Center for
Mathematical Economics
Working Papers

520

August 2014

Probabilistic Transitivity in Sports

Johannes Tiwisina and Philipp Külpmann

Center for Mathematical Economics (IMW)
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld · Germany

e-mail: imw@uni-bielefeld.de
http://www.imw.uni-bielefeld.de/wp/
ISSN: 0931-6558

Probabilistic Transitivity in Sports[*]

Johannes Tiwisina[†] Philipp Külpmann[‡]

August 28, 2014

Abstract

We seek to find the statistical model that most accurately describes empirically observed results in sports. The idea of a transitive relation concerning the team strengths is implemented by imposing a set of constraints on the outcome probabilities. We theoretically investigate the resulting optimization problem and draw comparisons to similar problems from the existing literature including the linear ordering problem and the isotonic regression problem. Our optimization problem turns out to be very complicated to solve. We propose a branch and bound algorithm for an exact solution and for larger sets of teams a heuristic method for quickly finding a „good“ solution. Finally we apply the described methods to panel data from soccer, American football and tennis and also use our framework to compare the performance of empirically applied ranking schemes.

Keywords: stochastic transitivity, trinomial, geometric optimization, ranking, branch and bound, linear ordering problem, elo, tabu search, football, soccer, tennis, bundesliga, nfl, atp

JEL Classification: L83, C61, C63, C81

1 Introduction

In many situations we are confronted with data about a certain set of objects which only include an array of comparisons about two of these objects at a time. Then all too often the task arises to find the "fairest" or "most legitimate" ranking among all of the objects in the considered set reaching from the "best" one to the "worst" one.

The probably most popular application of such paired comparisons is sports. In most sports games two opponents face each other in a duel. The result can be a win for one of the teams or, depending on the sport, also a tie. Different

[*] We would like to thank Christoph Kuzmics for valuable comments and suggestions. This research was carried out within the International Research Training Group "Economic Behavior and Interaction Models“ (EBIM) financed by the German Research Foundation (DFG) under contract GRK 1134/2.

[†] Center for Mathematical Economics, Bielefeld University, johannes.tiwisina@uni-bielefeld.de.

[‡] Center for Mathematical Economics, Bielefeld University, philipp.kuelpmann@uni-bielefeld.de.

kinds of data sets can arise. An experiment in a league where each team plays each other team a fixed number of times will be called complete. That this property is not naturally given can be seen for instance in American college football, where each team of the 119 teams (in division 1) plays only a small fraction of their competitors.

An important attribute of a ranking is that it expresses a transitive relation between all of its objects. This means that if object or team A precedes B and B precedes C, it automatically implies that A precedes C. In contrast to this, paired comparison data can include circular relations, which seem to be inconsistent with this property. In a tournament it is possible that A beats B, B beats C, but C beats A. It is easy to imagine that as the number of teams rises, the probability of the occurrence of such inconsistencies rapidly increases. Especially in the not very recent literature many suggestions have been made to overcome these inconsistencies and find a ranking with a good fit according to different concepts. A good overview of the classical models for obtaining rankings from data sets gives Brunk [1960b]. One approach that deserves attention is the one proposed by Slater [1961]. Here the observed number of inconsistencies (in the sense mentioned above) is minimized. This nontrivial problem later became known as a particular form of the so called linear ordering problem. For a good survey on the linear ordering problem see for example Charon and Hudry [2010].

The major issue concerning the mentioned approaches is that despite all of them having some intuitive appeal, they seem to be rather arbitrary in finding the "right" ranking. The difference of our approach is that we assume that there actually *exists a correct ranking.* Of course we cannot directly observe it, but we can try to find the ranking which is most likely identical to it. To be more precise, we first of all make the assumption that the outcome of each match follows a trinomial distribution, with a fixed probability for a loss, a tie, and a win. These unobservable probabilities fulfill a certain form of transitivity. Applying the respective conditions we can then use a likelihood function to gauge the chance of the observed set of results given a particular set of probabilities. Maximizing this likelihood function while fulfilling the transitivity conditions answers the question about the most likely *correct* ranking.

In the literature there can be found plenty of works using the concepts of the so called weak and strong stochastic transitivity. These are definitions, which transfer the very intuitive concept of transitivity to the world of probabilities. Because in our model we consider ties and also home/away asymmetries, we are forced to define our own concept which goes beyond WST and SST.

At this point the optimization problem, which is the main object of the paper is completely defined by the set of probabilities for three outcomes for each game, the likelihood function which shall be maximized, and finally the set of constraints imposed by the stochastic transitivity defined above. We are not the first authors trying to find a maximum likelihood ranking while applying probabilistic transitivity conditions. Thompson and Remage [1964] propose a similar problem of ranking pairwisely compared objects. The analysis is extended in Singh and Thompson [1968] by the incorporation of ties. However, Thompson uses only constraints of WST.[1] This contributes a lot to the simplicity of the

[1] After the incorporation of ties he naturally can't use the WST constraints, but has to alter his concept. However, it still differs substantially from ours which makes a comparison

problem and enables Decani [1969] to formulate it as a linear program and later propose in Decani [1972] a branch and bound algorithm to solve the problem even more efficiently.

Unfortunately the new set of constraints make things much more complicated. Increasing the number of teams leads to a huge number of constraints. And it is straightforward to see that the space of transitive probability sets of a particular dimension is not convex. So it is not a surprise that state of the art solvers do not succeed in finding the optimal solution to this non-linear, non-convex problem as soon as the number of teams is increased to more than 5 or 6.

This is why we split up the problem in two parts. The first one is to find the probability sets and the likelihood for a fixed ranking and the second one is to find the ranking with the greatest likelihood.

When the goal is to find probabilities for a fixed ranking, while still sticking to the transitivity definition, the constraints become much simpler.

The problem we arrive at is now very close to the so called isotonic regression problem in which a set of probabilities needs to be estimated, while one knows their order according to their magnitude (see Barlow and Brunk [1972] or Van Eeden [1996] for an overview). A reference much closer to the subject of this paper is Brunk [1955b]. Here the random variables (in our case the match results) are assumed to follow a distribution belonging to a an exponential family. The single distribution parameter follows a function depending monotonically on potentially multiple variables. These variable would in this work correspond to the two teams that are playing. The very efficient method developed in this paper later became known as the pool adjacent-violators algorithm (PAVA). The major difference of Brunk's paper to our approach is that the trinomial distribution we will be using does not belong to the exponential family he is referring to. It also has not one but two distribution parameters. So we are very unfortunate to not being able to apply the PAVA. To be able to estimate not only ordered binomial but also ordered multinomial distribution parameters Jewell and Kalbfleisch [2004] developed a modification of this algorithm, the so called m-PAVA. This algorithm is technically able to solve our first problem, but turns out to be very inefficient and slow. But there is an alternative. Lim et al. [2009] finds that a program of the kind we are facing can be formulated as a geometric program, which then can be transformed into a convex program. By applying state of the art interior point solvers, we are then able to find a solution very efficiently.[2]

The second part of the problem is more complicated. If we increase the number of teams, the possible number of orderings rises very quickly. For 4 teams there are 24 possibilities, for 5 teams there are 120 and for 18 teams there are more than 6×10^{15}. But even if we're not able to find the optimal ranking, we are still able to compare different rankings created by the application of empirically relevant ranking systems. And this is exactly what we do in the empirical section of the paper. Among the candidates are the classical "three points for a win" and "two points for a win" systems from soccer and also the Elo system applied e.g. in chess.

To be able to make a good judgment about the true quality of the systems

very difficult.

[2] In Lim et al. [2009] investigations geometric programming is more than 150 times faster.

when applied to different sports, we develop two kinds of statistical tests. The first one assumes the trueness of the null hypothesis stating that one of two ranking systems under consideration is able to find the correct ordering. Then we estimate all the probabilities and simulate a test statistic. Combined with the empirically observed likelihoods, we are then ideally able to reject the null hypothesis which lets us state that here the considered system is not able to generate the correct ranking.

Our second approach of investigation focuses directly on the probability that a particular empirically applied ranking system is correct, close to correct or not so correct after all. Again, we use the system to order the teams in an empirically observed data set and then estimate the probabilities using the maximum likelihood approach. Then we simulate a very large set of results and calculate a statistic of disagreements also used in the linear ordering problem.

The paper proceeds as follows.

2 Setup

All sports described above have in common that n teams are competing in a number of repeated one-on-one games. The results of these games should be aggregated to one final complete ranking. Let p_{ij} be the probability that team i beats team j.
Naturally, we must have $\forall\, i, j \in \{1, \dots, n\}$

$$\begin{aligned} p_{ij} &\in [0,1] \\ p_{ij} + p_{ji} &\leq 1 \end{aligned} \tag{1}$$

It can be observed that playing at home (meaning in i's stadium) and playing away makes a difference to the wining probabilities. Therefore we introduce different probabilities for at home and away games: p_{ijh} is the probability that i beats j at home and p_{jia} that team j wins against i at i's stadium.
Therefore (1) changes to

$$p_{ijh} + p_{jia} \leq 1 \qquad \forall i, j \in \{1, \dots, n\}$$

Since in many sports, there exists the possibility of a draw, there is no strict equality. In fact, the probability of a draw is

$$q_{ijh} = q_{jia} = 1 - p_{ijh} - p_{jia}.$$

In this paper, we want to make only one assumption concerning a set of those probabilities. This assumption is based on the concept of weak and strong stochastic transitivity, which formalizes the very intuitive thought that if team i is better than team j and j is better than k then i has to be better than k, as well. In a model of symmetric paired comparison without ties this can be translated fairly easily into stochastic terms.

$$p_{ij} \geq 1/2 \quad \wedge \quad p_{jk} \geq 1/2 \implies p_{ik} \geq 1/2 \tag{WST}$$
$$p_{ij} \geq 1/2 \quad \wedge \quad p_{jk} \geq 1/2 \implies p_{ik} \geq \max\{p_{ij}, p_{jk}\} \tag{SST}$$

Where (SST) is equivalent to

$$p_{ij} \geq 1/2 \implies p_{ik} \geq p_{jk}.$$

The concept of stochastic transitivity has been widely used in the literature on paired comparisons, especially in the 60s and 70s (see e.g. Tversky [1969], Chung and Hwang [1978], Morrison [1963] or Davidson and Solomon [1973]).

The introduction of ties and in addition to that the introduction of a home/away asymmetry forbid to use this concept directly. (SST) is best interpreted by saying "if team i is better than team j, it has to have a higher chance of beating any third team k". But in a world with draws and home advantage we cannot interpret "being better" as $p_{ij} > 1/2$. Thats why one has to alter this point. This is done in the following definition.

Definition 1 (Transitivity)**.** A set of probabilities will be called transitive if the following holds for every $i, j, k, l \in \{1, \dots, n\}, x, y \in \{a, h\}$ and $\exists i', j', k', l' \in \{1, \dots, n\}$:

$$\begin{aligned} p_{ikx} \geq p_{jkx} &\Leftrightarrow p_{ily} \geq p_{jly} \\ p_{kix} \geq p_{kjx} &\Leftrightarrow p_{liy} \geq p_{ljy} \\ p_{i'k'x} > p_{j'k'x} &\Rightarrow p_{l'j'x} > p_{l'i'x} \end{aligned} \tag{2}$$

The set of transitive probability sets will be called $\mathcal{T}$.

Definition 2 (Transitive Ranking)**.** A ranking will be called transitive if for all i ranked above j the following holds:

$$p_{ikh} \geq p_{jkh},\ p_{kih} \leq p_{kjh},\ p_{ika} \geq p_{jka},\ p_{kia} \leq p_{kja} \qquad \forall\, k \in \{1, \dots, n\} \backslash i, j$$

The set of probability sets according to this definition will be called $\mathcal{T}'$.

Proposition 1. *Definition 1 is, when assigning* 0 *to all draw probabilities and ignoring away/home differentiation, equivalent to (SST).*

The fact that a transitive ranking has a set of transitive probabilities and every set of transitive probabilities has a transitive ranking is established in the following Proposition.

Proposition 2. *A set of probabilities P is in $\mathcal{T}$ if and only if it is in $\mathcal{T}'$.*

For the proofs of propositions 1 and 2 see appendix A.

The structure of the constraints and hereby the problem we have to solve becomes clearer, if we write down the set of p_{ijx} values in matrix form and add the constraints using one particular ranking.

3 Optimization under a known ranking

Note that the probabilities depicted in figure 1 are only the constraints that apply for one ranking. So the optimization problem can be split into first finding the optimal (i. e., likelihood maximizing) probabilities that satisfy the monotonicity constraints from the matrix and second finding the best ranking. It should become clear that if we consider the indices as variables of the functions $p_h(i, j)$ and $p_a(i, j)$, then this function is monotone non-increasing in the first variable and monotone nondecreasing in the second one. In the considered case the two matrices are only insofar dependent on each other as the sum of an element of the upper right half of the first matrix and the corresponding element of the bottom left half of the second matrix has to be less than or equal to unity.

$$\begin{pmatrix} * & \leq p_{12h} & \leq p_{13h} & \leq \cdots & \leq p_{1nh} \\ \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} \\ p_{21h} \leq & * & \leq p_{23h} & \leq \cdots & \leq p_{2nh} \\ \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} \\ \cdots & \leq \cdots & \leq \cdots & \leq \cdots & \leq \cdots \\ \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} \\ p_{n1h} & \leq p_{n2h} & \leq p_{n3h} & \leq \cdots & \leq * \end{pmatrix}, \begin{pmatrix} * & \leq p_{12a} & \leq p_{13a} & \leq \cdots & \leq p_{1na} \\ \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} \\ p_{21a} \leq & * & \leq p_{23a} & \leq \cdots & \leq p_{2na} \\ \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} \\ \cdots & \leq \cdots & \leq \cdots & \leq \cdots & \leq \cdots \\ \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} & \rotatebox{90}{$\leq$} \\ p_{n1a} & \leq p_{n2a} & \leq p_{n3a} & \leq \cdots & \leq * \end{pmatrix}$$

Figure 1: Transitivity matrices for home and away probabilities

$$\begin{pmatrix} * & 0 & 1 & 0 \\ 0 & * & 1 & 0 \\ 1 & 1 & * & 1 \\ 1 & 0 & 0 & * \end{pmatrix} \qquad \begin{pmatrix} * & \frac{1}{2} & \frac{3}{5} & \frac{3}{5} \\ \frac{1}{2} & * & \frac{3}{5} & \frac{3}{5} \\ \frac{1}{2} & \frac{1}{2} & * & \frac{3}{5} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & * \end{pmatrix}$$

Figure 2: PAVA example: Result matrix and p-Matrix

3.1 Transitivity without draws

Now, let us again compare the original problem to the one in the much simpler case without ties. Here, the problem of estimating the probabilities is much easier. Given the above assumptions, the number of wins when two teams play each other a particular amount of times follows an elementary binomial distribution. This instant allowed Brunk [1955a] to develop an algorithmic approach, building the foundation of what later became known as the Pool Adjacent Violators Algorithm (PAVA). See also Brunk [1960a] for an application to paired comparisons. It follows a short description of the estimation procedure.

A lower interval is the set of all points (i, j) for which $i \geq i'$, $j \leq j'$. So it includes a point in one of the above matrices as well as all the points in its south-west quadrant. An upper interval is analogously defined. A lower layer is a union of lower intervals and an upper layer is a union of upper intervals.

The procedure is now to find the largest upper layer within which the average number of wins is maximized. That is, we have to find an upper layer with the property that the number of wins divided by the number of games it comprises is maximal. For each p_{ij} in this layer the maximum likelihood estimate under the monotonicity constraints we defined is this average number of wins. Next step is to repeat the procedure on the remaining set of the matrix of results.

To illustrate the approach, consider the following example of a tournament of 4 teams in which each two teams played each other once. (Even though it wouldn't be a problem to consider home and away games, for simplicity we just assume that the row team plays at home, here.)

On the left there is the matrix of tournament results. The solid line shows the first upper layer with an average number of wins of 3/5, giving us the p-value listed in the right matrix. The second layer includes all the numbers above and to the right of the dashed line. Here the average value is 1/2 and so on. Having the p-Matrix at hand, it is straightforward to calculate the maximum likelihood of the tournament to be 0.018.

Please note that this algorithm, while being very efficient at finding the probabilities for a fixed ranking, does not help finding the optimal permutation of the teams. To find it, one is still forced to apply this algorithm $4! = 24$ times for this example.

Unfortunately including the chance of draws forbids to use this very simple and efficient procedure. In the next section we show how to arrive at a solution nonetheless.

3.2 Solution process for the case including draws

Again focusing on the part of the problem where the ranking is already fixed, allowing for ties makes the solution procedure much more complicated. Now, the task is not to estimate ordered binomial, but rather ordered trinomial distribution parameters. Jewell and Kalbfleisch [2004] developed an extension of the PAV algorithm discussed above. The Authors call this algorithm the modified- or m-PAV algorithm. In the process the problem is iteratively broken down into many one dimensional optimization problems. Since the number of these subproblems grows very quickly with the number of teams and also the number of adjacent violators, the required computational effort also does. This is the main reason for Lim et al. [2009] to to reconsider the problem finding that it can be formulated as a geometric program. Then it can be transformed into a convex optimization problem, for which one can find a global solution very efficiently with the help of e.g. interior-point algorithms. Lim et al. [2009] compare the computational efficiency of the two approaches and find that geometric programming is much faster than the m-PAV algorithm. These findings facilitate the choice for us in this paper.

Let us take a look at it in detail. Consider the optimization problem for a fixed ranking in its raw form.[3]

$$\begin{aligned}
\min_p \; & \prod_{(i,j)\in E} p_{ijh}^{-w_{ijh}} p_{jia}^{-w_{jia}} (1 - p_{ijh} - p_{jia})^{-(1-w_{ijh}-w_{jia})} \\
s.t. \; & \frac{p_{ijx}}{p_{ikx}} \leq 1 \quad \forall \quad (i,j) \in E, (i,k) \in E, j \succeq k, x \in \{h,a\} \\
& p_{ijh} + p_{jia} \leq 1 \\
& p_{ijx} \geq 0
\end{aligned} \tag{3}$$

This is a geometric program. The objective function as well as the the left side of the first constraint are monomial and the left side of the second constraint are polynomials. The third constraint reflects the fact that the domain of our objective function is positive, as in all geometric programs. The program can easily be transformed to a convex optimization problem.

$$\begin{aligned}
\min_p \; & \sum_{(i,j)\in E} w_{ijh} \ln(p_{ijh}) + w_{jia} \ln(p_{jia}) + (1 - w_{ijh} - w_{jia}) \ln(1 - p_{ijh} - p_{jia}) \\
s.t. \; & \ln(p_{ijx}) - \ln(p_{ikx}) \leq 0 \quad \forall \quad (i,j) \in E, (i,k) \in E, j \succeq k, x \in \{h,a\} \\
& \ln(e^{\ln(p_{ijh})} + e^{\ln(p_{jia})}) \leq 0
\end{aligned}$$

[3] The only change made is the conversion to a minimization instead of a maximization problem.

It is straightforward to show that the logarithm of a posynomial is convex in $\ln(x)$, which proves the fact that this is indeed a convex program. To solve this kind of program we make use of the software package IPOPT (see Wächter and Biegler [2006]). In addition to the program it requires the input of the Jacobian and Hessian matrices of the constraints. It then applies an interior point algorithm and solves our problem very efficiently.

4 Comparing rankings

4.1 Optimal Ranking

By assumption, each outcome in a set of paired comparisons is trinomially distributed. We define w_{ijh} to be the empirically observed number of times team i beats team j at home and $t_{ijh} = t_{jia}$ as the number of times team i ties team j. Let m_{ij} be the total number of games between i and j at i's stadium. The probability distribution is

$$Pr\{x_{ij} = w_{ij}\} = p_{ijh}^{w_{ijh}} p_{jia}^{w_{jia}} (1 - p_{ijh} - p_{jia})^{m_{ij} - w_{ijh} - w_{jia}} \tag{4}$$

where w_{ij} is the vector consisting of the elements w_{ijh} and w_{jia} and x_{ij} is the analogously defined vector of a realization of the corresponding random variable. (4) tells us the probability of a certain outcome of a game between two particular teams in one particular stadium. By taking the exponential of the natural logarithm of the left side, we can write the above equation as

$$\begin{aligned} Pr\{x_{ij} = w_{ij}\} = &\exp(w_{ijh} \ln(p_{ijh}) + w_{jia} \ln(p_{jia}) \\ &+ (m_{ij} - w_{ijh} - w_{jia}) \ln(1 - p_{ijh} - p_{jia})) \end{aligned}$$

In large parts of the empirical section we restrict our selves to the case where $m_{ij} = 1$. But as we will see the other cases are treated analogously.

Let

$$F[x_{ij}, p_{ij}] := w_{ijh} \ln(p_{ijh}) + w_{jia} \ln(p_{jia}) + (1 - w_{ijh} - w_{jia}) \ln(1 - p_{ijh} - p_{jia})$$

The likelihood of a set of particular results to occur will be

$$Pr\{(x_{ij}, \dots, x_{i'j'}) = (w_{ij}, \dots, w_{i'j'})\} = \exp(F[w_{ij}, p_{ij}] + \cdots + F[w_{i'j'}, p_{i'j'}])$$

Let E be the set of all valid (i, j) combinations $E = \{(i, j) | i, j \in \{1, ..., n\}, i \neq j\}$. Then (2) implies that, in order to maximize the likelihood of a set of outcomes, we have to solve the following maximization problem

$$\max_{p_{ij}} J[p] = \sum_{(i,j) \in E} F[w_{ij}, p_{ij}] \qquad s.t. \quad \{p_{ijx} \,|\, (i, j) \in E, x \in \{h, a\}\} \in \mathcal{T}$$

This is a rather complicated optimization problem, first because the objective function (the log of the likelihood function) is not linear, and second because we have a huge number of non-linear constraints, which make the space we are dealing with highly convoluted and non-convex. We can achieve convexity by fixing a particular ranking of teams. In this case we face a total number of $2(2(n-2)n+(n-1))$ constraints. In this case we face a total number of $2(2(n-2)n+(n-1))$ constraints. Note that a simple transformation of parameters cannot help us making the problem convex. Also it cannot make the problem linear after fixing a ranking. In this highly simplified case, where the untransformed constraints can be expressed in a linear form, a logarithmic transformation would make the objective function linear but take away linearity from the constraints. More details on this will follow in section 3.2.

4.2 Ranking methods

4.2.1 The Linear Ordering Problem

At this point before proceeding with our efforts of finding solutions to the proposed problem it makes sense to consider a related, but as we will see, clearly different problem. As one of the classical combinatorial optimization problems the linear ordering problem (LOP) attracted many authors resulting in a huge amount of literature on it. See for example Marti and Reinelt [2011] for a good introduction to the problem as well as a review of suitable algorithms. Also feel referred to Charon and Hudry [2010] for a detailed survey.

If one is given a complete directed graph $D_n = (V_n, A_n)$ with arc weights c_{ij} for every ordered pair $(i,j) \in V_n \times V_n$, the linear ordering problem consists of finding an acyclic tournament T (which corresponds to a permutation of the set of objects or teams), which maximizes the sum of the arcs which are in agreement with the direction of the arcs from D_n. So the sum $\sum_{(i,j)\in T} c_{ij}$ has to be maximal. Equivalently one could formulate the problem as minimizing the so called remoteness corresponding to minimizing the arc weights pointing in the opposite direction.

A more illustrative representation of the problem is the maximization of the sum of superdiagonal elements in a matrix by manipulating the row/column ordering. This is the so called Triangulation problem.

The reader might already be able to grasp a sense of similarity here. To establish a direct connection between the LOP and the problem dealt with in this paper, consider a situation where we fix the probabilities of wins and losses at homogeneous values below and above the diagonal of the matrix independently of which teams are in question. This means we set $p_{ijh} = \overline{p}_h$ above diagonal and $p_{ijh} = \underline{p}_h$ below it and analogously for the away probabilities. Let us consider the case where $\overline{p}_h > \underline{p}_h$ and $\overline{p}_a > \underline{p}_a$. Remember that the goal is to maximize

$$\begin{aligned}
&\sum_{(ij)\in E} w_{ijh}\ln(p_{ijh}) + w_{jia}\ln(p_{jia}) + (1 - w_{ijh} - w_{jia})\ln(1 - p_{ijh} - p_{jia}) \\
&= \sum_{(ij)\in \overline{E}} w_{ijh}\ln(\overline{p}_h) + w_{ija}\ln(\overline{p}_a) + t_{ijh}\ln(1 - \overline{p}_h - \underline{p}_a) \\
&\quad + \sum_{(ij)\in \underline{E}} w_{ijh}\ln(\underline{p}_h) + w_{ija}\ln(\underline{p}_a) + t_{ijh}\ln(1 - \underline{p}_h - \overline{p}_a)
\end{aligned}$$

where $\overline{E}$ and $\underline{E}$ represent the sets of elements above and below the diagonals, respectively.

The results of a particular team in his two games against a particular opponent makes a certain contribution to the sum. This contribution might be higher because it is multiplied by higher probabilities if the records are super-diagonal. So we are confronted with a triangulation problem just like the one described above. Many Authors suggest an application of the LOP in sports rankings (see e.g. Marti and Reinelt [2011]). And since it indeed seems well suited for our purposes, we will include it in the analysis.

4.2.2 Branch and Bound Algorithm

Branch and Bound Algorithms are particularly well suited for combinatorial optimization problems. As opposed to the other methods we are proposing, this one leads with certainty to the optimal ranking. For an early survey on Branch and Bound methods feel referred to Lawler and Wood [1966].

The following steps describe the execution of the algorithm:

1. Take the next team from the list of all teams
2. Put it in the list of previously selected teams at each possible position
3. For each position calculate an upper bound $\overline{L}$ above which the likelihood cannot rise going further down the tree (i.e. after all teams were inserted)
4. Leave the team at the position with the highest upper bound
5. If all teams are inserted go to 6., otherwise go to 1.
6. Compare the likelihood to the best one found so far
7. Cut of the tree at all nodes where $\overline{L}$ is below the best likelihood
8. Go to the best of the lowest hanging nodes that could not be deleted and start with 1. from there

Before asking how the upper bound estimate $\overline{L}$ is calculated, lets first focus on the procedure itself. To understand it better, consider a simple example of three teams "a" "b" and "c".

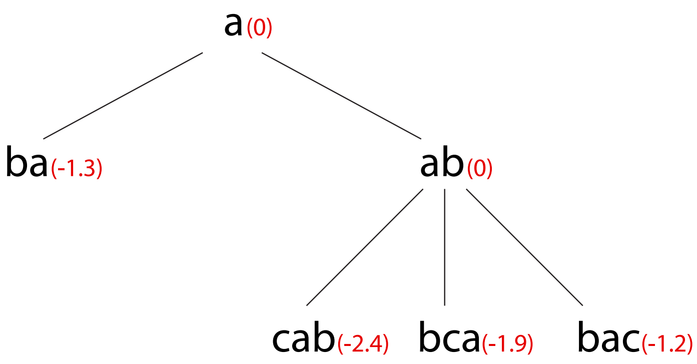


Figure 3: Branch and Bound Algorithm: An example

We start by inserting team "a". The upper bound for the log likelihood at this point is still 0, which is indicated in brackets in Figure 3. Then team "b" is added at each possible position. In Figure 1 we see upper bounds of -1.3 and 0, respectively. So we continue by leaving "b" at the second position and then insert team c at each possible location. Since the example only includes three teams, we can now calculate the value of the real objective function instead of calculating $\overline{L}$ the way it was done previously. The highest value of the objective function is found using the ordering "bac". This value of -1.2 now enables us to cut of all hanging nodes, which have an upper bound below -1.2. So we cut of the tree at "ba", since there is no way, we could get a better likelihood going down the tree from this node. It is easy to see how the procedure can save computational effort (even in this tiny example) compared to calculating the MLE for all permutations.

The upper bound $\overline{L}$ is calculated as follows. First the optimization problem (for a fixed ranking) is applied to the teams that have been inserted so far.

Lemma:. *Adding an additional team into an existing ranking without changing the relative order of the already existing teams can not increase $\overline{L}$.*

Proof. It is trivial to see that adding a variable (team) to the maximization problem without adding additional constraints (results) does not change the maximum likelihood (i.e., we are multiplying by 1). Now, adding additional constraints without changing the objective function or changing the other constraints can never increase the maximum likelihood and therefore the new $\overline{L}$ has to be less or equal to the $\overline{L}$ with 1 team less. □

At this stage we could already use this maximum likelihood of the considered subset of teams for $\overline{L}$. But there is a way to reduce the upper bound even further and thereby make the algorithm a lot more efficient. For each team that is still pending to be inserted we already know a subset of the constraints that will be applied to the corresponding probabilities when going further down the tree, no matter where this particular team will be inserted. Consider a situation where teams $1, ..., k$ have already been inserted. Now, for each team $l \in \{k+1, ..., n\}$ we know that $p_{ilx} \leq p_{i'lx}$ and $p_{lix} \geq p_{i'lx}$ for every $i, i' \in \{1, ..., k\}$ and $x \in \{h, a\}$ such that i is ranked above i'. For $k = 3$ this is depicted in Figure 4.

$$\begin{pmatrix} * & p_{12h} & p_{13h} \\ p_{21h} & * & p_{23h} \\ p_{31h} & p_{32h} & * \end{pmatrix} \quad \begin{array}{c} p_{1lh} \\ \rotatebox{90}{$\leq$} \\ p_{2lh} \\ \rotatebox{90}{$\leq$} \\ p_{3lh} \end{array} \qquad\qquad \begin{pmatrix} * & p_{12a} & p_{13a} \\ p_{21a} & * & p_{23a} \\ p_{31a} & p_{32a} & * \end{pmatrix} \quad \begin{array}{c} p_{1la} \\ \rotatebox{90}{$\leq$} \\ p_{2la} \\ \rotatebox{90}{$\leq$} \\ p_{3la} \end{array}$$

$$p_{l1h} \leq p_{l2h} \leq p_{l3h} \qquad\qquad\qquad p_{l1a} \leq p_{l2a} \leq p_{l3a}$$

Figure 4: Calculation of the upper bound $\overline{L}$

For every team that has not been inserted yet, we know this subset of constraints. So we have another optimization problem for each team. The results of these optimization problems (, having the form of a log likelihood values) can be added to the value $\overline{L}$.

As mentioned, the algorithm leads for sure to the optimal ordering. The drawback is that despite of the fairly sophisticated upper bound that we are suggesting, it is still not efficient enough to be applied to tournaments with more than 9-10 teams[4]. Nevertheless the branch and bound algorithms deserves to

[4] Depending on the structure of results in the tournament as well as the users patience.

be included in the empirical section of this paper.

4.2.3 Tabu Search

The third ranking algorithm we are suggesting is a heuristic search method. The advantage of tabu search lies in the combination of local search and a diversification mechanism. The local search systematically browses through neighborhood solutions, checking for a possible improvement of the objective value. That the algorithm doesn't get stuck in local optima is assured by a memory structure, avoiding previously visited regions of the solution space, giving a tendency for diversification. A reference with a somewhat related application is Laguna et al. [1999].

The algorithm works as follows:

1. Start from a randomly generated order of teams (call it ρ)
2. Calculate the maximum likelihood for the current ranking $L(\rho)$
3. Randomly select a team that is not on the "Tabu List" and remove it from the order
4. Insert the team at position i and calculate difference between the maximum likelihood of the new and the original ranking: $MoveValue = L(\rho') - L(\rho)$
5. Repeat 4. for $1 \leq i \leq n$ except for the original position
6. Insert the team at the position with the highest $MoveValue$
7. Put the team on the "Tabu List" so that it won't be selected for the next "TabuTenure" iterations
8. Go to 2.

Basically what the algorithm does is taking a team from the ranking and trying out every possible position for it, except for the original one. Important is that the best among the new positions is selected even if the "$MoveValue$" is negative. Different convergence criteria are possible for the procedure. Since in our analysis the computational effort in each iteration is fairly large, we use a fixed number of iterations for the algorithm, so that we can best control the amount of time it takes for the algorithm to finish.

4.2.4 Popular Ranking Methods

Finally we want to take a more practical approach and compare different ranking systems, which have been used in different fields of sports. We chose the 3 point system (also known as "Three points for a win"), which awards zero points for a loss, one point for a draw and three points for a win. The sum of the points together with the goal difference as a tie breaker then decides upon the ranking. This system has been used in most soccer leagues since it was officially adopted in 1995 by FIFA.[5] Before the 3 point system was introduced, the analogously structured 2 point system had been widely used in soccer. Here the only difference is that two instead of three points are awarded for a win.

[5] England introduced the system already in 1981. The first time it was used internationally was in the 1994 World Cup finals.

These two systems are fairly easy to apply and (unfortunately) also very similar to each other. So as a third candidate for a ranking scheme, we use the Elo rating system. The Elo Rating system is a system invented by Elo [1978] originally intended as a rating system for chess. Today it is not only used as for different chess organizations, including the FIDE and the United States Chess Federation, but also the European Go Federation, many different computer games and even the National Collegiate Athletic Association, the organization which is responsible for the organization of many American college sport programs, notable college football and college basketball.
The main differences between the three points for a win is that it factors in the strength of the opponent: winning against a strong opponent yields more points than winning against a weak one. This results in the mayor weakness for our needs: a relatively high number of games is needed to give meaningful results and the order in which the teams play matters a lot.

4.3 Comparing the explanatory power of these rankings

To further enhance our comparative analysis of ranking systems, we will apply a statistical hypothesis test. In this test two ranking systems are compared, call them system a and system b. We solve problem (3) for both rankings. The p-matrix calculated with the constraints generated by one of the rankings, say system a, will yield a likelihood for the observed season at least as great as the one generated by the other one, say system b.

$$L(\hat{P}_a(w)|w) \geq L(\hat{P}_b(w)|w)$$

where $\hat{P}_a(w)$ and $\hat{P}_b(w)$ are the estimated p-matrices. So we could say, a allows one to calculate a p-matrix with a higher explanatory value, so it must be the better system. But in fact, it might have happened by chance, that this ranking system performed better than the other one. The central question concerns the degree of the odds that a performed better than b by the observed amount. Let us define the likelihood ratio as follows

$$LR_{a,b} = log(L(\hat{P}_a(w)|w))) - log(L(\hat{P}_b(w)|w)).$$

We assume a Hypothesis H_0 stating that "b is the correct ranking system". Correct means that it allows us to estimate the right p-values. Using these probabilities for each match, we simulate a complete season and get a new tournament $\hat{w}$ for which we again calculate the likelihoods given $\hat{P}_a(w)$ and $\hat{P}_b(w)$. This way a few thousand seasons are simulated and we receive a distribution over the difference of the log-likelihood. In the ideal case, the probability (suggested by the simulated distribution) of the observed difference between the likelihoods is small enough to be able to reject H_0 with this very test size α.

$$P[LR_{a,b}(\hat{w}) \leq LR_{a,b}(w)] < \alpha$$

So, roughly what we do is assuming that one of the systems is correct, and then we try to reject this hypothesis, by showing that the probability for another system to be as much better as empirically observed is very small.

The weakness of this approach is pretty obvious. We are only able to reject the hypothesis that a particular system is perfectly correct. Even though the data allows us to make a guess about it, the test does not allow us to make a statement about which of the two systems in consideration is actually better. So in fact, both of the systems might be incorrect, but we are only able to reveal the inadequacy of one of them.

5 Data

We obtain the data from different sources. For soccer we focus on the German Bundesliga and the British Premier League. For the former we have data from the seasons 1968/69 till 2012/13, for the latter the sample from the seasons 1997/98 till 2012/2013. The scores for all matches, which are translated to win/draw/loss data, are obtained from the website www.kicker.de. Notable about the soccer data is that each team plays each other team exactly once at home and once away in each season. This introduces a symmetry to the data which, even though not it is not necessary, might be considered as desirable and certainly influences the results of our analysis.

Regarding tennis, we face a different situation. Since there is no league of players in which each player faces another one a fixed number of timer per season we have to go a different way. We will focus only on the top 10 players according to the official ATP ranking at the end of each year (obtained from http://www.atpworldtour.com). Then we collect the data for all the ATP matches played in this season from http://www.tennis-data.co.uk. Of course these data sets will be highly asymmetric, because some players play each other more that once, and some might not play each other at all during a season. Another special fact about the tennis data is that we don't have a real home away situation.[6] Even more importantly, in tennis there is no possibility of ties. So we face only a binomial distribution for the outcome of each match which considerably facilitates the optimization procedure.

Concerning American football, we will focus exclusively on the NFL. We have data on the scores of every NFL game since 1978 from the website http://www.repole.com/sun4cast/data.html. The NFL comprises from 28 (in the season 1978) to 32 teams in 2012. This is by far the largest group of teams. Almost naturally it follows that among the samples there is a huge number of teams that don't face each other during a season. Which team is playing which is determined by a complicated system, which shall not be further discussed here. In football draws are possible, but only happen very rarely. Along with the fact that American football enjoys great popularity, this makes NFL data very interesting for our analysis.

6 Empirical Analysis

We now want to apply the presented methods to real data from sports. Countless different types of sports are imaginable and probably the readers preferences for what he would like to see in this section are very heterogeneous. Nevertheless for reasons of space we want to focus on three types, namely soccer, tennis

[6] Of course some players might feel more at home when a tournament is taking place in their country of origin. But since this is very different to the situation of a team playing in his very own stadium in its city, we will assume that every game takes place on neutral ground.

and American football. The main questions we seek to answer are, "Is there a tendency for one of the ranking schemes to be superior to the others according to the criterion we defined?", "If yes, which one is it?", "Does it depend on the type of sport?" and finally "Are we able to improve on the rankings found by the simple ranking methods using one of the algorithms presented in section 4.2?"

6.1 Soccer in Austria: Finding an Optimal Ranking

With the branch and bound algorithm we find our selves equipped with a very powerful instrument to find the optimal ranking. Unfortunately this algorithm can only be applied to sets of teams that have a limited size. The first object of our investigation shall be the Austrian Bundesliga. Its size of 10 teams enables us to apply the discussed bnb-method. During a season each team plays each other team four times, two times at home and two times away. This is different from most other soccer leagues, but doesn't increase the computational complexity by much. Here, we consider the season 2012/2013. To draw a first comparison between the performances of the other ranking schemes, Table 1 shows the maximum likelihoods that have been calculated.

Method	BnB	2-Point	3-Point	LOP	Elo	Tabu-Search
MLE	-129.844	-131.742	-135.561	-140.024	-131.703	-130.465

Table 1: Log likelihood values for the Austrian Bundesliga 12/13

While the ranking corresponding to the solution to the linear ordering problem gives a relatively low likelihood, the two point system as well as the Elo-system seem to explain the results a lot better. Nevertheless, none of systems generates the optimal ranking found by the branch and bound algorithm. The ranking produced by the Tabu Search gives a higher likelihood than all the systems, but still is not the optimal one.

Figure 5 compares the optimal ranking that we found with the actually applied order, namely the 3-Point ranking. One can see that there are indeed some differences. Perhaps most striking is that in this season SV Mattersburg was relegated, while in the optimal ranking Wacker Insbruck would have been relegated. This team was actually ranked 8th.

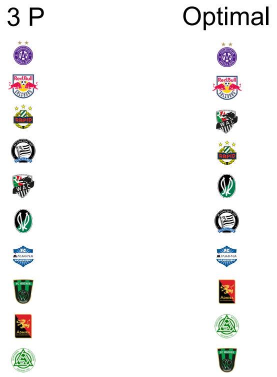


Figure 5: Rankings resulting from 3-point system and Branch and Bound algorithm

Unfortunately most leagues are larger that the Austrian Bundesliga. The resulting computational effort makes it virtually impossible for us to find optimal rankings. which is why in the next section we focus on the other methods and compare the different ranking schemes across panel data from different leagues in different sports.

6.2 Ranking Systems and Maximum Likelihood Estimates

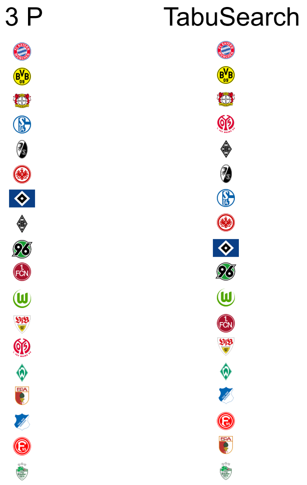
Before the values of the probability matrices can be estimated, the teams need to be put into an order. To give the reader an impression of how such a matrix of outcome probabilities for each game looks like after the optimization, Figure 6 depicts the probabilities for home game wins for the Bundesliga season 2012/13 estimated using the "three points for a win" system.

0.18	0.31	0.69	0.8	0.8	0.8	0.8	0.8	1	1	1	1	1	1	1	1	1	1
0	0.31	0.69	0.69	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.92	0.92	0.92	0.92	0.92	1
0	0.31	0.46	0.69	0.69	0.69	0.69	0.69	0.71	0.71	0.71	0.71	0.71	0.92	0.92	0.92	0.92	1
0	0.31	0.31	0.46	0.46	0.46	0.52	0.52	0.71	0.71	0.71	0.71	0.71	0.92	0.92	0.92	0.92	0.95
0	0.31	0.31	0.46	0.46	0.46	0.46	0.52	0.7	0.7	0.7	0.7	0.7	0.86	0.92	0.92	0.92	0.95
0	0.31	0.31	0.46	0.46	0.46	0.46	0.49	0.49	0.49	0.49	0.61	0.7	0.85	0.85	0.85	0.85	0.85
0	0.31	0.31	0.46	0.46	0.46	0.46	0.49	0.49	0.49	0.49	0.57	0.7	0.85	0.85	0.85	0.85	0.85
0	0	0.31	0.31	0.31	0.46	0.46	0.48	0.49	0.49	0.49	0.53	0.58	0.58	0.85	0.85	0.85	0.85
0	0	0.31	0.31	0.31	0.31	0.46	0.47	0.47	0.49	0.49	0.49	0.49	0.58	0.85	0.85	0.85	0.85
0	0	0.31	0.31	0.31	0.31	0.31	0.41	0.41	0.45	0.49	0.49	0.49	0.49	0.49	0.85	0.85	0.85
0	0	0.31	0.31	0.31	0.31	0.31	0.41	0.41	0.41	0.43	0.44	0.44	0.44	0.44	0.44	0.44	0.44
0	0	0.19	0.31	0.31	0.31	0.31	0.41	0.41	0.41	0.41	0.42	0.42	0.43	0.44	0.44	0.44	0.44
0	0	0.19	0.19	0.19	0.19	0.31	0.38	0.41	0.41	0.41	0.42	0.42	0.42	0.44	0.44	0.44	0.44
0	0	0.075	0.19	0.19	0.19	0.31	0.36	0.36	0.36	0.36	0.36	0.42	0.42	0.42	0.42	0.44	0.44
0	0	0.056	0.19	0.19	0.19	0.19	0.19	0.29	0.29	0.29	0.36	0.36	0.42	0.42	0.42	0.44	0.44
0	0	0.037	0.19	0.19	0.19	0.19	0.19	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.36	0.44	0.44
0	0	0	0	0	0.19	0.19	0.19	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.35	0.44
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.094	0.17	0.3

Figure 6: MLE for p_{ijh} using 3-point system

Generally, a striking feature about the structure of the estimated probably matrices is the occurrence of homogeneous values in certain areas of the matrix, reminding of the layer structure discussed in section 2.3. Remarkable in this particular matrix is the large number of "1"s in the upper right corner and "0"s in the lower left corner. The reader might be tempted to argue that these values are fairly unrealistic, because intuition tells us that even if the strongest team plays the weakest one, in the current case Bayern München against Greuter Fürth, the chance of the former to win against the latter will be high, but never 100%. But the point is that we only hold this intuition, because probably at some point in the past we have seen top teams occasionally loosing against teams that were ranked very low. But since this kind of information is not part of our estimation procedure, it is only natural that estimates look like this.[7]

Next, we want to try to improve this ranking by using one of the algorithms presented in section 2.5. Unfortunately the sample of 18 teams is to large for an application of the branch and bound algorithm, which would technically allow us to find the optimal ranking. So we use the Tabu search method, which we run for 100 iterations. The resulting ordering as well as the corresponding maximum likelihoods are shown in Figure 7.



$ln(L(\hat{P}_{3P}|W)) = -237.515$ $ln(L(\hat{P}_{Tabu}|W)) = -229.408$

Figure 7: Rankings resulting from 3-point system and Tabu Search

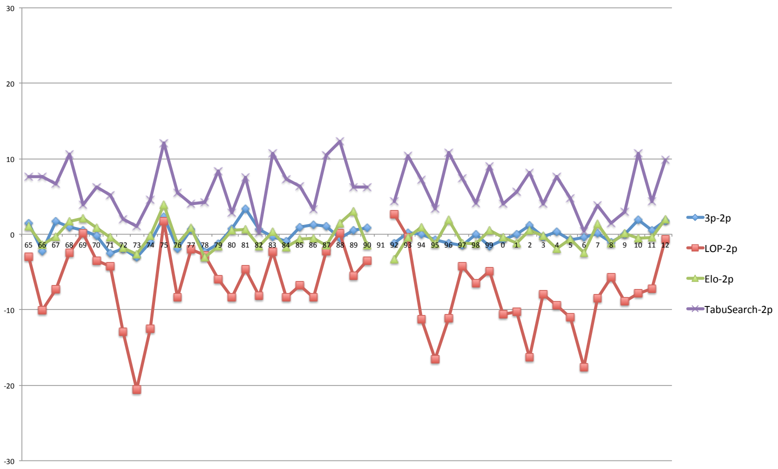
The Tabu Search finds a ranking that is partly very different from the one determined using the 3-point system. The biggest difference is the position of "Mainz 05" jumping from the 13th position to the 4th. This difference can only be that "Mainz 05" has won the matches in this season that were particularly important in the sense of being in accordance with the team having fairly high winning probabilities in general. However, despite of differences in parts, a great similarity between the rankings can be observed. This similarity can be measured using Spearman's rank correlation coefficient defined as $\rho = 1 - \frac{6\sum(r_i - s_i)^2}{n(n^2-1)}$[8] with r_i being the original (3 point) ranking of team i and the ranking and s_i the ranking with the highest maximum likelihood has calculated with the Tabu Search algorithm. The correlation between the 3-point ranking and the one found by the Tabu Search is indeed fairly high with a value of about

[7] We have to add, that in case he has seen Bayern München play in the season 2012/13, he most certainly would agree that estimating some probabilities in the right of the upmost row with a value of 1 most probably only involves a very small error.

[8] Note that, as we do allow for ties in the results, we do not have ties in the rankings.

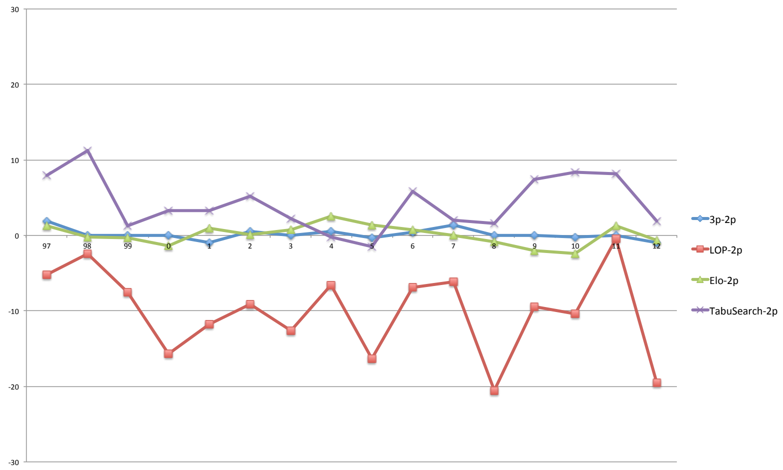
0.87616. The difference between the maximum likelihood values however, is in fact very large. The probabilities found using the Tabu Search ranking make the observed season 3318 more likely compared to the probabilities found using the 3 point ranking.

As mentioned above, we have data not only on this one Bundesliga season, but on the ones from the last 50 years.[9] For every season that we have data on, we calculated the maximum likelihood p-matrices as well as the objective function values using the "2 points for a win", the "3 points for a win", Elo system and the ranking from the solution to the linear ordering problem. Finally we used the Tabu search method to find out, whether or not one is able to improve on one or all of the ranking schemes. Because from season to season the likelihood values fluctuate heavily, it makes sense to use the likelihood found by one of the systems as a reference value and plot the differences to these values in a diagram. As opposed to just plotting the absolute likelihoods of every system in each year, this technique allows us to better compare the quality of the rankings throughout the panel data. The system of reference will be "2 points for a win".



(a) Bundesliga

[9] Because in the seasons 1963/64, 1964/65 and 1991/92 the number of teams in the Bundesliga was different from 18, we excluded these seasons from the sample. Sacrificing these three data points for a higher comparability seems reasonable.



(b) Premier League

Figure 7: Maximum Likelihoods for Bundesliga and Premier League panel data

Figure 7 (a) and (b) reveal that the two and three point systems are in fact very close in the maximum likelihoods they "produce". This is not least because in most cases the rankings determined by the two systems only differ in a few spots. And if the rankings do not differ much, it's only natural that the likelihood values won't be very far apart either. The two point system allows for a calculation of p-matrices that make the observed seasons on average across the Bundesliga samples by about 9.8% more likely than when using three point system. In the Premier League the three point system has a 5.2% higher explanatory power. The Elo-system also gives us likelihoods in the same range, indicated by the green lines. Actually this is a bit of a surprise, since there were some hopes that the intuitively very reasonable mechanism of getting more points for winning against relatively strong opponents would enable us to explain the observed results better. Still it is not worse than the conventional two and three point systems. But because of its higher complexity we clearly refrain from making a recommendation for using this system in soccer. The ranking resulting from solving the linear ordering problem is by far the worst performer in the diagram. One observes it to yield likelihoods that are on average more that 1000 times smaller that the ones from the two point system. So we have to clearly reject the suggestion for a possible application of the LOP in soccer that has been made in the literature.

Another striking feature about the graphs is the position of the likelihood curves corresponding to the tabu search. The heuristic algorithm is able to improve on every single ranking from the sample, except for the Premier League seasons 04/05 and 05/06. On average it helps to explain the results about 457 times better. The graph shows us that even though the simple ranking schemes produce fairly "good" orderings in the sense of a high correlation (as seen above), they are far away from being the most likely correct ones.

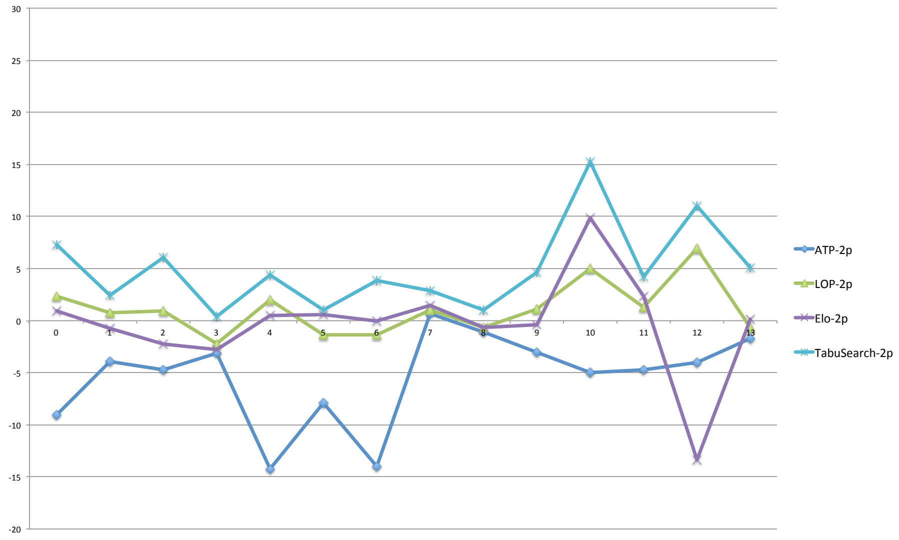


Figure 8: Maximum Likelihoods for ATP panel data

Next, Figure 8 shows the analogue results for the tennis panel data from the last 14 years. The first thing to note is that the two and three point systems produce the same likelihoods throughout the whole sample, which is why in this graph there is no curve comparing the two, since it would lie on the x-axis. This comes natural, because in Tennis we do not have draws, so in both systems the players are only ranked according to their number of victories. In Figure 7, in addition to the curves from Figure 6, the likelihoods from the official ATP ranking from the end of each year are listed. This ranking is determined by awarding different amounts of points for a stage that is reached in the Grand Slam Tournaments, the ATP World Tour Finals, the Masters 1000, Olympics etc. Of course this method is very sophisticated and includes also the results of the matches of the top 10 players against others that might not be in the top 10. This data is not part of the other systems we are analyzing. According to the criterion of this work, the ATP ranking performs fairly bad in explaining the observed results. Interestingly in this tennis sample, the linear ordering ranking produces fairly high likelihoods, in fact on average higher ones than the n-point and Elo system. Again, in every year the tabu search algorithm is able to improve on all of the discussed rankings.

Finally, Figure 9 illustrates the results form the same calculations as above, now for American football results from the National Football League in the US. There is little difference between 2- and 3 point systems, because draws are very unlikely to occur. However, the 3 point system is almost at every point at least as good as the 2 point system. The LOP and Elo systems operate in the same range of likelihoods as well. With NFL data, applying the tabu search is more effortful and thus takes more time for the same number of iterations, because of the higher number of teams. However, again, the tabu search improves upon all the rankings in the sample.

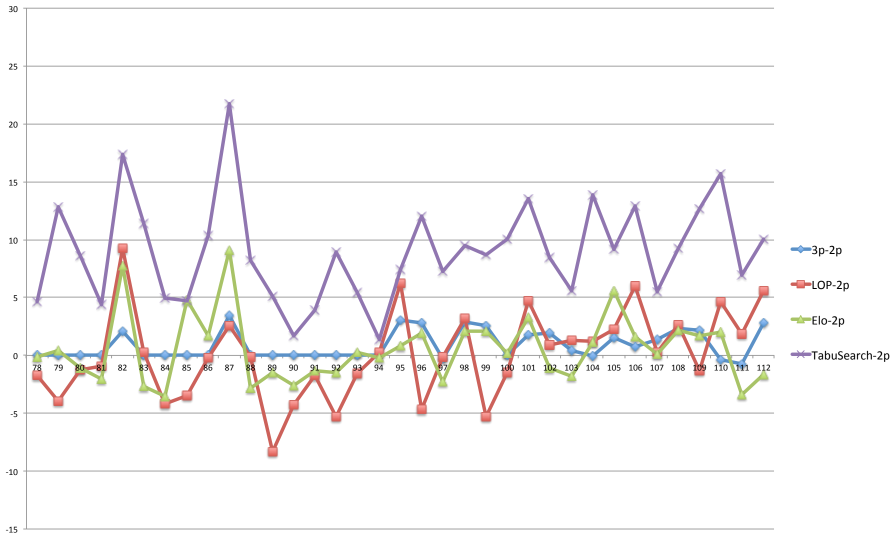


Figure 9: Maximum Likelihoods for NFL panel data

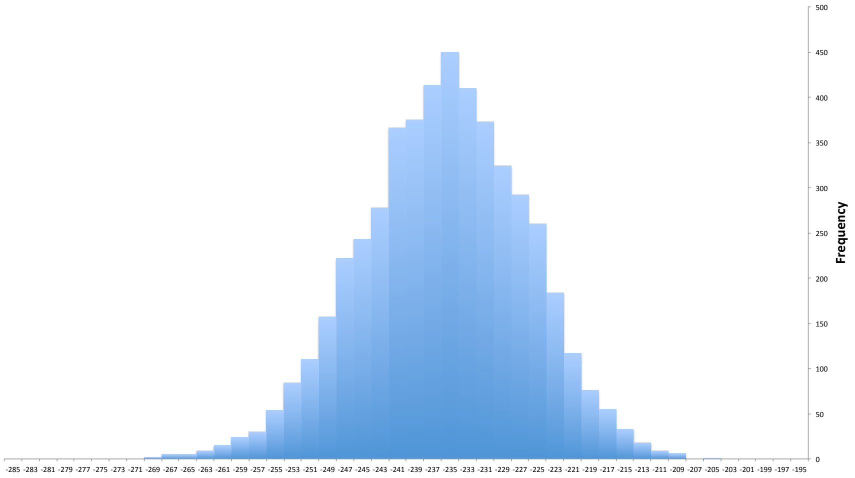
In general, the difference in the relative likelihoods when applying the Elo/LOP system and the n-point systems between soccer on the one hand and tennis and American football on the other could be due to the heterogeneity in the number of games played between the teams in tennis and football as opposed to the symmetric situation in soccer. Certainly a system like "two points for a win" doesn't seem to be particularly well suited in a situation where teams play different amounts of matches. And as explained further above, here it could be justified to give 1 and -1 points instead of 0 and 2 for a win and a loss, respectively. However, implementing this changes not much and even reduces the average likelihood a bit. Another explanation could be the sport itself. It might be due to the result generating probabilities themselves, that for one sport different ranking schemes are better suited then others. Indeed, it is easy to show that in the space of transitive probability matrices, there are areas where each of the considered systems is most likely to generate a ranking closest to the real one. This is an interesting direction for further theoretical research.

6.3 Hypothesis Testing

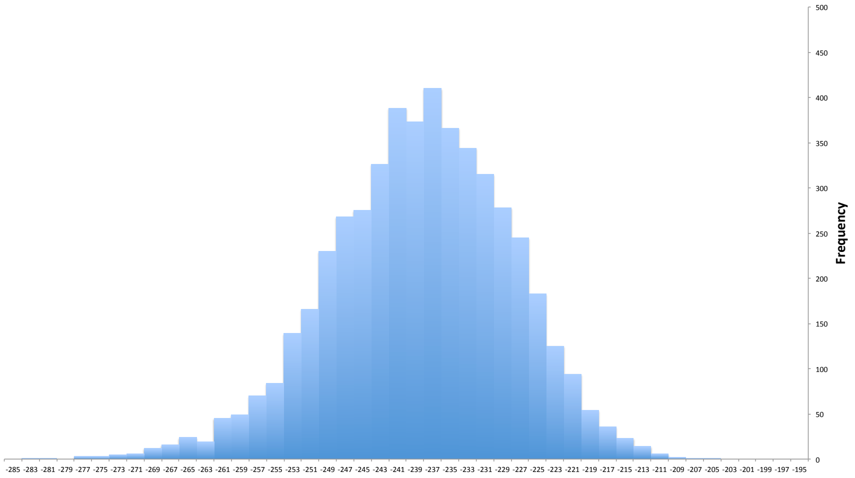
Now we are going further in the analysis of ranking systems than just observing which ordering scheme is able to gen erate a higher maximum likelihood value. We will consider two examples, which will help deepen the understanding of the problem, but will also clearly highlight the limitations of this hypothesis testing approach.

Consider the Bundesliga season 2011/2012. Looking at Figure 7 reveals that for this data set the 3 point system performed better than the 2 point system. The difference between the two maximum likelihood logs is 0.564. But the central question is "did this MLE difference appear because the underlying unobservable probabilities make the 3 point system more appropriate than the 2 point system in this season or could it in fact be the other way around with the observation just happening by chance?".

To answer this question, assume the correctness of the Hypothesis H_0: "The 2 point system puts the teams in the correct order". We will test H_0 against the alternative Hypothesis H_1: "The 3 point system puts the teams in the correct order". Now, for the two systems the probability matrices $\hat{P_{2p}}(w)$ and $\hat{P_{3p}}(w)$ are estimated. Using $\hat{P_{2p}}(w)$ 5000 seasons are simulated. Then $L(\hat{P_{2p}}(w)|\hat{w})$ and $L(\hat{P_{3p}}(w)|\hat{w})$ are calculated for each of the seasons. Their respective frequency distribution is depicted in Figure 9 (a) and (b). The distribution of their difference, which corresponds to the ration of the likelihoods without logs is plotted in Figure 9 (c).

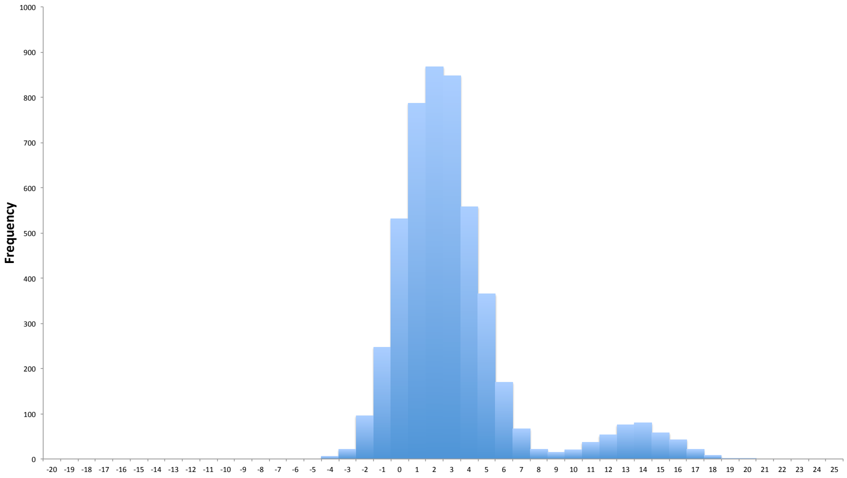


(a) $L(\hat{P_{2p}}(w)|\hat{w})$



(b) $L(\hat{P_{3p}}(w)|\hat{w})$

e



(c) $LR_{2p,3p}(\hat{w})$

Figure 9: Simulated test statistic for Bundesliga hypothesis test

Looking closely at the first two diagrams reveals that the distribution of $L(\hat{P}_{3p}(w)|\hat{w})$ is shifted a little bit to the left relative to the one of $L(\hat{P}_{2p}(w)|\hat{w})$. This is intuitively correct because it is only natural that the probability matrix that generated the seasons of the sample gives the higher likelihood values than the matrix $\hat{P}_{3p}(w)$, which has nothing to do with the season simulation. Now to find out the confidence level with which we would be able to reject H_0 one has to compare the observed likelihood ratio to the likelihood ratio distribution in Figure 9 (c). This procedure shows us that assuming the correctness of H_0, the probability of the likelihood ratio being ≤ 0.564 is only 11%. So we are able to reject the Hypothesis that the 2 point system gives the correct ranking with test size $\alpha = 0.11$, meaning that the probability of not making an error of the first kind is 0.89. One has to be careful not to misinterpret this result. It means that we are able to reject the hypothesis that the 2 point system gives the correct ranking. However, this does by no means imply that the 3 point system gives the correct ranking.

Now let us conduct a second hypothesis test, this time using tennis data. A good experiment would be to test for the correctness of the LOP system against the 3 point/2 point system in the year 2012. In this year the LOP produced a considerably higher likelihood than the 2 point system (see Figure 8), so we would like to know if this was just a random result or if we can actually conclude that the underlying probabilities favor the LOP scheme in the sense of telling us the truth about the ordering of tennis players. The hypothesis are:

H_0: "The 2 point system puts the teams in the correct order"

H_1: "The solution to the LOP puts the teams in the correct order"

Assuming the correctness of H_0, we again estimate the probability matrices and then simulate 5000 seasons. Hereby we always assume that the m_{ij} values stay constant, i.e. the amount of times players meet is the same in every simulation. We proceed as above by calculating the test statistic for the likelihood ratio and then comparing it to the empirically observed one. We have:

$$LR_{2p,LOP} = log(L(\hat{P}_{2p}(w)|w))) - log(L(\hat{P}_{LOP}(w)|w)) = -6.9103$$

The simulated test statistic tells us that in case H_0 is correct, the probability of an occurrence of such a small likelihood ratio is only 0.02%. It follows that we can reject H_0 with test size $\alpha = 0.02$ (i.e. a confidence level of 99.98%).

7 Conclusion

We constructed a statistical model describing the outcomes of sports matches. The model assumes a transitive relationship between the relative strengths of the teams. The resulting constraints turn out to be very restrictive, which is illustrated by the rapidly shrinking size of the parameter space.The incorporation of ties as well as home/away asymmetries makes our model much more complicated than the related isotonic regression problem. The discussed branch and bound algorithm is capable of solving the problem for up to 10 teams. For larger data sets, a tabu search heuristic has been proposed. The empirical section of the paper first illustrates the structure of an optimized probability matrix with an example. We have shown that in the example the maximum likelihood produced by the tabu search is more that 3000 times higher than the one resulting from an application of the 3-point system. But this does not mean that the two rankings are strongly uncorrelated as seen from the high value of Spearman's rank correlation coefficient. Panel data has been used to compare different ranking systems in three types of sports. In soccer, data from German Bundesliga and English Premier League have shown that the 2- and 3-point systems are very close to each other in the maximum likelihoods they produce, which is not a surprise when considering their structural similarity. Hopes were higher for the performance of the Elo system, because as opposed to the traditional point systems it considers the opponents strength. However, on average the generated MLEs were in the same range as the ones from the n-point systems. This result also applies for ATP tennis and NFL American football data. So the additional degree of complexity seems to be enough of a justification for not giving a recommendation towards an introduction of the Elo system. A difference worth mentioning is that the ranking, which results from the LOP performs fairly well in tennis and American football, but worse than everything else in soccer. We show that almost in every sample across all considered types of sports we are able to improve on the rankings produced by the considered systems by using tabu search. This illustrates that there might be a system that is much better at finding the most likely correct ranking, possibly without the inclusion of a great complexity. As a final remark, we want to mention that the framework presented in this paper has its natural limitations and leaves out many important aspects that should be considered when choosing or designing a ranking scheme. Things like opponents incentives during a match and the resulting effects on the observers level of thrill or the occurrence of winning decision as late as possible during a season could be interesting points for further research.

A Proofs

Proof of proposition 1: Ignoring the away/home differentiation, we can write p_{ikx} as p_{ik}. With 0 probabilities of draws, equation 1 is now

$$p_{ik} = 1 - p_{ki}$$

and therefore equation 2 is then equivalent to

$$\begin{aligned} p_{ik} \geq p_{il} \Leftrightarrow p_{jk} \geq pjl \\ p_{i'k'} > p_{j'k'} \Rightarrow p_{l'j'} > p_{l'i'} \end{aligned} \tag{5}$$

Now we have to show that $(SST) \implies (5)$ and $(5) \implies (SST)$.

$(SST) \implies (5)$:

We are dividing this case into two cases: For $p_{ik} \geq \frac{1}{2} \geq p_{jk}$ we can see:

$$p_{ij} \geq p_{kj} = 1 - p_{jk} \geq \frac{1}{2} \xRightarrow{SST} p_{ix} \geq p_{jx} \quad \forall x$$

For every other case we can assume wlog that $p_{ik} \geq p_{jk} \geq \frac{1}{2}$

$$\begin{aligned} p_{ij} \geq \frac{1}{2} \xRightarrow{SST} p_{ix} \geq p_{jx} \quad \forall x \\ p_{ij} < \frac{1}{2} \implies p_{ji} > \frac{1}{2} \xRightarrow{SST} p_{jk} > p_{ik} \end{aligned}$$

Which is a contradiction to the assumption, therefore $p_{ij} \geq \frac{1}{2}$.

$(5) \implies (SST)$:

$$\begin{aligned} p_{jk} > p_{ik} \xRightarrow{(5)} p_{li} > p_{lj} \qquad \forall l \\ \Rightarrow p_{ii} > p_{ij} \xRightarrow{p_{ii}=1/2} p_{ij} < \frac{1}{2} \end{aligned}$$

□

Proof of proposition 2: Define a ranking from best to worst $\rho(i) : \{1, .., n\} \multimap \{1, ..., n\}$ such that $p_{ikx} \geq p_{jkx} \Rightarrow \rho(i) < \rho(j)$ and $p_{kix} \leq p_{kjx} \Rightarrow \rho(i) < \rho(j)$.

$$\begin{aligned} p_{ikx} \geq p_{jkx} \Leftrightarrow \rho(i) < \rho(j) \Leftrightarrow p_{ily} \geq p_{jly} \quad \forall i, j, k, l, x, y \\ p_{kix} \geq p_{kjx} \Leftrightarrow \rho(j) < \rho(i) \Leftrightarrow p_{ily} \geq p_{jly} \quad \forall i, j, k, l, x, y \\ p_{i'k'x} > p_{j'k'x} \Leftrightarrow \rho(j) < \rho(i) \text{ and } \rho(i) \not\leq \rho(j) \exists i', j', k', x \end{aligned}$$

□

B Parameter space

In this section we explore the effect of transitivity conditions on the parameter space of winning probabilities to illustrate the limitations enforced by it. To do that we compare the size of the parameter space with transitivity to the space of unrestricted winning probabilities $\overline{S}_n$, e.g. every p_{ij}, p_{ji} fulfilling $p_{ij} + p_{ji} = 1$. The space of parameters including the transitivity conditions is a subset of this set $\overline{S}_n$. $S_n(R)$ is hereby defined as the size of this space relative to $\overline{S}_n$ only considering the restrictions for $p_{ij} \in R$. The unrestricted parameter space is in this simple case: $\overline{S}_n = [0,1]^{\frac{n(n-1)}{2}}$ which can be easily seen by the fact that every p_{ji} is completely determined by p_{ij}. The restricted space for n players and the transitivity conditions for every $(i,j) \in K_n$ with $K_n = \{(i,j)|i,j \in \{1,2,...,n\}, i < j\}$ is therefore

$$S_n(K_n) = \int\limits_{b_{i+1,j}}^{b_{i,j+1}} S_n(K_n \backslash \{(i,j)\}) \mathrm{d}p_{ij}$$

with

$$S_n((i_0,j_0)) = \int\limits_{b_{i_0+1,j_0}}^{b_{i_0,j_0+1}} \mathrm{d}p_{i_0 j_0}$$

and

$$b_{i,j} := \begin{cases} p_{ij}, & \text{for } (i,j) \in K_n \\ 0.5, & \text{for } i = j \\ 0, & \text{else} \end{cases}$$

As this fairly complicated recursive integral may be hard to interpret, table 2 gives the values for the relative size of the transitive parameter space for up to five teams. It can be seen that the size rapidly shrinks and it is not hard to imagine that for a league comprising e.g. 18 teams the conditions are in this sense very strict.

n	2	3	4	5	6	7
relative size	1	$\frac{1}{4}$	$\frac{1}{120}$	$\frac{1}{40320}$	$\frac{1}{203212800}$	$\frac{1}{19313344512000}$
Approximation	1	0.25	8.3×10^{-3}	2.5×10^{-5}	4.9×10^{-9}	5.2×10^{-14}

Table 2: Relative size of the transitive parameter space

References

RE Barlow and HD Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

H. D. Brunk. Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26(4):pp. 607–616, 1955a. ISSN 00034851. URL `http://www.jstor.org/stable/2236374`.

HD Brunk. Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26(4):607–616, 1955b.

Hugh D Brunk. Mathematical models for ranking from paired comparisons. *Journal of the American Statistical Association*, 55(291):503–520, 1960a.

Hugh D Brunk. Mathematical models for ranking from paired comparisons. *Journal of the American Statistical Association*, 55(291):503–520, 1960b.

Irène Charon and Olivier Hudry. An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Annals of Operations Research*, 175(1):107–158, 2010.

FRK Chung and FK Hwang. Do stronger players win more knockout tournaments? *Journal of the American Statistical Association*, 73(363):593–596, 1978.

Roger R Davidson and Daniel L Solomon. A bayesian approach to paired comparison experimentation. *Biometrika*, 60(3):477–487, 1973.

John S Decani. Maximum likelihood paired comparison ranking by linear programming. *Biometrika*, 56(3):537–545, 1969.

John S Decani. A branch and bound algorithm for maximum likelihood paired comparison ranking. *Biometrika*, 59(1):131–135, 1972.

Arpad E Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

Nicholas P Jewell and John D Kalbfleisch. Maximum likelihood estimation of ordered multinomial parameters. *Biostatistics*, 5(2):291–306, 2004.

Manuel Laguna, Rafael Marti, and Vicente Campos. Intensification and diversification with elite tabu search solutions for the linear ordering problem. *Computers & Operations Research*, 26(12):1217–1230, 1999.

Eugene L Lawler and David E Wood. Branch-and-bound methods: A survey. *Operations research*, 14(4):699–719, 1966.

Johan Lim, Xinlei Wang, and Wanseok Choi. Maximum likelihood estimation of ordered multinomial probabilities by geometric programming. *Computational Statistics & Data Analysis*, 53(4):889–893, 2009.

Rafael Marti and Gerhard Reinelt. The linear ordering polytope. In *The Linear Ordering Problem*, pages 117–143. Springer Berlin Heidelberg, 2011.

H William Morrison. Testable conditions for triads of paired comparison choices. *Psychometrika*, 28(4):369–390, 1963.

Jagbir Singh and WA Thompson. A treatment of ties in paired comparisons. *The Annals of Mathematical Statistics*, 39(6):2002–2015, 1968.

Patrick Slater. Inconsistencies in a schedule of paired comparisons. *Biometrika*, 48(3/4):303–312, 1961.

WA Thompson and Russell Remage. Rankings from paired comparisons. *The Annals of Mathematical Statistics*, 35(2):739–747, 1964.

Amos Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

C Van Eeden. Estimation in restricted parameter spaces-some history and some recent developments. *CWI Quarterly*, 9(1):69–76, 1996.

Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.