clay compared to non-clay. Djokovic, the first

[6] This may be attributed to its position in the seasonal calendar. The Paris Masters is the last tournament before ATP World Tour Finals. Top players often choose to skip this tournament to prepare for ATP World Tour Finals which is more prestigious. This has led to some surprising results, e.g., David Ferrer, a strong clay player, won the Paris Masters in 2012 (even though the Paris Masters is a hard court indoor tournament).

Table 5. Learned $\mathbf{\Lambda} = \mathbf{WH}$ matrix for first 10 men players

Tournament	Novak Djokovic	Rafael Nadal	Roger Federer	Andy Murray	Tomas Berdych	David Ferrer	Stan Wawrinka	Jo-Wilfried Tsonga	Richard Gasquet	Juan Martin del Potro
Australian Open	2.16E-02	1.54E-02	1.47E-02	9.13E-03	2.47E-03	2.67E-03	3.34E-03	3.97E-03	1.77E-03	4.41E-03
Indian Wells Masters	2.29E-02	1.42E-02	1.68E-02	1.06E-02	2.13E-03	2.30E-03	2.88E-03	4.63E-03	1.72E-03	4.84E-03
Miami Open	2.95E-02	2.30E-02	1.90E-02	1.17E-02	3.80E-03	4.12E-03	5.15E-03	5.07E-03	2.55E-03	5.89E-03
Monte-Carlo Masters	1.19E-02	1.53E-02	4.46E-03	2.27E-03	2.90E-03	3.14E-03	3.92E-03	9.12E-04	1.46E-03	1.94E-03
Madrid Open	1.38E-02	1.51E-02	6.63E-03	3.75E-03	2.75E-03	2.97E-03	3.72E-03	1.57E-03	1.50E-03	2.45E-03
Italian Open	1.19E-02	1.84E-02	2.78E-03	1.00E-03	3.59E-03	3.89E-03	4.87E-03	3.23E-04	1.68E-03	1.71E-03
French Open	1.39E-02	1.43E-02	7.12E-03	4.11E-03	2.57E-03	2.79E-03	3.48E-03	1.74E-03	1.45E-03	2.52E-03
Wimbledon	2.63E-02	1.66E-02	1.91E-02	1.20E-02	2.50E-03	2.71E-03	3.39E-03	5.27E-03	2.00E-03	5.54E-03
Canadian Open	1.16E-02	2.40E-03	1.11E-02	7.32E-03	0.00E+00	1.26E-39	2.42E-51	3.25E-03	5.31E-04	2.81E-03
Cincinnati Masters	1.82E-02	1.43E-02	1.17E-02	7.17E-03	2.36E-03	2.56E-03	3.20E-03	3.10E-03	1.58E-03	3.62E-03
US Open	1.17E-02	9.42E-03	7.38E-03	4.51E-03	1.58E-03	1.71E-03	2.13E-03	1.95E-03	1.03E-03	2.31E-03
Shanghai Masters	8.12E-03	4.38E-03	6.29E-03	4.01E-03	6.09E-04	6.59E-04	8.24E-04	1.76E-03	5.64E-04	1.76E-03
Paris Masters	7.29E-03	9.37E-03	2.73E-03	1.39E-03	1.77E-03	1.92E-03	2.40E-03	5.58E-04	8.94E-04	1.19E-03
ATP World Tour Finals	1.13E-02	8.13E-03	7.63E-03	4.74E-03	1.31E-03	1.41E-03	1.77E-03	2.06E-03	9.29E-04	2.30E-03

Table 6. Learned $\mathbf{\Lambda} = \mathbf{WH}$ matrix for last 10 men players

Tournament	Marin Cilic	Fernando Verdasco	Gilles Simon	Milos Raonic	John Isner	Philipp Kohlschreiber	John Isner	Feliciano Lopez	Gael Monfils	Nicolas Almagro
Australian Open	2.36E-03	2.24E-03	2.87E-03	1.84E-03	2.21E-03	4.38E-04	1.47E-03	1.86E-03	1.09E-03	7.23E-04
Indian Wells Masters	2.79E-03	2.42E-03	2.72E-03	2.06E-03	2.42E-03	3.77E-04	1.37E-03	2.12E-03	9.39E-04	8.56E-04
Miami Open	2.98E-03	3.02E-03	4.20E-03	2.43E-03	2.95E-03	6.75E-04	2.18E-03	2.43E-03	1.68E-03	9.12E-04
Monte-Carlo Masters	4.10E-04	1.11E-03	2.58E-03	6.96E-04	9.77E-04	5.14E-04	1.43E-03	5.95E-04	1.28E-03	1.26E-04
Madrid Open	8.34E-04	1.34E-03	2.59E-03	9.37E-04	1.23E-03	4.87E-04	1.41E-03	8.64E-04	1.21E-03	2.56E-04
Italian Open	0.00E+00	1.05E-03	3.03E-03	5.47E-04	8.63E-04	6.38E-04	1.71E-03	3.95E-04	1.59E-03	7.68E-07
French Open	9.48E-04	1.36E-03	2.49E-03	9.82E-04	1.27E-03	4.57E-04	1.34E-03	9.22E-04	1.14E-03	2.91E-04
Wimbledon	3.17E-03	2.77E-03	3.17E-03	2.36E-03	2.77E-03	4.45E-04	1.59E-03	2.42E-03	1.11E-03	9.72E-04
Canadian Open	2.05E-03	1.32E-03	6.84E-04	1.27E-03	1.40E-03	2.26E-07	2.62E-04	1.38E-03	2.46E-19	6.27E-04
Cincinnati Masters	1.82E-03	1.86E-03	2.60E-03	1.49E-03	1.81E-03	4.20E-04	1.35E-03	1.49E-03	1.04E-03	5.58E-04
US Open	1.14E-03	1.19E-03	1.71E-03	9.49E-04	1.16E-03	2.80E-04	8.94E-04	9.42E-04	6.97E-04	3.49E-04
Shanghai Masters	1.08E-03	8.69E-04	8.72E-04	7.62E-04	8.82E-04	1.08E-04	4.26E-04	7.92E-04	2.69E-04	3.29E-04
Paris Masters	2.51E-04	6.78E-04	1.58E-03	4.26E-04	5.97E-04	3.14E-04	8.72E-04	3.64E-04	7.82E-04	7.73E-05
ATP World Tour Finals	1.22E-03	1.17E-03	1.51E-03	9.61E-04	1.15E-03	2.32E-04	7.76E-04	9.70E-04	5.77E-04	3.75E-04

Table 7. Learned dictionary matrix **W** for the women's dataset

Tournaments	Row Normalization		Column Normalization	
Australian Open	1.00E-00	3.74E-26	1.28E-01	3.58E-23
Qatar Open	6.05E-01	3.95E-01	1.05E-01	4.94E-02
Dubai Tennis Championships	1.00E-00	1.42E-43	9.47E-02	3.96E-39
Indian Wells Open	5.64E-01	4.36E-01	8.12E-02	4.51E-02
Miami Open	5.86E-01	4.14E-01	7.47E-02	3.79E-02
Madrid Open	5.02E-01	4.98E-01	6.02E-02	4.29E-02
Italian Open	3.61E-01	6.39E-01	5.22E-02	6.63E-02
French Open	1.84E-01	8.16E-01	2.85E-02	9.04E-02
Wimbledon	1.86E-01	8.14E-01	3.93E-02	1.24E-01
Canadian Open	4.59E-01	5.41E-01	5.81E-02	4.92E-02
Cincinnati Open	9.70E-132	1.00E-00	5.20E-123	1.36E-01
US Open	6.12E-01	3.88E-01	8.04E-02	3.66E-02
Pan Pacific Open	1.72E-43	1.00E-00	7.82E-33	1.57E-01
Wuhan Open	1.00E-00	6.87E-67	1.41E-01	1.60E-61
China Open	2.26E-01	7.74E-01	4.67E-02	1.15E-01
WTA Finals	1.17E-01	8.83E-01	9.30E-03	5.03E-02

man in the "Open era" to hold all four Grand Slams on three different surfaces (hard court, clay and grass) at the same time (between Wimbledon 2015 to the French Open 2016, also known as the Nole Slam), is more of a balanced top player as his skill levels are high in both columns of $\mathbf{H}^T$. Federer won the most titles on tournaments played on grass and, as expected, his skill level in the first column is indeed much higher than the second. As for Murray, the $\mathbf{H}^T$ matrix also reflects his weakness on clay. Wawrinka, a player who is known to favor clay has skill level in the second column being much higher than that in the first. The last column of Table 4 lists the total number of matches that each player participated in (within our dataset). We verified that the skill levels in $\mathbf{H}^T$ for each player are not strongly correlated to how many matches are being considered in the dataset. Although Berdych has data of more matches compared to Ferrer, his scores are not higher than that of Ferrer. Thus our algorithm and conclusions are not skewed towards the availability of data.

The learned skill matrix $\mathbf{\Lambda} = \mathbf{WH}$ with column normalization of $\mathbf{W}$ is presented in Tables 5 and 6. As mentioned in Sec. 2.1, $[\mathbf{\Lambda}]_{mi}$ denotes the skill level of player i in tournament m. We observe that Nadal's skill levels are higher than Djokovic's only for the French Open, Madrid Open, Monte-Carlo Masters, Paris Masters and Italian Open, which are tournaments played on clay except for the Paris Masters. As for Federer, his skill level is highest for Wimbledon, which happens to be the only tournament on grass; here, it is known that he is the player with the best record in the "Open era". Furthermore, if we consider Wawrinka, the five tournaments in which his skill levels are the highest include the four clay tournaments. These observations again show that our model has learned interesting latent variables from $\mathbf{W}$. It has also learned players' skills on different types of surfaces and tournaments from $\mathbf{H}$ and $\mathbf{\Lambda}$ respectively.

4.4 Results for Women Players

We performed the same experiment for the women players except that we now consider $M = 16$ tournaments. The factor matrices $\mathbf{W}$ and $\mathbf{H}$ (in its transpose form) are presented in Tables 7 and 8 respectively.

It can be seen from $\mathbf{W}$ that, unlike for the men players, the surface type is not a pertinent latent variable since there is no correlation between the values in the columns and the surface type. We suspect that the skill levels of top women players are not as heavily influenced by the surface type compared to the men. However, the tournaments in Table 7 are ordered in chronological order and we notice that there is a slight correlation between the values in the column and the time of the tournament (first half or second half of the year). Any latent variable would naturally be less pronounced, due to the sparser

Table 8. Learned transpose $\mathbf{H}^T$ of coefficient matrix for the women's dataset

Players	matrix $\mathbf{H}^T$		Total Matches
Serena Williams	5.93E-02	1.44E-01	130
Agnieszka Radwanska	2.39E-02	2.15E-02	126
Victoria Azarenka	7.04E-02	1.47E-02	121
Caroline Wozniacki	3.03E-02	2.43E-02	115
Maria Sharapova	8.38E-03	8.05E-02	112
Simona Halep	1.50E-02	3.12E-02	107
Petra Kvitova	2.39E-02	3.42E-02	99
Angelique Kerber	6.81E-03	3.02E-02	96
Samantha Stosur	4.15E-04	3.76E-02	95
Ana Ivanovic	9.55E-03	2.60E-02	85
Jelena Jankovic	1.17E-03	2.14E-02	79
Anastasia Pavlyuchenkova	6.91E-03	1.33E-02	79
Carla Suarez Navarro	3.51E-02	5.19E-06	75
Dominika Cibulkova	2.97E-02	1.04E-02	74
Lucie Safarova	0.00E+00	3.16E-02	69
Elina Svitolina	5.03E-03	1.99E-02	59
Sara Errani	7.99E-04	2.69E-02	58
Karolina Pliskova	9.92E-03	2.36E-02	57
Roberta Vinci	4.14E-02	0.00E+00	53
Marion Bartoli	1.45E-02	1.68E-02	39

dataset for women players (cf. Table 2). A somewhat interesting observation is that the values in $\mathbf{W}$ obtained using the row normalization and the column normalization methods are similar. This indicates that the latent variables, if any, learned by the two methods are the same, which is a reassuring conclusion.

By computing the sums of the skill levels for each female player (i.e., row sums of $\mathbf{H}^T$), we see that S. Williams is the most skilful among the 20 players over the past 10 years. She is followed by Sharapova and Azarenka. As a matter of fact, S. Williams and Azarenka have been year-end number one 4 times and once, respectively, over the period 2008 to 2017. Even though Sharapova was never at the top at the end of any season (she was, however, ranked number one several times, most recently in 2012), she had been consistent over this period such that the model and the longitudinal dataset allow us to conclude that she is ranked second. In fact, she is known for her unusual longevity being at the top of the women's game. She started her tennis career very young and won her first Grand Slam at the age of 17. Finally, the model groups S. Williams, Sharapova, Stosur together, while Azarenka, Navarro, and Vinci are in another group. We believe that there may be some similarities between players who are clustered in the same group. The $\mathbf{\Lambda}$ matrix for women players can be found in Tables S-5 and S-6 in the supplementary material [21].

4.5 Comparison to BTL and mixture-BTL

Finally, we compared our approach to the BTL and mixture-BTL [14, 15] approaches for the male players. To learn these models, we aggregated our dataset $\{b_{ij}^{(m)}\}$ into a single matrix $\{b_{ij} = \sum_m b_{ij}^{(m)}\}$. For the BTL model, we maximized the likelihood to find the optimal parameters. For the mixture-BTL model with $K = 2$ components, we ran an Expectation-Maximization (EM) algorithm [22] to find approximately-optimal values of the parameters and the mixture weights. Note that the BTL model corresponds to a mixture-BTL model with $K = 1$.

The learned skill vectors are shown in Table 9. Since EM is susceptible to being trapped in local optima and is sensitive to initialization, we ran it 100 times and reported the solution with likelihood that is close to the highest one.[7]

[7] The solution with the highest likelihood is shown in Trial 2 of Table S-7 but it appears that the solution there is degenerate.

Table 9. Learned $\boldsymbol{\lambda}$'s for the BTL ($K = 1$) and mixture-BTL ($K = 2$) models

Players	$K = 1$	$K = 2$	
Novak Djokovic	2.14E-01	7.14E-02	1.33E-01
Rafael Nadal	1.79E-01	1.00E-01	4.62E-02
Roger Federer	1.31E-01	1.35E-01	1.33E-02
Andy Murray	7.79E-02	6.82E-02	4.36E-03
Tomas Berdych	3.09E-02	5.26E-02	2.85E-04
David Ferrer	3.72E-02	1.79E-02	4.28E-03
Stan Wawrinka	4.32E-02	2.49E-02	4.10E-03
Jo-Wilfried Tsonga	2.98E-02	3.12E-12	1.08E-01
Richard Gasquet	2.34E-02	1.67E-03	2.97E-03
Juan Martin del Potro	4.75E-02	8.54E-05	4.85E-02
Marin Cilic	1.86E-02	3.37E-05	2.35E-03
Fernando Verdasco	2.24E-02	5.78E-02	8.00E-09
Kei Nishikori	3.43E-02	5.37E-08	3.58E-02
Gilles Simon	1.90E-02	7.65E-05	5.16E-03
Milos Raonic	2.33E-02	2.61E-04	6.07E-03
Philipp Kohlschreiber	7.12E-03	1.78E-25	3.55E-03
John Isner	1.84E-02	2.99E-02	1.75E-08
Feliciano Lopez	1.89E-02	1.35E-02	3.10E-04
Gael Monfils	1.66E-02	5.38E-10	6.53E-03
Nicolas Almagro	7.24E-03	1.27E-15	1.33E-03
Mixture weights	1.00E+00	4.72E-01	5.28E-01
Log-likelihoods	-682.13	-657.56	

The solution for mixture-BTL is not stable; other solutions with likelihoods that are very close to the maximum one result in significantly different parameter values. Two other solutions with similar likelihoods are shown in Table S-7 in the supplementary material [21]. As can be seen, some of the solutions are far from representative of the true skill levels of the players (e.g., in Trial 2 of Table S-7, Tsonga has a very high score in the first column and the skills of the other players are all very small in comparison) and they are vastly different from one another. This is in stark contrast to our BTL-NMF model and algorithm in which Theorem 1 states that the limit of $\{(\mathbf{W}^{(l)}, \mathbf{H}^{(l)})\}_{l=1}^{\infty}$ is a stationary point of (9). We numerically verified that the BTL-NMF solution is stable, i.e., different runs yield $(\mathbf{W}, \mathbf{H})$ pairs that are approximately equal up to permutation of rows and columns.[8] As seen from Table 9, for mixture-BTL, neither tournament-specific information nor semantic meanings of latent variables can be gleaned from the parameter vectors. The results of BTL are reasonable and expected but also lack tournament-specific information.

5 Conclusion and Future Work

We proposed a ranking model combining the BTL model with the NMF framework as in Fig. 1. We derived MM-based algorithms to maximize the likelihood of the data. To ensure numerical stability, we "regularized" the MM algorithm and proved that desirable properties, such as monotonicity of the objective and convergence of the iterates to stationary points, hold. We drew interesting conclusions based on longitudinal datasets for top male and female players. A latent variable in the form of the court surface was also uncovered in a principled manner. We compared our approach to the mixture-BTL approach [14,15] and concluded that the former is advantageous in various aspects (e.g., stability of the solution, interpretability of latent variables).

In the future, we plan to run our algorithm on a larger longitudinal dataset consisting of pairwise comparison data from more years (e.g., the past 50 years) to learn, for example, who is the "best-of-all-time" male or female player. In

[8] Note, however, that stationary points are not necessarily equivalent up to permutation or rescaling.

addition, it would be desirable to understand if there is a natural Bayesian interpretation [19, 23] of the ϵ-modified objective function in (5).

References

1. A. E. Elo. *The Rating of Chess Players, Past and Present.* Ishi Press International, 2008.
2. R. Bradley and M. Terry. Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika*, 35:324–345, 1952.
3. R. Luce. *Individual choice behavior: A theoretical analysis.* Wiley, 1959.
4. D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.
5. A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation.* John Wiley & Sons, Ltd, 2009.
6. J. I. Marden. *Analyzing and modeling rank data.* CRC Press, 1996.
7. D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In *Neural Information Processing Systems*, pages 535–541, 2000.
8. C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis. *Neural Computation*, 21(3):793–830, Mar 2009.
9. M. W. Berry and M. Browne. Email surveillance using non-negative matrix factorization. *Computational and Mathematical Organization Theory*, 11(3):249–264, 2005.
10. A. Geerts, T. Decroos, and J. Davis. Characterizing soccer players' playing style from match event streams. In *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2018 workshop*, pages 115–126, 2018.
11. D. R. Hunter and K. Lange. A tutorial on MM algorithms. *American Statistician*, 58:30–37, 2004.
12. R. Zhao and V. Y. F. Tan. A unified convergence analysis of the multiplicative update algorithm for regularized nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 66(1):129–138, 2018.
13. M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.
14. S. Oh and D. Shah. Learning mixed multinomial logit model from ordinal data. In *Neural Information Processing Systems*, pages 595–603, 2014.
15. N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018.
16. C. Suh, V. Y. F. Tan, and R. Zhao. Adversarial top-K ranking. *IEEE Transactions on Information Theory*, 63(4):2201–2225, 2017.
17. W. Ding, P. Ishwar, and V. Saligrama. A topic modeling approach to ranking. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 214–222, 2015.
18. C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.
19. V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1592–1605, 2013.
20. D. R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.
21. R. Xia, V. Y. F. Tan, L. Filstroff, and C. Févotte. Supplementary material for "A ranking model motivated by nonnegative matrix factorization with applications to tennis tournaments". `https://github.com/XiaRui1996/btl-nmf`, 2019.
22. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1–38), 1977.
23. F. Caron and A. Doucet. Efficient Bayesian inference for generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.