Bayesian inference and simulation approaches improve the assessment of Elo-scores in the analysis of social behaviour

Adeelia S. Goffe[1] Julia Fischer[1,2]

Holger Sennhenn-Reulen[1,2]

[1]Cognitive Ethology Laboratory, German Primate Center – Leibniz Institute for Primate Research, Germany

[2] Leibniz ScienceCampus Göttingen 'Primate Cognition', Göttingen, Germany

July 7, 2017

Abstract

1. The construction of rank hierarchies based on agonistic interactions between two individuals ("dyads") is an important component in the characterization of the social structure of groups. To this end, winner-loser matrices are typically created, which collapse the outcome of dyadic interactions over time, resulting in the loss of all information contained in the temporal domain. Methods that track changes in the outcome of dyadic interactions (such as "Elo-scores") are experiencing increasing interest. Critically, individual scores are not just based on the succession of wins and losses, but depend on the values of starting scores and an update ("tax") coefficient. Recent studies improved existing methods by introducing a point estimation of these auxiliary parameters on the basis of a maximum likelihood (ML)

approach. For a sound assessment of the rank hierarchies generated this way, we argue that measures of uncertainty of the estimates, as well as a quantification of the robustness of the methods, are also needed.

2. We introduce a Bayesian inference (BI) approach using "partial pooling", which rests on the assumption that all starting scores are samples from the same distribution. We compare the outcome of the ML approach to that of the BI approach using real-world data. In addition, we simulate different scenarios to explore in which way the Elo-score responds to social events (such as rank takeovers), and low numbers of observations.

3. Estimates of the starting scores based on "partial pooling" are more robust than those based on ML, also in scenarios where some individuals have only few observations. Our simulations show that assumed rank differences may fall well within the "uncertain" range, and that low sampling density, unbalanced designs, and coalitionary leaps involving several individuals within the hierarchy may yield unreliable results.

4. Our results support the view that Elo rating can be a powerful tool in the analysis of social behaviour, when the data meet certain criteria. Assessing the uncertainty greatly aids in the interpretation of results. We strongly advocate running simulation approaches to test how well Elo scores reflect the (simulated) "true" structure, and how sensitive the score is to "true" changes in the hierarchy.

Keywords: Agonistic behaviour, Bayesian inference, dominance, Elo rating, maximum likelihood estimation, partial pooling, rank hierarchy, simulation, social groups.

1 Introduction

Dominance hierarchies based on agonistic (or approach and retreat interactions) are often used to characterize the social structure of groups (e.g. Asiatic wild asses (*Equus*

hemionus) (Ganslosser and Dellert, 1997), dogs (*Canis familiaris*) (Cafazzo et al., 2010), geladas (*Theropithecus gelada*) (Johnson et al., 2014), Guinea baboons (*Papio papio*) (Kalbitzer et al., 2015), Przewalski horses (*Equus ferus przewalskii*) (Tilson et al., 1988), spotted hyaenas (*Crocuta crocuta)*) (Tilson and Hamilton, 1984), plains zebras (*Equus quagga*) (Ganslosser and Dellert, 1997), and sea lions (*Zalophus californianus*) (Schusterman and Dawson, 1968)). In addition, characteristics of social hierarchies such as the extent of the linearity in dominance interactions have been used in inter-specific comparisons (e.g. macaques (*Macaca* spp.) (Adams et al., 2015), and equids (*Equus* spp.) (Ganslosser and Dellert, 1997; Proops et al., 2012)).

Traditionally, scholars in the field of Animal Behaviour have generated social hierarchies by pooling sequential data into matrices, from which a quasi-static rank order can be derived for the time period under consideration (e.g. David (1987, 1988); de Vries (1998)). Implicitly, such an approach assumes that rank relationships are invariable. Due to migration events, death, or coalitionary upheaval, this assumption is rarely, if ever the case (e.g. Cheney et al. (2004); Haunhorst et al. (2017)). To acknowledge changes in the rank hierarchy, one common approach was to compare the hierarchies over different periods of time (Arseneau-Robar et al., 2017), or before and after specific events, such as the immigration of a new subject (Zhu et al., 2016). Yet, this method fails to track the potentially continuous changes in the rank order of subjects. Breaking down the assessment of rank hierarchies into ever shorter windows of time is no option, however, because short sampling periods result in a lack of data to infer the rank hierarchy.

In recent years, researchers in animal behaviour studies have therefore turned to the Elo-rating method, which allows tracking the outcome of individual dyadic interactions and assessing their influence on the overall rank hierarchy (Elo, 1961, 1978; Albers and de Vries, 2001; Neumann et al., 2011). Elo-rating was initially developed to assess the rating of chess players (Elo, 1961, 1978) and has been further applied to a variety of other sports.

One fundamental problem with a dynamic approach, such as Elo-rating, is that the

sampling begins at some arbitrary point in time where the animals in stable groups already possess (unknown) rank relationships (experimental approaches where subjects are grouped at the discretion of the researcher are not afflicted by this problem, e.g Tung et al. (2016)). As a solution for studies of already existing groups, an arbitrary score is assigned to each subject at the beginning of the study (Neumann et al., 2011), akin to Elo-rating in sports (Elo, 1978). Because this arbitrary score most likely deviates from the score that would be obtained if information about the past had been available, it may lead to considerable bias in the estimates of the Elo scores, especially in the first period of the study. The period until a certain equilibrium is reached – which is itself difficult to judge in this dynamic approach – is also known as the "burn-in" phase, which is often discarded for varying reasons (Neumann et al., 2011; Franz et al., 2015).

To avoid such loss of information, Foerster et al. (2016a) recently introduced a maximum likelihood (ML) approach to estimate the starting scores, as well as the winning/loosing tax constant (k), on the basis of the observed interactions (x_j), such that the complete course of Elo scores most plausibly matches this sequence of observed interaction outcomes. This is an important contribution, since it overcomes a shortcoming of the classical way of calculating Elo scores that was depending on artificially determined starting scores, as well as an artificially determined tax constant. The ML approach is able to by-pass the problems generated by having predetermined/artificial starting values (see Neumann et al. (2011)). Yet this method creates a novel 'downstream' problem. Specifically, in order to see whether a group of individuals already has a clearly developed hierarchy at the beginning of a study, separate models – with and without estimating starting scores – are fitted. This generates a decision task between those two models. Because decisions are always accompanied by uncertainties, these need to be incorporated in inferences drawn from the data.

Further, little is know about the elasticity of the Elo rating method. One of the benefits of this method is the possibility to assess the temporal dynamics in winner-loser interactions, yet we know very little how fluctuations in Elo scores emerge and how well they reflect "true" scores. Three aspects of the temporal responsiveness of the Elo rating

method are of specific importance: (1) the response to changes in the true hierarchy, (2) the response to variations in the predictability of each single outcome, and (3) the response to varying interaction rates within the group.

In order to address these concerns we suggest using an approach that: (i) gives a direct quantification of the variation in the starting hierarchy, (ii) shows how this interplays with the winning/loosing update coefficient (k), and (iii) directly introduces the above "model decision uncertainty" into a single estimation result. We achieve this by a Bayesian inference (BI) approach applying the concept of *partial pooling*, which assumes that all starting scores are samples from the same distribution with a shared variation parameter σ. Partial pooling is the current state-of-the-art concept for tasks where a compromise is searched between modeling the full individual variation (no pooling, equal to the ML result), and modeling no individual variation (full pooling, only one single population coefficient) within a population: Each individual is assumed to have a different ability for winning an encounter, but the data for all of the observed individuals informs the estimate of each individual.

Here, we use real-world data from Foerster et al. (2016a) in order to compare the utility of ML estimation to our newly introduced BI approach (using partial pooling) in the prediction of starting values. Further, we use simulations to explore issues associated with how well Elo scores 1) represent the "true" rank order and rank distances in unbalanced designs in which dyads vary in their interaction rates, and 2) respond to changes in the "true" hierarchy.

The remainder of this manuscript is organized as follows: We begin by introducing a BI approach for estimating starting scores and the Elo score update coefficient (k). Subsequently, we present the results of a re-analysis of real-world data, followed by simulation studies to evaluate several aspects that are important when working with Elo scores on behavioural data. The Stan (Carpenter et al., 2017b) code implementing the BI approach is included in supplementary material S1, and the full R code implemented in the statistical software environment R (R Development Core Team, 2016) used for the purpose of

this article is available under `https://github.com/holgersr/Bayesian-inference-and-simulation-of-Elo-scores-in-analysis-of-social-behaviour`.

2 Methods

2.1 Calculating Elo scores

Immediately after interaction $j = 1, \ldots, J$, the Elo scores ($\text{Elo}_{i,j}$ of individuals $i = 1, \ldots, n$) are recursively defined (following Franz et al. (2015)) as a function of Elo scores before the first interaction ($\text{Elo}_{i,0}$), all interaction outcomes in between, and a winning/loosing tax coefficient (k), by:

$$\text{Elo}_{A_j,j} = \text{Elo}_{A_j,j-1} + k \cdot \left(1 - \frac{1}{1 + \exp\left(-0.01 \cdot \left(\text{Elo}_{A_j,j-1} - \text{Elo}_{B_j,j-1}\right)\right)}\right),$$

for individual A_j winning over individual B_j ($A_j, B_j \in \{1, \ldots, n\}$) – the Elo score of A_j increases –, and by:

$$\text{Elo}_{B_j,j} = \text{Elo}_{B_j,j-1} - k \cdot \left(1 - \frac{1}{1 + \exp\left(-0.01 \cdot \left(\text{Elo}_{A_j,j-1} - \text{Elo}_{B_j,j-1}\right)\right)}\right),$$

for B_j loosing against A_j – the Elo score of B_j decreases. Here, A_j and B_j denote the two individuals participating in interaction j. For all remaining individuals $i \notin \{A_j, B_j\}$, the Elo score remains unchanged, ie. $\text{Elo}_{i,j} = \text{Elo}_{i,j-1}$. Note that for notational convenience, we will use only A and B in the following to refer to A_j and B_j.

In an Elo update as described by the above equations, the central role is played by the factor:

$$1 - \frac{1}{1 + \exp\left(-0.01 \cdot \left(\text{Elo}_{A,j-1} - \text{Elo}_{B,j-1}\right)\right)},$$

which describes the probability that A looses interaction j against B. This is a complicated way of writing down the logistic function, known as response function from the logistic regression model (e.g. in Fahrmeir et al. (2013), Section 5.1.1):

$$1 - \frac{1}{1 + \exp\left(-0.01\left(\mathrm{Elo}_{A,j-1} - \mathrm{Elo}_{B,j-1}\right)\right)} = \frac{\exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)}{1 + \exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)}$$

$$= \begin{cases} < 0.5, & \text{if } \mathrm{Elo}_{A,j-1} > \mathrm{Elo}_{B,j-1}, \\ 0.5, & \text{if } \mathrm{Elo}_{A,j-1} = \mathrm{Elo}_{B,j-1}, \\ > 0.5, & \text{if } \mathrm{Elo}_{A,j-1} < \mathrm{Elo}_{B,j-1}\,. \end{cases}$$

See Supplement S2 for a derivation of this equation.

By the recursive definition of Elo scores, the Elo scores before the first interaction $\mathrm{Elo}_{i,0}$ (denoted as *starting scores* in the following) become 'active' at the interaction where individual i is interacting for the first time. So, the index terminology with 0 for the interaction index does not strictly refer to the interaction index before the first observed interaction of the whole group (i.e., it does not necessarily mean $j = 0$), but refers to the interaction previous to the first observed interaction of individual i.

In the approach introduced by Foerster et al. (2016a), this estimation is achieved by maximizing the log-likelihood function l, which is defined by the sum of $\log\left(p_{A,B,j}\right)$ across all observed interactions j, i.e.

$$l\left(\mathbf{x}|\,\mathrm{Elo}_{1,0}, \ldots, \mathrm{Elo}_{n,0}, k\right) = \sum_{j=1}^{J} \log\left(p_{A,B,j}\right),$$

where A and B denote the individuals involved in interaction j, $\mathbf{x}$ denotes an observation vector containing all interaction outcomes $x_{A,B,j}$, $j = 1, \ldots, J$, and $p_{A,B,j}$ is the probability that A wins against B, i.e.

$$p_{A,B,j} = \frac{\exp\left(0.01\left(\mathrm{Elo}_{A,j-1} - \mathrm{Elo}_{B,j-1}\right)\right)}{1 + \exp\left(0.01\left(\mathrm{Elo}_{A,j-1} - \mathrm{Elo}_{B,j-1}\right)\right)}.$$

2.2 Reformulation of the winning/loosing probability term

In the upper definition of Elo scores, the factor:

$$p_{A,B,j} = \frac{1}{1 + \exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)} = 1 - \frac{\exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)}{1 + \exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)}$$

determines the probability that A wins interaction j against B (this is according to the definition by Albers and de Vries (2001); see also Franz et al. (2015); Neumann et al. (2011)).

We use a slightly different formulation for this winning probability equation that replaces the factor $\delta = 0.01$ in the denominator by $\delta = 1$:

$$p_{A,B,j} = \frac{1}{1 + \exp\left(\text{Elo}_{B,j-1} - \text{Elo}_{A,j-1}\right)}.$$

This definition is equivalent to the well-known logistic distribution function for a random variable defined by dyadic difference in Elo scores, i.e. by $\text{Elo}_{B,j-1} - \text{Elo}_{A,j-1}$, and has some benefits for interpretation: For $\text{Elo}_A - \text{Elo}_B$ measuring the difference in Elo scores between two individuals A and B, the neutral odds $1 = 1/1$ for A winning over B are multiplied by $\exp(\text{Elo}_A - \text{Elo}_B)$, i.e.:

$$\frac{P(A \text{ wins} \mid \text{Elo}_A - \text{Elo}_B)}{P(B \text{ wins} \mid \text{Elo}_A - \text{Elo}_B)} \Big/ \frac{P(A \text{ wins} \mid \text{Elo}_A = \text{Elo}_B)}{P(B \text{ wins} \mid \text{Elo}_A = \text{Elo}_B)} = \exp(\text{Elo}_A - \text{Elo}_B).$$

By using $\delta = 1$, differences between Elo scores become directly interpretable on the familiar logistic scale: a difference in Elo scores of the value 5 – such as one individual has an Elo score of 1 and the other individual has an Elo score of 6 – leads to a probability of 99.3% (R command `plogis(5)`) that the individual with the higher Elo score wins an interaction between these two individuals, and a probability of 0.7% to loose it. For the difference sequence 0, 1, 2, ..., 5, the call to the `plogis()` function returns: 50.0%, 73.1%, 88.1%, 95.3%, 98.2%, and 99.3%.

It only needs basic calculation rules for random variables to show (Supplement S3) the standard deviation equivalence by varying δ, and also how k scales with δ.

2.3 Bayesian estimation of starting scores and tax coefficient (k)

In contrast to Foerster et al. (2016a), we use a fully Bayesian statistical inference approach to estimate starting scores and k (Supplement S1 gives Stan code for a software implementation). With this, we are able to make probabilistic prior statements – for all

of the coefficients $(\text{Elo}_{1,0}, \ldots, \text{Elo}_{n,0}, k)$ included in the probabilistic formulation of the data generating mechanism, i.e. included in the likelihood – in the form of probability densities:

$$p\left(\text{Elo}_{1,0}, \ldots, \text{Elo}_{n,0}, \sigma, k\right) \propto p\left(\text{Elo}_{1,0}, \ldots, \text{Elo}_{n,0} \,|\sigma\right) \cdot p\left(\sigma\right) \cdot p\left(k\right).$$

With the use of Bayes' theorem, we may base post-estimation calculations as well as inferences for the starting scores $(\text{Elo}_{i,0})$ and the tax coefficient (k) on a direct probability density function termed the *posterior density*:

$$p\left(\text{Elo}_{1,0}, \ldots, \text{Elo}_{n,0}, \sigma, k|\mathbf{x}\right) \propto l\left(\mathbf{x}|\, \text{Elo}_{1,0}, \ldots, \text{Elo}_{n,0}, \sigma, k\right) p\left(\text{Elo}_{1,0}, \ldots, \text{Elo}_{n,0}, \sigma, k\right).$$

2.3.1 Prior statement for starting scores

In application, individuals are observed with a different number of interactions, where it is rather the rule than the exception that some individuals only have a very small number of observed interactions. For those rarely observed individuals, only scant information is available, leading to increased standard errors in a direct ML estimation. Importantly, in extreme cases where an individual either looses or wins all interactions, the marginal likelihood is completely flat, and the starting score estimation become non-identifiable.

Instead of estimating the starting scores completely independently (as performed in ML estimation), a solution to the above problem is to estimate a scale parameter for a shared prior distribution of starting scores. This approach is the previously mentioned *partial pooling*, suitable in various application scenarios (Gelman et al., 2016), and conceptually very close to Bayesian Ridge Regression (Fahrmeir et al., 2010).

We choose the starting values $\text{Elo}_{i,0}$ coming from a shared population distribution:

$$\text{Elo}_{i,0} \,|\sigma \sim \text{N}\left(0, \sigma^2\right), \quad \text{with} \quad \sigma \sim \text{N}_+\left(0, 1\right), \quad \sigma \geq 0,$$

where, without any loss of generality, we fix the mean of the prior for the starting scores to 0, since the mean Elo score is not well-defined (recall that in the definition of Elo scores, the value of a score only matters relative to the other scores). For the scale parameter σ, the distribution $\text{N}_+\left(0, 1\right)$ denotes a truncated (beyond at 0) standard

normal distribution.

This "shared-distribution-assumption" acts as an informative prior to the starting scores, where a general rule of Bayesian statistics applies that the strength of a prior is relative to the flatness of the marginal likelihood: for the influence of a prior, it only matters how flat the prior is in the region of high likelihood, and vice versa.

- For an individual with a compact region of high marginal likelihood (an 'informed individual'), the prior has a very small gradient across this range, the prior has only a weak influence.

- For an individual with a wide region of high marginal likelihood (an 'un-informed individual'), the prior incorporates its full shape, the prior has a strong influence (since the mean of the partial pooling prior is equal to the mean of the starting scores, this prior has a regularizing influence).

By this, partial pooling cures the problem of inflated standard errors in a very natural way ("at a cost of (typically small) bias"(Fahrmeir et al., 2013, p. 238)): when the knowledge about an individual is very sparse, the most plausible assumption is to assign it a value that does not differ substantially from the population mean, i.e. the individual is ranked in the center of the hierarchy, and does not stand out as either very dominant or subordinate. Partial pooling thus allows a single starting score to speak with a louder voice if it is based on strong information, i.e. a large group of observations from the same individual. Conversely, if little information is available, the starting score does not have a strong influence.

The use of the logistic distribution function is again helpful to interpret the scale parameter σ in the distribution of starting scores: σ is the standard variation from a normal distribution, and $[-2 \cdot \sigma, 2 \cdot \sigma]$ roughly spans an inner 95% probability interval of starting scores. Applying the logistic quantile function, we can transform this to an inner 95% probability interval for winning probabilities in comparison to an individual that is expected to win against the one half of the group, and to loose against the other

half; such an individual would have the central Elo score of 0. This view on σ from the perspective of assumed differences in Elo scores, and the differences in probabilities they refer to, acts as a helpful tool while thinking about priors for σ.

Although the use of a truncated standard normal distribution prior for σ is not uncommon, we want to point out that, as for any prior assumption in practice, sensitivity checks should be performed. For example by the use of Hellinger distances in the spirit of Roos and Held (2011), or by refitting using a different prior formulation (see Figure 3 for an example).

2.3.2 Prior statement for winning/loosing tax coefficient

The winning/loosing tax coefficient (k) plays a central role in allowing for – and measuring – dynamics of hierarchies by Elo scores.

Neumann et al. (2011) write:

> "Previous experience of individuals plays an important role in the outcome of agonistic encounters in many animal taxa: the winner of a previous interaction is more likely to win a future interaction, whereas losers are more likely to lose future interactions."

Albers and de Vries (2001), as well as Neumann et al. (2011) used different selections for k in order to compare the changes in Elo scores and hierarchies that go with it, however. Neumann et al. (2011) further write:

> "Its value is usually set between 16 and 200 and, once chosen, remains at this value throughout the rating process."

Foerster et al. (2016a) bound $\exp_e(k)$ between -10 and 10 (in the dryad archive they refer to in their article), and by this k between approximately 0.000 045 and 22 000. This shows that they allow a much wider spread to k than Neumann et al. (2011), but certainly not all of those values make, a priori, equal sense.

This can be demonstrated using two toy-examples:

- if k is 1, and two individuals A and B with scores $\text{Elo}_A = 1$ and $\text{Elo}_B = 3$ ($\delta = 1$) have an encounter (`plogis(2)` is 0.88), then the scores change to $\text{Elo}_A = 0.88$ and $\text{Elo}_A = 3.12$ if A wins, and to 1.88 and 2.12 if B wins. The post encounter probability (that the originally higher scored individual B wins) is 0.90 in the first case, and 0.55 in the second case.

- everything unchanged, but now $k = 2$ instead of $k = 1$, the scores change to 0.76 and 3.24 if B wins, and to 2.76 and 1.24 if A wins. The post encounter probability (that the originally higher scored individual B wins) is 0.92 (`plogis(3.24-0.76)`) in the first case, and 0.18 (`plogis(1.24-2.76)`) in the second case.

Thus, the selection of $k = 2$ (for $\delta = 1$) results in a complete turn-around of the two scores in the case of the unexpected outcome. Consequently, any larger values will hardly yield sensible results.

Because of this, we attribute the same prior to k as we have done to σ:

$$k \sim \text{N}_+ (0, 1), \quad k \geq 0,$$

where $\text{N}_+ (0, 1)$ again denotes a truncated (beyond at 0) standard normal distribution. This informative prior is hopefully strong enough to avoid pathologically large posterior samples for k, but hopefully still weak enough to "let the data speak for themselves". Again, sensitivity checks to the prior choice are recommended in practice.

2.3.3 The effect of newly introduced members on the ranking of "silent members"

In the analysis of long-term data on agonistic interactions of female eastern chimpanzees (*Pan troglodytes schweinfurthii*) (Foerster et al., 2016a), individuals quitting (through emigration or mortality) typically had achieved high Elo scores, while newly entering individuals (through immigration or birth) had low Elo scores. Remaining individuals who had not been involved in any interactions for a long period of time (we call them "silent members"), automatically "march through" the hierarchy (see Figures 1 (e, f) and 2 of the female eastern chimpanzee data from Foerster et al. (2016a)).

We therefore decided to subtract the mean Elo score of the present individuals at each interaction from the Elo scores of the present individuals at this interaction (i.e. not present individuals don't change). Each individual's Elo score is then directly interpretable – at any time – as the difference to the mean of the current group, which is 0. By this, changes in Elo scores of "silent members" become more clearly visible.

This "correction" of the mean is also important in the Bayesian estimation of starting scores, since the applied partial pooling approach assumes that all starting scores are sampled from a distribution with equal mean, which we need to guarantee in the algorithm (see appendix S1).

2.4 Real-World Application

We first conducted a re-analysis of long-term data (Foerster et al., 2016b) on agonistic interactions of female eastern chimpanzees, studied at Gombe National Park, Tanzania. The study population consisted of $n = 44$ female individuals and data were collected between the years 1969 – 2013. As in Foerster et al., we did not include the first 100 interactions in order to facilitate methodological comparison between the ML and BI approaches; a total of $J = 915$ agonistic interactions were included in the analysis.

This dataset also allowed us to study the two estimation approaches in a scenario in which one individual contributes with losses (or wins) only. The occurrence of only wins or losses is inadvertently the case when extremely strict and completely stable hierarchies are analysed. Foerster et al. (2016a) addressed this issue by removing 13 females that incorporated solely wins or losses in order to "facilitate model fitting and interpretation" (p.2).

2.5 Build up for simulation 1–High-dimensional "estimation problem"

We simulated different numbers of interactions in the scenario given by the Foerster et al. (2016a) female data ($N = 44$, $k = 0$). The marginal probability that an individual is involved in one of the simulated interactions is equal for all the individuals, which should

result in a more balanced data-set as used in Foerster et al. (2016a). This was achieved by randomly sampling one individual for each of the simulated interactions. The opponent identity was realized by sampling from the respective remaining individuals, with the probability being inversely proportional to the true underlying difference in starting scores to the first individual. The true underlying starting scores were sampled according to our result on the female eastern chimpanzee data ($\text{Elo}_{0,i} \sim \text{N}\left(0, \sigma^2 = 2.62^2\right)$, when fixing k to 0). Since there were no changes in the underlying scores estimated in the original study, a true tax coefficient of 0 is utilized in this simulation scenario. The numbers of interactions vary by $250, 500, 1000, 1500$, and the results of comparing the BI with the ML approach are illustrated in Figure 3.

2.6 Build up for simulation 2–General unbalanced design

In comparison to the above build-up, the marginal probability that an individual is involved in one of the simulated interactions is proportional to the true underlying scores, i.e. the lower the score, the lower the interaction rate. We reduced the number of individuals to $N = 10$, to obtain a clearer picture of the uncertainties in this scenario. To introduce a tax coefficient different from 0, we swapped the scores of neighboring ranks as 2 with 1, 4 with 3, ..., 10 with 9. We further based this simulation on an equidistant grid of underlying Elo scores from -6 to 6 of a length $N = 10$. This lead to individual specific numbers of interactions between 398 and 1243 in the first run, and between 410 and 1207 in the second run.

2.7 Build-up for Social Instability Simulations

In order to assess the flexibility of Elo rating to adapt to changes in social dynamics, we then performed three further simulations representing scenarios similar to hierarchical social changes which may naturally occur in animal groups. All estimations were performed using the BI approach. Statistical assessment of instability was performed using by calculating the proportion of mismatches.

2.7.1 Simulation 3A–External Takeover

In order to simulate a change in social hierarchy which might occur in the event of an external rank takeover, for instance as a result of an immigration event (Marty et al., 2015; Cheney et al., 2004), we used an equidistant grid of underlying Elo scores from -6 to 6 with a length of $N = 10$. The simulation is divided into two periods. In the first period of 2000 interactions we included 10 individuals and a stable social hierarchy. In the second period of 2000 interactions, the formerly dominant individual dropped to the bottom position, and another individual (the "immigrant") is introduced at the top, taking over the alpha position. During this second period, the underlying social hierarchy was also stable.

2.7.2 Simulation 3B–Coalitionary Leap

In the second scenario we generated a dataset which represented a coalitionary-based rank change, such as one that might occur in a nepotistic society (Weiß et al., 2011). In the first period of 1000 interactions we included 25 individuals, divided into five subgroups of five individuals each, with a stable social hierarchy. In the second period of 2000 interactions, the third ranked group was moved to the top position; for the sake of simplicity we did not introduce a change in position within the subgroups. We based this simulation on an equidistant grid of underlying Elo scores from -6 to 6 with a length of $N = 25$.

2.7.3 Simulation 3C–Mortality-Instability-Recovery

In the third scenario we mimicked a mortality event followed by social upheaval and subsequent stability (Kaburu et al., 2013). In the first period of 1000 interactions we included 10 individuals in a stable social hierarchy (equidistant grid of underlying Elo scores from -6 to 6). In the second period of 1000 interactions, the top two ranked individuals were removed from the hierarchy and the ranks of the remaining 8 individuals were unstable. In the third period of 1000 interactions the social hierarchy was again stable.

3 Results

3.1 Real world application: female agonistic interaction data

The uncertainty of Elo scores – as illustrated by the shades in Figure 1 – indicated a considerable degree of uncertainty in the estimated coefficients. Bayesian estimation allowed us to answer the questions whether to start at the same/very close values, and to have a small or large k within one single estimation run, rather than a two-step model comparison approach. This transfers the increased uncertainty – introduced by selecting among different models – into the estimation problem.

When we calculated correctly predicted interaction proportions (Foerster et al., 2016a), we yielded 90.7% in comparison to 89.4% (we were not able to reproduce the value of 89.8% as given in Table 1 in Foerster et al. (2016a)). As a further in-sample prediction criterion, we calculated mean Brier scores:

$$\bar{\text{bs}} = \frac{1}{915} \sum_{i=1}^{915} (1 - \hat{p}_{A_i,B_i,i})^2 \,,$$

which not only incorporates the correct sign of the Elo score difference, as does the correctly predicted interaction proportion, but further takes into account whether the Elo scores were able to clearly distinguish the winning from the loosing individuals. A large positive difference is a greater success in comparison to a small positive difference; a small negative is less worse in comparison to a large negative difference. In comparison to the ML approach the BI approach led to a reduction from 0.085 to 0.075.

Bayesian estimation also allowed us to achieve our primary aim of reliably estimating Elo score estimates. This came directly from the distribution/spread of posterior samples. What seemed to be a clear hierarchy (Foerster et al., 2016a, Fig. 1f), was in fact unstable with a high degree of overlap between certain individuals (Figure 1).

The BI approach did not yield strong evidence for the claim $k = 0$ (Figure 2). The visual comparison of the differences between the ML and the BI approach did not suggest fundamental consequences at the practical level, with only a slight advantage for the BI

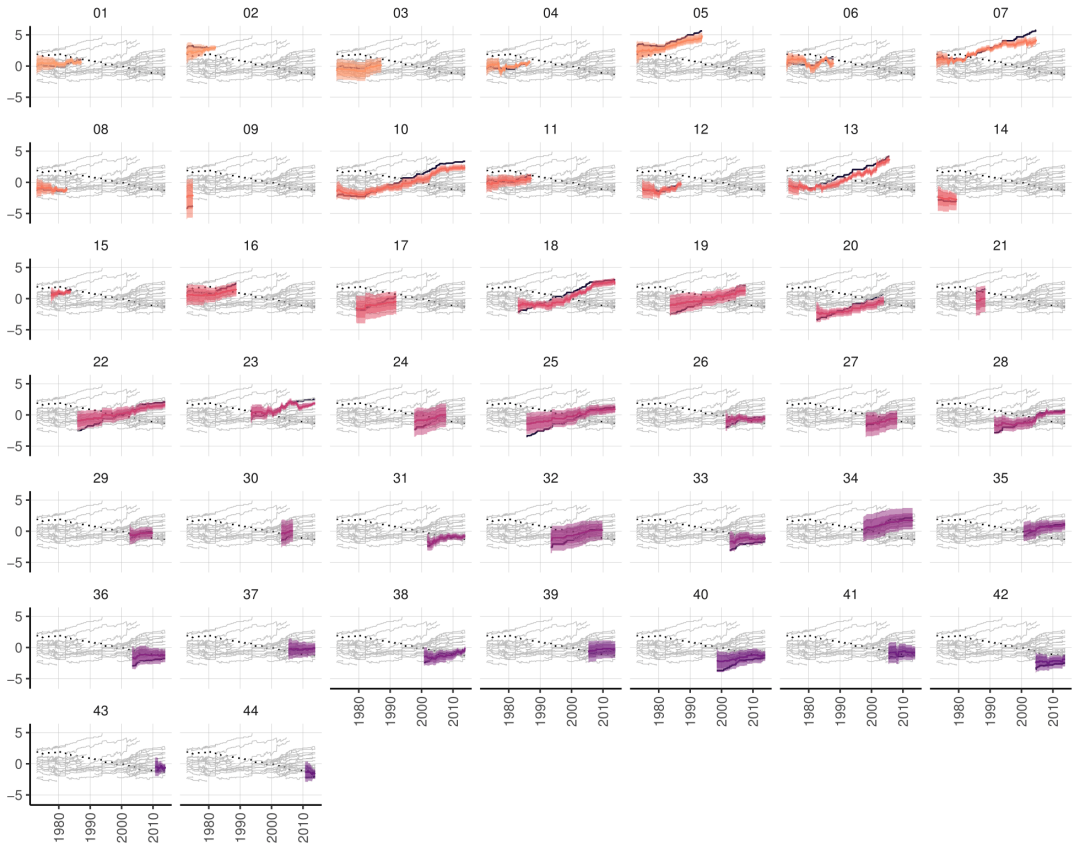

Figure 1: Path plots for the Foerster female data: Posterior means of our Bayesian estimation approach are shown as solid coloured step functions, results from Foerster et al. (2016a) as solid black step functions. The areas show 95% and 80% credible intervals, incorporating the uncertainties about starting scores as well as about k.

approach.

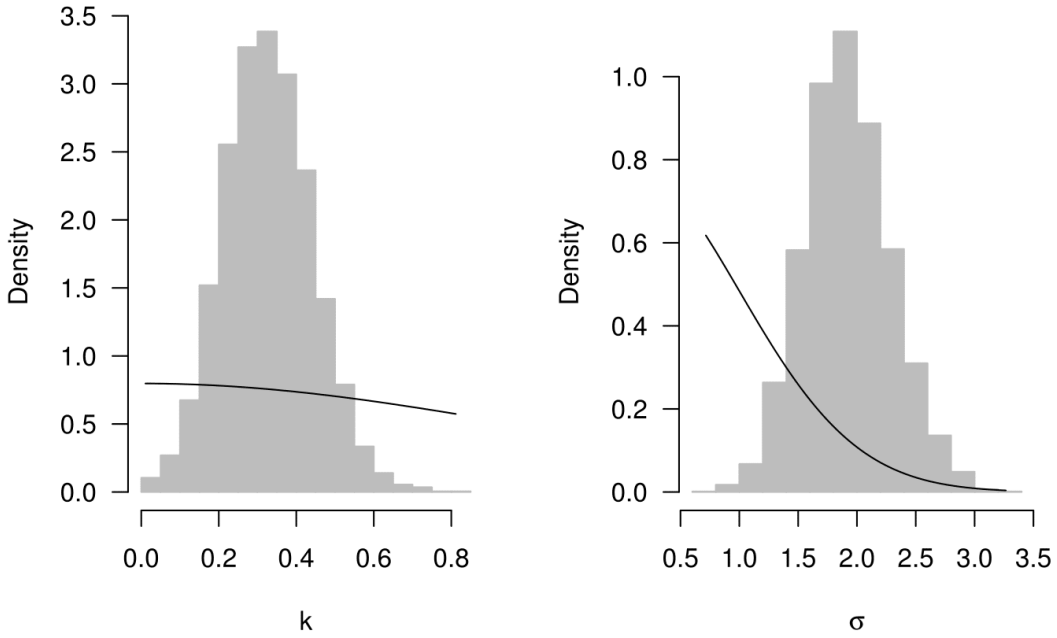

Figure 2: Comparison of posterior density (grey shaded histograms) to the prior density (solid black line) $k \sim \mathrm{N}_+(0,1)$, and $\sigma \sim \mathrm{N}_+(0,1)$.

3.2 Simulation results

For a quantification of the ability of Elo scores to correctly rank the individuals at each temporal stage j, we calculated the mean absolute error (Ranks MAE) in relation to the underlying true ranks:

$$\text{Ranks MAE}_j = \frac{1}{n}\sum_{i=1}^{n} |\,\text{RevOrder}\,(\text{Elo}_{ij}) - i|,$$

where the RevOrder (for "reversed order") function gives a value of 1 to the highest Elo score at the current interaction index, and a value of n to the lowest Elo score (note that this definition only works if prior to calculation the individual index $i = 1, \ldots, n$ already gives the rank in the truly underlying hierarchy).

3.2.1 Simulation 1–Maximum likelihood vs Bayesian estimation in a high-dimensional design

As illustrated by the results in Figure 3, the estimation based on the BI approach led to an improved stability in comparison to the ML approach. Even for relatively high number of observed interactions, the ML estimation led to unstable point estimates that differed substantially from their underlying ground truth (dashed line).

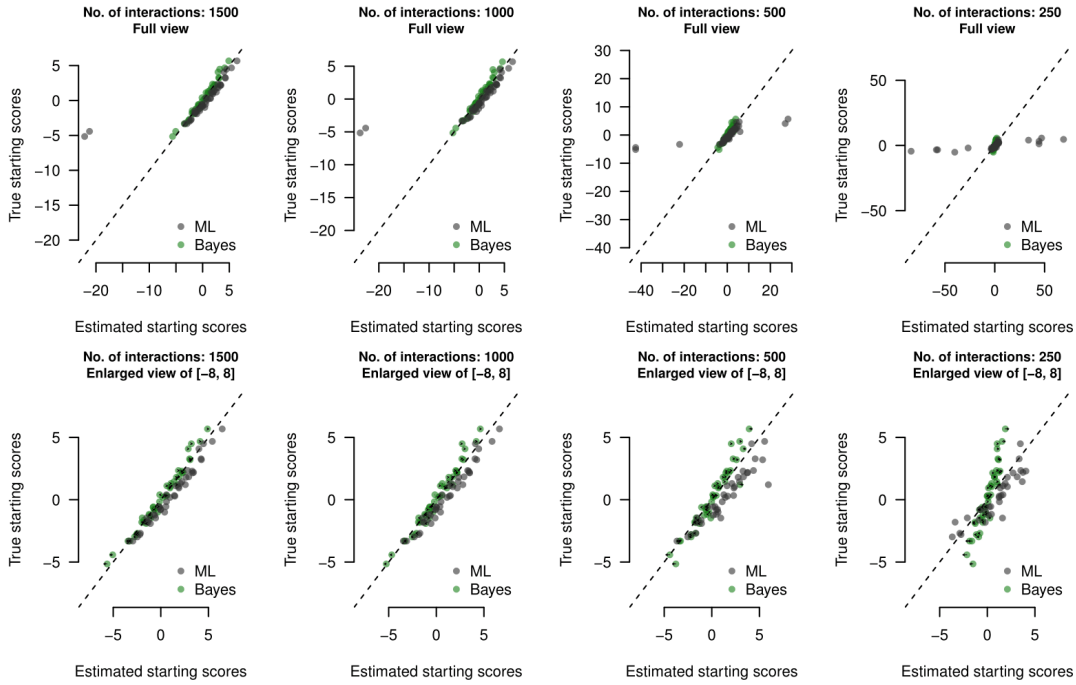

Figure 3: Results of starting score estimation for simulated data: x-axis values are true underlying starting scores, y-axis show estimated starting scores (posterior mean for Bayesian approach, ML estimates for the benchmark maximum-likelihood approach). Black arrows show the influence by changing the prior from $\sigma \sim \mathrm{N}_+(0,1)$ to $\sigma \sim \mathrm{N}_+(0,2)$: For small sample sizes, the larger prior standard deviation leads to a slight improvement in the bias of the estimates, for larger sample sizes, the posterior is more and more dominated by the Likelihood, and therefore the prior influence vanishes.

3.2.2 Simulation 2–General unbalanced design

If the number of interactions is highly unbalanced, it is likely that the estimation of the Elo scores of individuals with low interaction rates are inaccurate. In the present scenario, individuals with lower rank interacted at lower rates than more dominant individuals. For individuals with comparatively low sample sizes, Elo scores appeared to be more volatile and the lower portion of the dominance hierarchy was unstable (Figure 4).

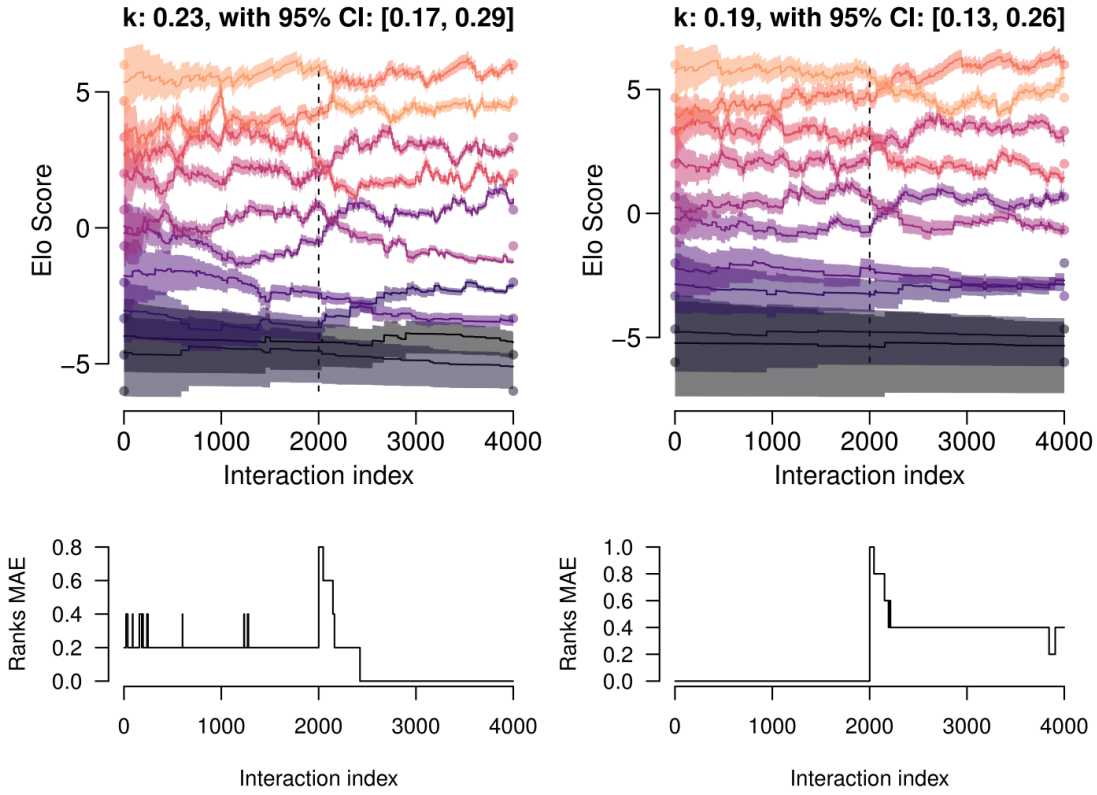

Figure 4: Results of two runs for simulation 2. Shaded areas give 95% probability intervals for the Elo scores at each interaction, based on the posterior samples for the starting values. The underlying true values within the two periods of 2000 interactions each are illustrated by bullet points at the beginning of the respective periods.

3.2.3 Recovery of Elo scores to changes in the hierarchy

Simulation 3A–External Takeover Figure 5 shows the results of two different simulation runs in which an external takeover of the alpha position occurs. At 200 interactions after the external takeover, Elo scores began to reflect the true underlying rank order. There was consistent underlying stochasticity in both stable periods.

The degree of uncertainty varied during different periods of the interaction sequence (Figure 5). Following the BI approach, uncertainty, as indicated by credible intervals, was initially higher in all 10 individuals at the start compared to the end of period 1. In period 2, the uncertainty of the new alpha was highest while the uncertainty of all ancillary individuals was reduced.

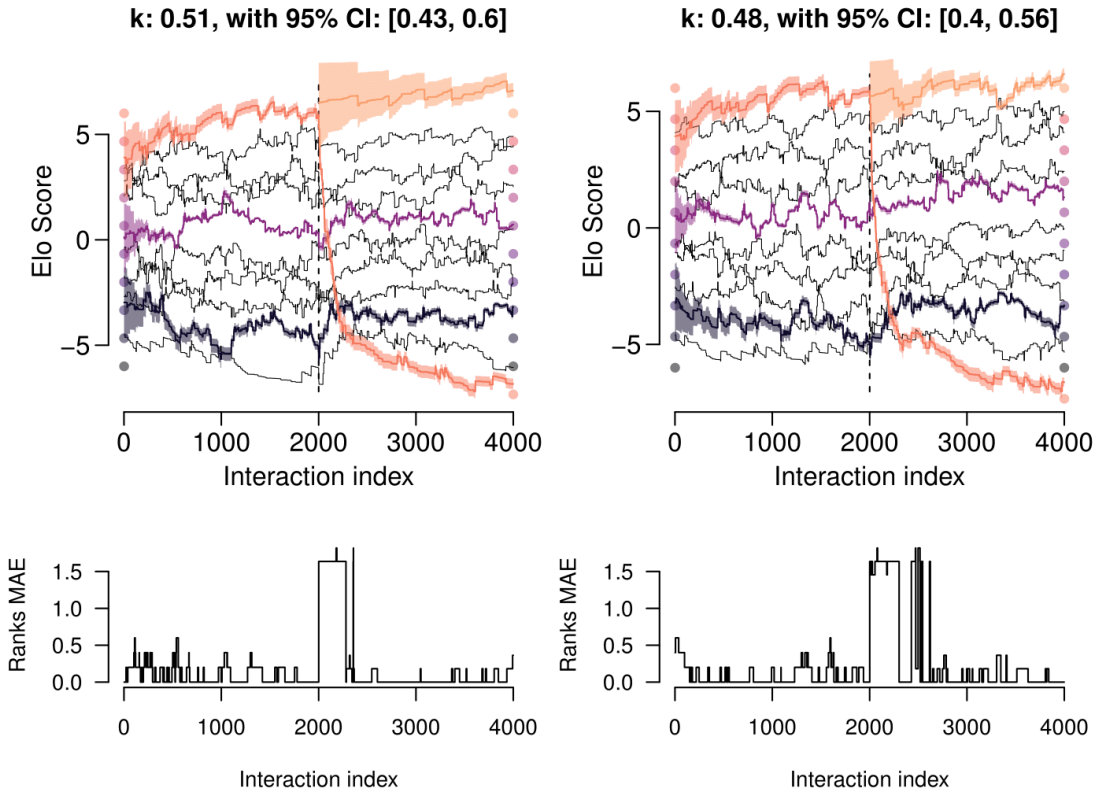

Figure 5: Results of two runs for simulation 3A ('external takeover'). In the top panels the two periods are separated visually by a vertical dotted line; ten individuals are present in the first period and eleven individuals are present in the second period. Each color represents a different individual and the line type indicates that the individual changed scores (dashed, N = 2) or retained consistent scores (solid, $N = 9$). Filled circles at the outer extremities of the two periods indicate the individuals' underlying Elo scores. The bottom panels show the evolution of the ranks' mean absolute error (MAE) across the 4000 interactions.

Simulation 3B–Coalitionary Leap Figure 6 shows the results for two different simulation runs. Although order and position remained constant within the 5 subgroups, the calculated Elo scores showed a high degree of stochasticity even 2000 interactions following the change in underlying scores.

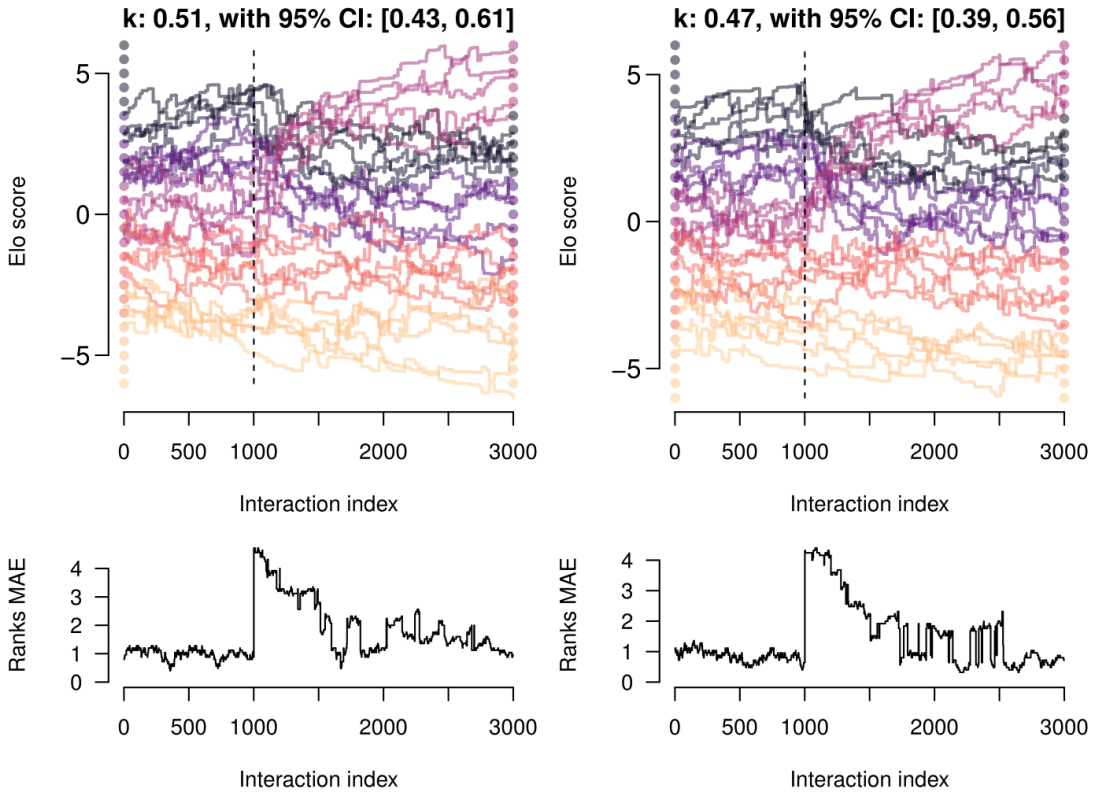

Figure 6: Results of two runs for simulation 3B ('coalitionary leap'). Filled circles at the outer extremities of the two periods indicate the individuals' underlying Elo scores and the dashed vertical line separates the two periods. The bottom panels show the evolution of the ranks' mean absolute error (MAE) across the 3000 interactions.

Simulation 3C–Mortality-Instability-Recovery Figure 7 shows the results for two different simulation runs depicting such a "mortality-instability-recovery" scenario. Here, we found that in spite of a prolonged period of social disruption, Elo scores (and the mean absolute error) returned within about the first 100 interactions to the levels before the event.

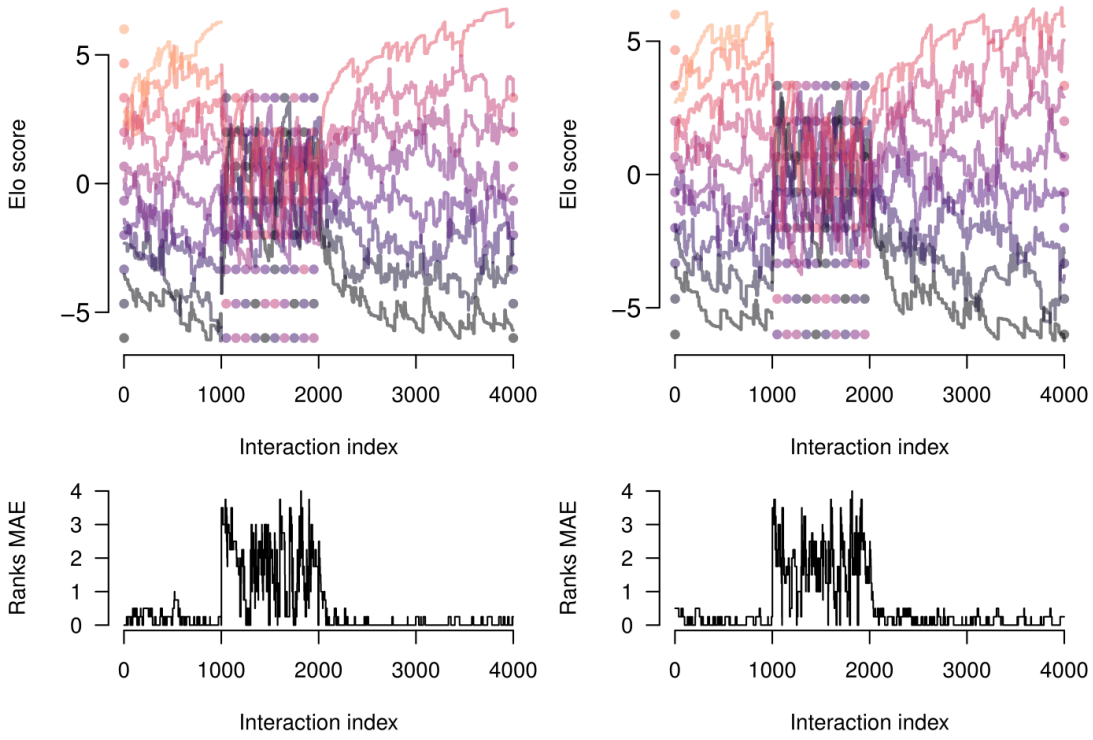

Figure 7: Results of two runs for simulation 3C ("mortality-instability-recovery"). Filled circles indicate the individuals' underlying Elo scores. The bottom panels show the evolution of the ranks' mean absolute error (MAE) across the 4000 interactions. The Ranks MAE calculation only incorporates the 8 individuals present in all three periods.

4 Discussion

Elo scores allow for the generation of dominance hierarchies while considering temporal dynamics and sequential events (Neumann et al., 2011). These scores therefore add the critical temporal dimension to the analysis of agonistic interactions between individuals. Our research extends previous applications in three critical ways: Firstly, we demonstrate that partial pooling within a Bayesian Inference approach provides better estimates of the starting scores than a ML approach. Secondly, the BI approach provides valuable additional information with regard to the assessment of uncertainty of the estimates. Third, simulation approaches allowed us to investigate the effects of different boundary conditions on Elo scores. Specifically, we found that (not surprisingly) in an unbalanced design, a small sample size increases the uncertainty of the estimate; that it may take a substantial amount of interactions (here: $N = 200$) after an external takeover until the hierarchy can be interpreted in a meaningful way; and that coalitionary leaps of whole groups may create havoc in the estimation of Elo scores, as it may take an excessive amount of interactions (in our simulation: 2000) until the hierarchy could be inferred with certainty. Finally, we found that the removal of an individual followed by a phase of instability may recover quickly.

Partial pooling appeared to be clearly superior to separate coefficient estimation. One compelling example is the ranking of baseball players by their estimated batting abilities (Carpenter et al., 2017a). Although the BI approach yielded more robust results than the ML approach for the assessment of starting scores, there was no substantial advantage of one method over the other in terms of the point estimates. Because the BI approach provides additional important additional information with regard to uncertainty, we suggest that this approach should be preferred.

The application of Bayesian inference estimation to a previously published data-set (Foerster et al., 2016b) allowed us to determine the extent of uncertainty in point estimates generated from real world social interactions. The assessment of uncertainty is critical for when drawing conclusions about current rank hierarchies and social dynam-

ics. For instance, what may look like a clear dominance hierarchy when only one point in time is considered, may vary wildly from time point to time point. In such cases, it is safer to abstain from the classification of individuals with regard to specific ranks.

The simulation scenarios allowed us to assess the consequences of specific events, while having full knowledge about the underlying social interactions. In our simulations, we found various causes for uncertainty in the reliability of Elo scores. When the true values were known, Elo scores varied with the proportion/number of interactions individuals contributed to the data-set (balance), the occurrence of social change (stability) and the type of change that occurred (e.g. number and position of individuals leaving the hierarchy). As in many other cases, the number of interactions considered is critical. Importantly, even when interactions were equally balanced and hierarchies were stable, it may take large numbers of interactions in order for Elo scores to approach the "true" underlying scores. Although larger sample sizes generally seem to be preferable, we are unable to make general recommendations regarding the number of interactions in relation to the number of dyads. We do, however, recommend using a simulation approach that is based on the data available to check the sensitivity and stability of the results.

4.1 Inclusion/exclusion of data

In the literature, it appears that individuals with comparatively sparse numbers of interactions may be excluded when assessing dominance hierarchies (e.g. Seyfarth et al. (2012); Foerster et al. (2016a)). This strategy may potentially bias the results since the number of observed interactions per individual may be depend on the underlying hierarchy position, resulting in apparent instability in certain regions of the dominance hierarchy (for example, simulation 2).

Moreover, the exclusion of an individual removes the information from winning/loosing of each of this individual's opponents. Although ML approaches dictate the exclusion of subjects with one type of interaction solely because models could not be fitted otherwise, this is highly problematic. Effectively, if one was interested in a society with a subject that was in the top position throughout the study period (no losses), this individual

would have to be excluded, and valuable information would be lost. This may even lead to a chain reaction where the second ranking subject whose losses against the former dominant subject are now removed from the analysis, also needs to be removed because it now only sports wins also.

All information, seemingly minute in detail, adds to the knowledge of social dynamics in the group and can thus assist in illuminating hierarchical fluctuations. The BI approach, which is able to deal with wins/losses only situation, clearly is the method of choice under such circumstances. We therefore discourage the exclusion of subjects, or at least strongly recommend running comparative analyses including and excluding this/these subjects, to assess the consequences of this step.

4.2 Uncertainty informs about the characteristics of the society

The extent to which we see uncertainty in Elo scores adds additional knowledge to our perception of the characteristics of a society. For instance, Barbary macaque (*Macaca sylvanus*) males have a rank hierarchy with clearly identifiable top and bottom subjects, while the rank for animals ranging in the middle is harder to discern (Henkel et al., 2010). For these animals, the uncertainty with which the rank can be estimated reflects precisely this fact, which would be misrepresented by a linear rank hierarchy only. Uncertainty estimates allow us to assess both the changes in rank order and the overlap of individuals with adjacent rank. In some societies, particularly those which are more egalitarian, the attribution of a fixed rank position to each individual may not be realistic. Rather, making probabilistic statements by assigning individuals to the dominant or subordinate half (or dominant, middle, and subordinate third) of the group with different (posterior) probabilities seems more appropriate.

4.3 The sensitivity of the Elo method and its drawbacks

On the one hand, Elo rating is sensitive and able to detect changes in social hierarchies resulting from social instability and demographic change. On the other hand, individual scores may be highly volatile, even in scenarios in which the hierarchy is actually stable

(e.g. the first periods in Figures 5, 6, and 7). Likely, this is due to the fact that individual A's scores are indirectly affected through the interaction of individual A's social partners (e.g., individuals B and C). Although this allows us to pool social knowledge, and gain information about individuals with low interaction rates, the inherent stochasticity in the Elo scores of others may increase the stochasticity of third parties (in our example, subject A). Future studies should examine to which extent group size and transitivity influence the sensitivity of Elo.

With the use of different simulation scenarios, we demonstrate that it takes time for a change in the underlying hierarchy to be reflected in the Elo scores. This may depend upon the extent to which the system is already established: the closer to the beginning the change occurs, the higher is the uncertainty in the Elo scores. The volatility of the system/degree of hierarchy stability may also influence the time it takes for the Elo scores to reflect the underlying change.

4.4 Summary and conclusion

Elo rating has been developed in the sphere of human sporting competition in which we can watch the contests between individuals and is a tool which has recently been applied to assessing dominance hierarchies in non-human animals. However, some of the challenges we face as those who study animal social dominance is a direct result of the type and complexities of the contests we observe and the limited window by which they are viewed. Thus far, Elo rating has proven to be a powerful tool in our understanding of temporal fluctuations in dominance hierarchies. However, the role of group size, interaction rate, sample size (i.e. number of interactions) and temporal sequence of events remain unclear. Here, we have taken the first steps in trying to understand these factors by presenting a tool, Bayesian inference, which allows us to see the variability in uncertainty around Elo scores and some of the ways in which uncertainty can vary in simple simulated dominance hierarchies.

References

Adams, M. J., Majolo, B., Ostner, J., Schülke, O., De Marco, A., Thierry, B., Engelhardt, A., Widdig, A., Gerald, M. S., and Weiss, A. (2015). Personality structure and social style in macaques. *Journal of Personality and Social Psychology*, 109(2):338–353.

Albers, P. C. H. and de Vries, H. (2001). Elo-rating as a tool in the sequential estimation of dominance strengths. *Animal Behaviour*, 61:489–495.

Arseneau-Robar, T. J. M., Taucher, A. L., Schnider, A. B., van Schaik, C. P., and Willems, E. P. (2017). Intra-and interindividual differences in the costs and benefits of intergroup aggression in female vervet monkeys. *Animal Behaviour*, 123:129–137.

Cafazzo, S., Valsecchi, P., Bonanni, R., and Natoli, E. (2010). Dominance in relation to age, sex, and competitive contexts in a group of free-ranging domestic dogs. *Behavioral Ecology*, 21(3):443–455.

Carpenter, B., Gabry, J., and Goodrich, B. (2017a). Hierarchical Partial Pooling for Repeated Binary Trials. URL https://cran.r-project.org/web/packages/rstanarm/vignettes/pooling.html [Accessed 16 June 2017].

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017b). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Cheney, D. L., Seyfarth, R. M., Fischer, J., Beehner, J., Bergman, T., Johnson, S., Kitchen, D. M., Palombit, R., Rendall, D., and Silk, J. B. (2004). Factors affecting reproduction and mortality among baboons in the Okavango Delta, Botswana. *International Journal of Primatology*, 25(2):401–428.

David, H. A. (1987). Ranking from Unbalanced Paired-Comparison Data. *Biometrika*, 74(2):432–436.

David, H. A. (1988). *The Method of Paired Comparisons*. Oxford University Press, New York.

de Vries, H. (1998). Finding a dominance order most consistent with a linear hierarchy: a new procedure and review. *Animal Behaviour*, 55(4):827–843.

Elo, A. (1961). The new U.S.C.F. rating system. *Chess Life*, 16:160–161.

Elo, A. (1978). *The Rating of Chess Players, Past and Present*. Arco, New York.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20(2):203–219.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer, Heidelberg.

Foerster, S., Franz, M., Murray, C., Gilby, I., Feldblum, J., Walker, K., and Pusey, A. (2016a). Chimpanzee females queue but males compete for social status. *Scientific Reports*, 6.

Foerster, S., Franz, M., Murray, C., Gilby, I., Feldblum, J., Walker, K., and Pusey, A. (2016b). Data from: Chimpanzee females queue but males compete for social status. *Dryad Digital Repository*. URL http://dx.doi.org/10.5061/dryad.r4g74 [Accessed 5 January 2017].

Franz, M., McLean, E., Tung, J., Altmann, J., and Alberts, S. C. (2015). Self-organizing dominance hierarchies in a wild primate population. *Proceedings of the Royal Society B*, 282.

Ganslosser and Dellert (1997). Experimental alterations of food distribution in two species of captive equids (*Equus burchelli* and *E. hemionus kulan*). *Ecology, Ethology & Evolution*, 9(1):1–17.

Gelman, A., Hill, J., and Yajima, M. (2016). Why We (Usually) Dont Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5:189–211.

Haunhorst, C. B., Heesen, M., Ostner, J., and Schülke, O. (2017). Social bonds with males lower the costs of competition for wild female Assamese macaques. *Animal Behaviour*, 125:51–60.

Henkel, S., Heistermann, M., and Fischer, J. (2010). Infants as costly social tools in male Barbary macaque networks. *Animal Behaviour*, 79(6):11991204.

Johnson, E. T., Snyder-Mackler, N., Beehner, J. C., and Bergman, T. J. (2014). Kinship and Dominance Rank Influence the Strength of Social Bonds in Female Geladas (*Theropithecus gelada*). *International Journal of Primatology*, 35(1):288–304.

Kaburu, S. S., Inoue, S., and Newton-Fisher, N. E. (2013). Death of the Alpha: Within-Community Lethal Violence Among Chimpanzees of the Mahale Mountains National Park. *American Journal of Primatology*, 75(8):789–797.

Kalbitzer, U., Heistermann, M., Cheney, D., Seyfarth, R., and Fischer, J. (2015). Social behavior and patterns of testosterone and glucocorticoid levels differ between male chacma and Guinea baboons. *Hormones and Behavior*, 75:100–110.

Marty, P. R., Hodges, K., Agil, M., and Engelhardt, A. (2015). Alpha male replacements and delayed dispersal in crested macaques (*Macaca nigra*). *American Journal of Primatology*.

Neumann, C., Duboscq, J., Dubuc, C., Ginting, A., Irwan, A. M., Agil, M., Widdig, A., and Engelhardt, A. (2011). Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo-rating. *Animal Behaviour*, 82(4):911–921.

Proops, L., Burden, F., and Osthaus, B. (2012). Social relations in a mixed group of mules, ponies and donkeys reflect differences in equid type. *Behavioural Processes*, 90(3):337–342.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ [Accessed 24 October 2016].

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278.

Schusterman, R. J. and Dawson, R. G. (1968). Barking, dominance, and territoriality in male sea lions. *Science*, 160(3826):434–436.

Seyfarth, R. M., Silk, J. B., and Cheney, D. L. (2012). Variation in personality and fitness in wild female baboons. *Proceedings of the National Academy of Sciences*, 109(42):16980–16985.

Tilson, R. L. and Hamilton, W. J. (1984). Social dominance and feeding patterns of spotted hyaenas. *Animal Behaviour*, 32(3):715–724.

Tilson, R. L., Sweeny, K. A., Binczik, G. A., and Reindl, N. J. (1988). Buddies and bullies: social structure of a bachelor group of Przewalski horses. *Applied Animal Behaviour Science*, 21(1-2):169–185.

Tung, J., Archie, E. A., Altmann, J., and Alberts, S. C. (2016). Cumulative early life adversity predicts longevity in wild baboons. *Nature Communications*, 7.

Weiß, B. M., Kotrschal, K., and Foerster, K. (2011). A longitudinal study of dominance and aggression in greylag geese (*Anser anser*). *Behavioral Ecology*, 22(3):616–624.

Zhu, P., Ren, B., Garber, P. A., Xia, F., Grueter, C. C., and Li, M. (2016). Aiming low: A resident male's rank predicts takeover success by challenging males in Yunnan snub-nosed monkeys. *American Journal of Primatology*, 78(9):974–982.

S1 Stan implementation of Bayesian estimation of starting scores and k

The following Stan (Carpenter et al., 2017b) code implements the Bayesian estimation of starting scores and winning/loosing tax coefficient k as introduced by this manuscript. For convenience, the observations $\mathbf{x}$ were stored in the form of two vectors $\mathbf{A}$ and $\mathbf{B}$, each of length n, where the integer values in $\mathbf{A}$ store the index of the interaction-specific winning individual, and the integer values in $\mathbf{A}$ store the index of the interaction-specific loosing individual. The remaining input arguments (see the `data` block) are the number of encounters `N`, the number of individuals `K`, a response vector `y` always taking on value 1 (needed for the probabilistic formulation of the likelihood as a Bernoulli random variable; always 1 since it's always the individual from `Ai` winning the encounter), and the Elo score difference factor `delta`.

```
functions {
  real[] ProbFunction(real[] EloStart, real k, matrix presence, int N, int K,
                      int[] Ai, int[] Bi, real diff_f) {
    real result[N];
    real toAdd;
    vector[K] EloNow;
    for (j in 1:K) {
      EloNow[j] = EloStart[j];
    }
    for (i in 1:N) {
      // centering:
      EloNow = EloNow - dot_product(row(presence,i),EloNow)/sum(row(presence,i));
      // likelihood contribution:
      result[i] = 1/(1 + exp(diff_f * (EloNow[Bi[i]] - EloNow[Ai[i]])));
      // update addend:
      toAdd = (1 - result[i]) * k;
      // update:
      EloNow[Ai[i]] = EloNow[Ai[i]] + toAdd;
```

```
      EloNow[Bi[i]] = EloNow[Bi[i]] - toAdd;
      }
    return result;
  }
}
data {
  int<lower=1> N; // number of encounters
  int<lower=1> K; // number of individuals
  int<lower=1> Ai[N]; // winner's index
  int<lower=1> Bi[N]; // losers's index
  matrix[N, K] presence;
  int<lower=0> y[N]; // always 1
  real<lower=0> diff_f; // Elo Score difference factor
}
parameters {
  real EloStart_raw[K];
  real<lower=0.0> k_raw;
  real<lower=0.0> sigma_raw;
}
transformed parameters {
  real EloStart[K];
  real<lower=0.0> k;
  for (i in 1:K) {
      EloStart[i] = EloStart_raw[i] - mean(EloStart_raw);
    }
  for (i in 1:K) {
      EloStart[i] = EloStart[i]/diff_f;
    }
  k = k_raw/diff_f;
  }
model {
  k_raw ~ normal(0, 1);
  sigma_raw ~ normal(0, 1);
```

```
  EloStart_raw ~ normal(0, sigma_raw);
  y ~ bernoulli(ProbFunction(EloStart, k, presence, N, K, Ai, Bi, diff_f));
  }
generated quantities{
  real<lower=0.0> sigma;
  sigma = sigma_raw/diff_f;
}
```

S2 Update factor is equal to Logistic function

Using basic algebra, one can show that the update factor used in the definition of the Elo score is a "rather complicated written" form of the logistic function:

$$\begin{aligned} 1 - \frac{1}{1+\exp\left(-0.01\left(\mathrm{Elo}_{A,j-1} - \mathrm{Elo}_{B,j-1}\right)\right)} &= 1 - \frac{1}{1+\exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)} \\ &= 1 - \frac{1}{1+\exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)}. \end{aligned}$$

With $x := 0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)$, we further get:

$$\begin{aligned} 1 - \frac{1}{1+\exp\left(x\right)} &= \frac{1+\exp\left(x\right)}{1+\exp\left(x\right)} - \frac{1}{1+\exp\left(x\right)} \\ &= \frac{1+\exp\left(x\right)-1}{1+\exp\left(x\right)} \\ &= \frac{\exp\left(x\right)}{1+\exp\left(x\right)}, \end{aligned}$$

and therefore:

$$\begin{aligned} 1 - \frac{1}{1+\exp\left(-0.01\left(\mathrm{Elo}_{A,j-1} - \mathrm{Elo}_{B,j-1}\right)\right)} &= \frac{\exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)}{1+\exp\left(0.01\left(\mathrm{Elo}_{B,j-1} - \mathrm{Elo}_{A,j-1}\right)\right)} \\ &= \begin{cases} < 0.5, & \text{if } \mathrm{Elo}_{A,j-1} > \mathrm{Elo}_{B,j-1}, \\ 0.5, & \text{if } \mathrm{Elo}_{A,j-1} = \mathrm{Elo}_{B,j-1}, \\ > 0.5, & \text{if } \mathrm{Elo}_{A,j-1} < \mathrm{Elo}_{B,j-1}\,. \end{cases} \end{aligned}$$

S3 Standard deviation equivalence by varying Elo score difference factor δ

Let $X = \text{Elo}_{A,j-1}$, and $Y = \text{Elo}_{B,j-1}$, and further $\tilde{X} = \delta X$, $\tilde{Y} = \delta Y$, with $0 < \delta \neq 1$. In the case of equal variances $\text{Var}(X) = \text{Var}(Y) = \sigma^2$, $\text{Var}\left(\tilde{X}\right) = \text{Var}\left(\tilde{Y}\right) = \tilde{\sigma}^2$, and independence, ie. $\text{Cov}(X, Y) = \text{Cov}\left(\tilde{X}, \tilde{Y}\right) = 0$:

$$\text{Var}(X - Y) = 2\sigma^2,$$

$$\text{Var}\left(\tilde{X} - \tilde{Y}\right) = \delta^2 \text{Var}(X) + \delta^2 \text{Var}(Y) = 2\delta^2\sigma^2$$

Since $\text{Var}\left(\tilde{X} - \tilde{Y}\right) = 2\tilde{\sigma}^2$, we get:

$$\tilde{\sigma}^2 = \delta^2\sigma^2 \Leftrightarrow \tilde{\sigma} = \delta\sigma.$$

By re-defining $\delta = 1$ in the winning/losing probability, differences in Elo scores that are 0.01 as large as in the original definition with $\delta = 0.01$ lead to the same winning/losing probabilities. Therefore, k will here also be only 0.01 as large as in the original definition, i.e. also the tax coefficient directly scales with δ:

$$\tilde{k} = \delta \cdot k.$$