Research Article

Ben Powell*

Generalizing the Elo rating system for multiplayer games and races: why endurance is better than speed

https://doi.org/10.1515/sample-YYYY-XXXX, Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: We introduce a non-standard generalization of the Elo rating system for competitions involving two or more participants. The new system can be understood as an on-line estimation algorithm for the parameters of a Plackett-Luce model which can be used to make probabilistic forecasts for the results of future competitions. The system's distinguishing feature is the way it treats competitions as sequences of elimination-type rounds that sequentially identify the worst competitors rather than sequences of selection-type rounds that identify the best. The significance of this important modelling choice is discussed and its consequences are explored. Finally, our generalized Elo system's predictive power is demonstrated using data from Formula One racing.

Keywords: Elo ratings, probabilistic forecasting, rank data, statistical modelling, time series.

1 Introduction

1.1 What is the Elo rating system and how can it be generalized?

Arpad Elo's rating system for quantifying the relative skill of chess players has been extremely popular and successful since its introduction in Elo (1967). It has proved simple enough for committed sports fans to engage with without requiring extensive mathematical training. At the same time it has proved sophisticated enough to provide predictions comparable to those from more complex models and those incorporating more contextual data (see Hvattum and Arntzen (2010)).

The conventional Elo system allocates to a competitor, labelled with subscript i, a score $\hat{R}_i$ that is understood as an approximate quantification of her match-winning ability. We then introduce a binary random variable '$X(\text{i beats j})$' taking value one if i beats competitor j, and zero otherwise. The win-probability for i is the prior expectation computed as

$$\mathbb{E}[X(\text{i beats j})] = \frac{e^{\hat{R}_i}}{e^{\hat{R}_i} + e^{\hat{R}_j}}, \tag{1}$$

which, given its introduction to the literature in Bradley and Terry (1952), is commonly referred to as a Bradley-Terry model for the variable $X(\text{i beats j})$. After a match is completed the players' scores are adjusted in the direction of the difference between the actual and expected values of $X(\text{i beats j})$. For competitor i, for example, her new adjusted score is

$$\hat{R}_i \leftarrow \hat{R}_i + k\left\{X(\text{i beats j}) - \mathbb{E}[X(\text{i beats j})]\right\}. \tag{2}$$

where k is some fixed constant and the leftwards-pointing arrow denotes the assignment of a new numerical value to a variable. Our notation here has been chosen to facilitate the interpretation of the Elo score

***Corresponding author: Ben Powell,** University of York, Department of Mathematics, York, Uk, e-mail: ben.powell@york.ac.uk

adjustment equations as approximate Bayesian updates to estimates of true but unknown ability parameters R_i. A comprehensive framing of the Elo system as a Bayesian calculation can be found in Glickman (1999), Ingram (2021) and Ebtekar and Liu (2021) for example. While a Bayesian formulation is not strictly necessary to motivate the Elo system, it will help us explain and justify our efforts to generalize it.

Variants of the Elo system have been employed in official and unofficial capacities to model the results of many physical sports such as American football (Silver (2014)), basketball (Silver and Fischer-Baum (2015)), ice hockey (Dabadghao and Vaziri (2022)), soccer (Lasek et al. (2013)) and tennis (McHale and Morton (2011)). Its application to e-sports has been arguably even more important, where it serves to quickly match huge numbers competitors of approximately equal ability in order to maximize the games' competitiveness. Although the system in its original form was designed to learn from and predict the results of pairwise comparisons, several attempts have been made to adapt it for multiplayer games and races. A notable contribution example comes from Moore et al. (2018), who choose to interpret a race with n competitors as a set of $n(n-1)/2$ independent head-to-head races which are each won by the faster racer. This modelling strategy, which is formalized in Mallows (1957) and used more recently in Weng and Lin (2011), allows for the direct use of the standard Elo machinery. However, it can be seen to effectively over-count the significance of race results involving a large number of competitors. The intuition here is that is unfair to repeatedly punish (reward) a competitor for every rival they lose (win) to when they have a single bad (good) race performance. More formally, we can think about competitors A, B and C who finish a race in this order. Once we learn that A beats B and B beats C, and we adjust the competitors' scores to better reflect these pairwise relationships, we already know that A has beaten C. We get no new information when we are told A has beaten C so it does not make sense to adjust the scores again. A different, principled, explicitly model-based approach to generalizing the Elo system is clearly needed to accommodate the dependence between the pairwise results and to avoid the over-counting of dependent events.

A particularly attractive class of models for doing this involves each of a competition's competitors simultaneously drawing 'performance scores' from their own distributions, whose location parameters are identified with a notion of ability or strength. The performance scores themselves are understood to be unobserved but their order determines a ranking, which is observed. In Yellott Jr (1977), Yellot showed that the class of shifted Gumbel distributions is the only one leading to the tractable Plackett-Luce model for competition winners, which we will meet properly in Section 3. The exponentials of the negated Gumbel random variables are Exponential random variables, which now vary according to their scales rather than locations as we switch between competitors. Concentrating on the Exponential variables helps us imagine the performance scores as waiting times whose relevance to real competitions is, for many important examples, more direct since they can be identified with times until match-winning maneuvers or match-losing errors.

Deviating from the Gumbel/Exponential model is possible but comes at considerable cost. The TrueSkill$^{\text{TM}}$ system, described in Herbrich et al. (2006) and designed with principal application to e-sports, is a proprietary algorithm for performing approximate Bayesian inference and prediction given Normal performance scores and Normal priors on their means. Approximating the relevant integrals here turns out to be a highly challenging endeavour that is tackled with an approximate message-passing algorithm over a graph encoding the parameter dependency structure. While technically impressive, this rating system is neither easy to reproduce nor easy to intuit. We argue that the latter problem is particularly important. We require an intuitive model because we want to use it to organize our thoughts about a competition in an intelligible way. We also want a model we can inspect, critique and explain to others because we want people to invest in it. This could mean competition organisers and competitors using the model to recognize performances worthy of reward, or gamblers using it to inform wagers.

So, sticking with the uniquely tractable Gumbel/Exponential model for performance scores, and noting the distribution's asymmetry, we are left with one seemingly innocuous question: should large values or small values determine the winners of a competition? The remainder of this paper essentially argues that large values should. The implications of this choice are discussed in theory and investigated in practice. They lead us to recommending a particular non-standard orientation of the Plackett-Luce model and a

corresponding generalization of the Elo system (the 'endure-Elo' system), which we argue is most relevant to competitive sports.

1.2 Is there a need to generalize the Elo rating system to multiplayer competitions?

A key characteristic of the event results that the Elo system serves to interpret is that they are only informative for the relative abilities of competitors. This seems obvious for zero-sum two-player games like chess or soccer, but less so for race-type competitions in which competitor interactions are less significant to the results. For highly standardized athletics events, we agree that absolute measures of a competitor's performance (i.e. finishing times) certainly do exist and tend to be relatively consistent from one event to another despite changes in venue and the set of fellow competitors. Here the existence of these dependable absolute metrics means that a scoring system based only on relative performance is arguably redundant. For sports like motor racing, however, the variation between tracks is such that finishing times for one event are not, by themselves, useful for predicting finishing times for another event. The relative performance (i.e. ranks) of the racers can be expected to remain consistent, and therefore useful for prediction, because the track effects are, to an extent, cancelled out.

An Elo system for races is also potentially useful when considering tournaments of two-player games. While the standard Elo system is undeniably useful for quantifying the relative performance of competitors in individual games, it is not entirely clear that the individual performances contribute independently to tournament success. For example, when the long-term objective of a soccer club is to accumulate enough points to be promoted into a higher league, or to avoid relegation into a lower one, their performance in a later match may be affected by those that preceded it and, more specifically, by the number of points scored from them. The point being that the tournament is perhaps better seen holistically as a race to accumulate points rather than a sequence of independent events, and that future tournament performances are better predicted from past tournament performances than from past match performances (divorced from the context of the tournament) alone.

1.3 Is there a prototypical form of multiplayer competition?

At this point it is also useful to mention two distinct types of multiplayer competition. The first are races (tests of speed) in which the first competitor to finish is the overall winner. The second are knock-out events (tests of endurance) in which the last competitor to finish is the overall winner. As we will see, both types of competition can be modelled using the same probabilistic devices by characterizing competitor performances in terms of time until success or time until failure and, correspondingly, by quantifying competitor abilities in terms of success rates or failure rates. Certain implications of the resulting models relating to how good and bad performances are interpreted, however, are reversed. The two modelling choices lead to two possible ways of generalizing the Elo rating system that we refer to as speed-Elo and endure-Elo. We will argue that the endure-Elo system is the most appropriate for most sporting events and hence deserves to be designated the preferred generalization.

Deciding whether to classify an actual race as a test of speed or a test of endurance is not necessarily as easy as it seems. Formula One (F1) racing, which we use in later sections as an example application for our methodology, is interesting in this regard because it arguably combines both sorts of competition. Here F1 drivers obviously compete to get the fastest overall race time but they also need to survive very many laps without completely or partially degrading their vehicle. The distinction between speed and endurance competitions is important when modelling competitors' performance probabilistically because the maximum and minimum of a set of random variables (and the identities of the variables that attain them) are liable to behave in fundamentally different ways. This is particularly true of exponential random variables.

Our theoretical arguments hinge on the assertion that the exponential distribution is better suited to describing failure times than finishing times. This is the case since the distribution effectively describes the time of a single competitor's first failure or finish event given that such events occur with an infinitesimally small but constant probability at any time. This assumption of a constant event rate is clearly not realistic for finishing events in most real sports but could be realistic for failure events, which we can also think of as errors that lead to unrecoverable set-backs in a race rather than complete destruction. Despite the exponential distribution being poorly suited to modelling times of successes rather than failures, it is commonly used (either explicitly or implicitly) for this purpose (in Plackett (1975), Beggs et al. (1981) and Gormley and Murphy (2008) for example). The reason for this appears to be that computing the distribution for the identity of the smallest of a set of exponential random variables is considerably easier than doing so for the largest. We surmise that modellers are inclined to pick the orientation of their ranks in order to make computing win-probabilities easier because win-probabilities are typically of primary interest. We encounter expressions for the probabilities for the smallest and largest variable in equations (9) and (60) respectively. The latter certainly takes more effort to compute as the number of competitors increases. Nevertheless, we consider the cost to be worthwhile for moderately large numbers of competitors (approximately 20), as in the example studied in Section 4.2, after which we can still make use of approximations as discussed in Appendix section A.

The differences between speed and endurance competitions are not just academic considerations for statistical modellers. They have qualitative effects on the resulting quantifications of competitor ability and on predictions for competition winners. More concretely, if a failure can happen at any time then competitors who fail early on in a competition may just have been unlucky. Accordingly, we should not read too much into these results. A competitor who outlasts all her competitors, however, is very unlikely to do so unless she had decreased her failure rate to an extremely low level. These ideas lead us to a generalized 'endure-Elo' system that is most sensitive to the best performances in a competition rather than the worst ones, and not the other way around as would be the case when finishing times are assumed to be exponentially distributed. An increasing level of discrimination among the best performances can also be seen in established rating systems without model-based justifications like the F1 Championship point system, suggesting that the same sort of reasoning has taken place among the sport's governing body. The results of endurance competitions being 'less random' for the best performing competitors is also reflected in the fact that the best competitors are more likely to win endurance competitions than speed competitions. In Appendix 3.5 we provide some theory and examples to express these ideas more formally.

We ought to emphasize here that these ideas are not unknown in the relevant academic communities, although they arguably remain underappreciated. Notably, Graves et al. (2003) reach the same conclusion about models for speed and endurance events based on the observation that, for the (reversed Plackett-Luce) endurance model, the observed Fisher Information is greatest for competitors who do the best in a competition and least for those who do the worst; and that this property is reversed for the (conventional Plackett-Luce) speed model. In this sense the estimation of ability parameters for the endurance model is informed predominantly by the drivers' best performances rather than their worst. Henderson and Kirrane (2018), who provide a valuable comparison study of variants of the Plackett-Luce model applied to F1 data, also question how informative the worst results in a race are for estimates of racers' abilities. Most relevant to our current work is their experimentation with a truncated form of the conventional Plackett-Luce model that effectively treats all finishing positions beyond a certain point as being equivalent. They find that the truncated (conventional Plackett-Luce) speed model outperforms the untruncated version when it comes to predicting the top positions in races, but is itself outperformed by the (reversed Plackett-Luce) endurance model.

2 Summary of endure-Elo scoring rules

In this section we present a concise summary of the endure-Elo scoring rules. The model motivating the rules is discussed in Section 3.

2.1 Base version

Our endure-Elo scoring system is premised on the idea that competitions in which all m competitors receive a rank are to be understood as a sequence of $m-1$ independent knock-out rounds in which the first competitor to make a mistake is eliminated and allocated the worst as yet untaken rank. Given endure-Elo scores $\hat{R}_{i,t}$ for a set of competitors Q at time t, we calculate the elimination and survival probabilities for a round in which they take part according to

$$P(\text{i eliminated in round}) = \frac{e^{-\hat{R}_{i,t}}}{\sum_{j\in Q} e^{-\hat{R}_{j,t}}}, \qquad P(\text{i survives round}) = 1 - P(\text{i eliminated in round}). \tag{3}$$

The quantities $\lambda_{i,t} = e^{-\hat{R}_{i,t}}$ above are to be understood as approximate error or failure rates.

When a competitor loses a round and is eliminated from the competition we update all the endure-Elo scores for competitors in that round according to

$$\hat{R}_{i,t} \leftarrow \hat{R}_{i,t} + k\left[\mathbb{I}(\text{i survives round}) - P(\text{i survives round})\right], \tag{4}$$

where k is a parameter that can be adjusted to maximize the product of probabilities (3) for the results of previously observed competitions. When competitions involve just two competitors (and so consist of a single elimination round) and the basic endure-Elo rating system reduces to the conventional Elo rating system.

2.2 Extended version

A more sophisticated treatment of the k-factors appearing in (4) involves recognizing them as variances for the true but unknown values that the $\hat{R}_{i,t}$ serve to estimate. Doing so motivates an extended version of the update procedure so that

$$k_{i,t}^{-1} \leftarrow k_{i,t}^{-1} + P(\text{i survives round})\left[1 - P(\text{i survives round})\right], \tag{5}$$

$$\hat{R}_{i,t} \leftarrow \hat{R}_{i,t} + k_{i,t}\left[\mathbb{I}(\text{i survives round}) - P(\text{i survives round})\right], \tag{6}$$

where the subscripted $k_{i,t}$ are effectively competitor-specifc k-factors.

If h time increments pass before the next competition we suggest implementing an information-discounting or forgetting step

$$k_{i,t+h} \leftarrow k_{i,t} + (1-\phi^{2h})(k_{i,\infty} - k_{i,t}), \tag{7}$$

$$\hat{R}_{i,t+h} \leftarrow \phi^h \hat{R}_{i,t}, \tag{8}$$

where ϕ is a discounting factor and $k_{i,\infty}$ is an asymptotic variance that quantifies the uncertainty for the ability of a competitor for whom we have no historical data.

3 A model to motivate endure-Elo scoring

3.1 The base model

Consider a competition for survival between competitors whose failure times are independently exponentially distributed. Each competitor, labelled with subscript $i \in Q = \{1, \ldots, m\}$, has a failure rate parameter

$\lambda_i > 0$. It is well known that in such a situation the probability of competitor k failing first is

$$P(k \text{ fails first}) = \frac{\lambda_k}{\sum_{a \in Q} \lambda_a}, \tag{9}$$

and that the time of the first failure is also exponentially distributed with rate parameter $\sum_{i \in Q} \lambda_i$.

The independence and the memoryless properties of the exponential distribution mean that once the first competitor fails, deciding the next competitor to fail effectively involves another independent competition involving only the remaining competitors. This means, for example, that if competitor i fails first in the first round (and so comes last place in the whole competition), the probability of competitor j failing first in the second round (and so coming second-to-last in the whole competition) is

$$P(j \text{ fails second in whole competition} \mid i \text{ fails first in whole competition}) = \frac{\lambda_j}{\sum_{a \in Q \setminus i} \lambda_a}. \tag{10}$$

This argument can be iterated, allowing us to compute the probability for any race result

$$\begin{aligned} P(\text{competition result}) =& P(\text{i fails first}) \times P(\text{j fails second}| \text{ i fails first}) \times & (11)\\ & P(\text{k fails third}| \text{ i fails first and j fails second}) \times & (12)\\ & \ldots \times P(\text{z fails last}|\text{all n-1 previous positions taken}) & (13)\\ =& \frac{\lambda_i}{\sum_{a \in Q} \lambda_a} \times \frac{\lambda_j}{\sum_{a \in Q \setminus i} \lambda_a} \times \frac{\lambda_k}{\sum_{a \in Q \setminus \{i,j\}} \lambda_a} \times \ldots \times 1. & (14) \end{aligned}$$

Such a distribution over rankings is referred to as a Plackett-Luce model, whose origins and properties are described in detail in Marden (1996). In that book, and in the majority of works that cite it, the model is explained in terms of sequentially allocating the best ranks rather than the worst ones. We explain in sections 1.3 and 3.5 why this is rarely a sensible approach for sports modelling.

Our endure-Elo rating system can be understood as an algorithm for estimating the failure rate parameters from the ranking results of competitions. More specifically, we iteratively adjust estimates of the (unconstrained) negative logged failure rate parameters $R_i = -\log(\lambda_i)$ by moving them in the direction of the gradient of their log-likelihood function evaluated at their current values. The negation of the logged rate parameter means that the larger a competitor's R_i the greater the probability they will win a competition. As a result the R_i retain the interpretation of a 'strength parameter'. For each competition this log-likelihood is the logarithm of (14), which decomposes into $n-1$ additive terms corresponding to each of the independent elimination rounds.

For every round that leads to competitor i being eliminated we move her $\hat{R}_i$ parameter, which we consider an estimate for R_i, by an amount proportional to

$$\begin{aligned} \frac{\partial}{\partial R_i} \log P(i \text{ eliminated in round}) =& -\frac{\partial}{\partial R_i} \log \left(1 + \sum_{a \in Q \setminus i} e^{-R_a + R_i} \right) & (15)\\ =& -\frac{\sum_{a \in Q \setminus i} e^{-R_a + R_i}}{1 + \sum_{a \in Q \setminus i} e^{-R_a + R_i}} & (16)\\ =& -(1 - P(i \text{ eliminated in round})) = -P(i \text{ survives round}) & (17) \end{aligned}$$

and for every round she participates in and in which a rival competitor $j (\neq i)$ is eliminated we move her $\hat{R}_i$ parameter by an amount proportional to

$$\begin{aligned} \frac{\partial}{\partial R_i} \log P(j \neq i \text{ eliminated}) =& -\frac{\partial}{\partial R_i} \log \left(1 + \sum_{a \in Q \setminus j} e^{-R_a + R_j} \right) & (18)\\ =& \frac{e^{-R_i + R_j}}{1 + \sum_{a \in Q \setminus j} e^{-R_a + R_j}} & (19)\\ =& P(i \text{ eliminated in round}) = 1 - P(i \text{ survives round}). & (20) \end{aligned}$$

Combining these results we derive an update rule

$$\hat{R}_i \leftarrow \hat{R}_i + k\left[\mathbb{I}(\text{i survives round}) - P(\text{i survives round})\right] \tag{21}$$

or

$$\hat{R}_i \leftarrow \hat{R}_i - k\left[\mathbb{I}(\text{i eliminated in round}) - P(\text{i eliminated in round})\right], \tag{22}$$

where the constant of proportionality k, which is commonly referred to as a k-factor by users of the Elo system, is a step-size parameter in what is effectively a form of coordinate-wise gradient descent algorithm for estimating the R_i. Strictly speaking, the probability in (21) is an estimated probability derived from substituting the unknown R_i parameters for estimates of them with the effect that the approximate expression

$$P(\text{i survives round}) = 1 - \frac{e^{-R_{i,t}}}{\sum_{j=1}^{n} e^{-R_{j,t}}} \approx 1 - \frac{e^{-\hat{R}_{i,t}}}{\sum_{j=1}^{n} e^{-\hat{R}_{j,t}}} \tag{23}$$

is treated as an equality.

3.2 Competitor-specific k-factors

It may be argued that the value of k in (21) and (22) ought to be decreased as a competitor takes part in more matches since the information in a single match outcome relative to her historical record of games decreases. Doing so can be seen as a consequence of the log-likelihood function for a competitor's strength parameter given all available data concentrating on particular values and, consequently, the precision of our parameter estimates increasing. This connection between the log-likelihood function and the precision is made precise by adopting a locally quadratic approximation to the log-likelihood for a parameter R_i and a corresponding Normal approximation for its distribution. Specifically, we identify the negative second derivative of the log-likelihood with the parameter's precision (or inverse variance).

After a competitor has taken part in a round of a competition the precision for her strength parameter increases by the negative second derivative of the log-likelihood given the result of that round. Denoting the precision for parameter R_i as k_i^{-1}, the resulting update rule is

$$k_i^{-1} \leftarrow k_i^{-1} + P(\text{i survives round})(1 - P(\text{i survives round})) \tag{24}$$

where $P(\text{i survives round})(1 - P(\text{i survives round}))$ is the negative second derivative with respect to R_i of both

$$\log P(i \text{ eliminated in round}) \qquad \text{and} \qquad \log P(j \neq i \text{ eliminated in round}), \tag{25}$$

i.e. the second derivative is independent of whether the competitor is or is not eliminated.

The approximate log-likelihood can now be optimized by adjusting the relevant parameter in the direction of its gradient by an amount proportional to the inverse of the precision. This Newton optimization step is written as

$$\hat{R}_i \leftarrow \hat{R}_i + k_i\left[\mathbb{I}(\text{i survives round}) - P(\text{i survives round})\right]. \tag{26}$$

It is here that we notice the correspondence between the inverse precisions (variances) with the Elo k-factors. We note that variable k-factors are also used in the popular Glicko scoring system of Glickman (1999) for two-player competitions, which we understand as implementing an alternative, more sophisticated Newton-type optimization step for the log-likelihood for a competitor-specific log-rate parameter. Comparable alternative models and inference procedures for dynamic rating systems can be found in Knorr-Held (2000) and Cattelan et al. (2013).

3.3 Time varying logged rate parameters

A further improvement to the points system can be made by acknowledging variation over time in the R_i parameters. A simple and convenient way to model this variation is to assume that the R_i evolve according autoregressive processes of order one so that

$$R_{i,t} = \phi R_{i,t-1} + \epsilon_{i,t}. \tag{27}$$

The innovation terms $\epsilon_{i,t}$ are understood to be independent Normal random variables with expectation zero and variance σ^2, and the autoregression coefficient $\phi \in [0, 1]$ is understood as encoding competitors' consistency. We can apply the expectation and variance operators to each side of equation (27) and iterate the resulting expressions to derive the h-step ahead expectation and variance

$$\hat{R}_{i,t} \leftarrow \phi^h \hat{R}_{i,t-h}, \qquad \operatorname{var}(R_{i,t}) \leftarrow \phi^{2h} \operatorname{var}(R_{i,t-h}) + \sigma^2 \frac{1-\phi^{2h}}{1-\phi^2}. \tag{28}$$

The first equation is readily interpretable as an inflation-type effect whereby accumulated points (both positive and negative) decrease in absolute value over time. Interpreting the second equation is made easier with the introduction of the asymptotic variance $k_\infty = \sigma^2/(1-\phi^2)$ which allows us to rewrite (28) as

$$\hat{R}_{i,t} = \hat{R}_{i,t-h} + (1-\phi^h)(0 - R_{i,t-h}), \qquad k_{i,t} = k_{i,t-h} + (1-\phi^{2h})(k_\infty - k_{i,t-h}). \tag{29}$$

Equations (29) tell us then when h time increments have passed we should adjust the expectation and variance in the direction of their asymptotic values by amounts that depend on ϕ.

3.4 Specifying and interpreting hyperparameters

Since competitor abilities are only indirectly observed through the race results, learning appropriate values for the model hyperparameters is challenging. Suitable values may be searched for so that the conditional probabilities of the observed competition results are maximized. We anticipate that considerable care will be needed for such a strategy, however, since it is difficult to theorize about the function being maximized. Whether the hyperparameters are specified a priori or searched for numerically, it is sensible to build up some intuition for their role in the model. In the following subsections we consider each of the hyperparameters more closely.

3.4.1 Specifying the k-factors

The hyperparameter k_∞ is the marginal variance for the $R_{i,t}$, and tells us about the range of abilities among a population of competitors. One route to specifying k_∞ is to consider the disparity between competitors at the first and third quartiles of the whole population of competitors. Assuming that this population is well described by a normal distribution, and calling the probability that the stronger competitor beats the weaker competitor q, it follows that the standard deviation for the distribution of competitors is

$$k_\infty^{1/2} = \frac{1}{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)} \log\left(\frac{q}{1-q}\right) \approx \frac{3}{4} \log\left(\frac{q}{1-q}\right), \tag{30}$$

where Φ^{-1} is the quantile function of the unit normal distribution.

In general, we can treat all the Elo k-factors as variances for the strength parameters R_i, conditional on historical competition results. As a special case, the parameter k_∞ is the variance for a competitor whose last historical result was observed infinitely long ago. This brings us to the tricky question of specifying a fixed k-factor appropriate for Elo-type algorithms. We suggest that it is sensible to specify such a value as a fraction of k_∞ representing the relative uncertainty for a competitor's ability score within the relevant

population of peers. For example, we might consider a sport in which the competitor ranked 25 of 100 has a probability of 0.95 of beating a competitor ranked 75 of 100 so that $k_\infty^{1/2} \approx 2.18$. We might then suppose that even given recent competition results it is reasonable only to pin down a competitor's true position within the population of peers to the nearest decile. At the centre of the normal distribution its deciles are approximately 0.25 apart meaning that a k-factor of $k \approx (0.25 \times 2.18)^2 \approx 0.25$ may be appropriate. Experimentation with k-factors around this initial guess is then advisable.

3.4.2 Specifying autoregression coefficient ϕ

The hyperparameter ϕ quantifies the correlation between the R_i at consecutive time points, and tells us about the consistency or longevity of competitors' abilities. Specifically, we can think about the hyperparameter ϕ in terms of the rate at which our expectation for a competitor's ability decays over time. Specifically, if our expectation for a competitors ability relative to population average, as encoded by their R_i parameter, halves over a period of $t_{1/2}$ time units then we should specify $\phi = 2^{1/t_{1/2}}$.

3.5 Some properties of the exponential model for failure times

Let us consider a variant of the endure-Elo system whereby adjustments for all a competition's rounds are made simultaneously. This could be thought of as a step in the direction of the gradient of the sum of log-likelihoods for all the rounds' results. If all competitors are a priori considered to be equally capable then the competitor who is eliminated in the u^{th} round of n (so comes $v^{th} = (n-u+1)^{th}$ place in the competition) has her endure-Elo score updated by an amount proportional to

$$\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{n-(u-2)} - \left(1 - \frac{1}{n-(u-1)}\right) = -1 + \sum_{k=1}^{u} \frac{1}{n+1-k} = -1 + \sum_{k=n}^{v} \frac{1}{k}, \qquad (31)$$

where the first $u-1$ summands on the left hand side of (31) are earned by surviving rounds and the u^{th} summand is a downwards adjustment due to being eliminated in the u^{th} round.

The number of extra points a competitor receives as her finishing position improves gets larger. More precisely, if a competitor's position were to be adjusted from $v+1$ to v her reward in endure-Elo points would increase by an increment of $1/v$, which increases as v decreases. In this sense we would say that competitors receive increasing returns for getting a better position. A generalized Elo system constructed with exponential distributions describing race finishing times rather than failure times, so that consecutive events allocate competitors the best positions rather than the worst, leads to the opposite result. This means that a model for exponential race finishing times leads to a rating system in which competitors receive diminishing returns as they get better positions. Scaled and integer-rounded endure-Elo score updates as specified by (31) for a race with $n = 20$ competitors are presented in Table 1. We note their qualitative similarity to those of the official F1 Championship point system insofar as the second differences of both sequences of scores being non-negative, while similarly computed speed-Elo score updates exhibit non-positive second differences.

Another way to appreciate how the model for exponential failure times associates the greatest significance with the best performances follows from considering two competitors labelled 1 and 2 with failure rates $\lambda_1 > \lambda_2$. Suppose we observe that one of them survives for at least c time units and we are asked to guess which competitor it is. The likelihood ratio for the two options is

$$\frac{P(X \geq c \mid X \sim \text{Exp}(\lambda_1))}{P(X \geq c \mid X \sim \text{Exp}(\lambda_2))} = e^{-(\lambda_1 - \lambda_2)c}, \qquad (32)$$

where X denotes the unknown competitor's time of failure. Clearly, the greater c is the more convinced we we ought to be that the data refers to competitor 2 with the lower failure rate. Phrased slightly differently,

the more impressive the performance we observe the more sure we should be that it was achieved by the stronger competitor.

We suggest that awarding increasingly large parameter adjustments for the best competition results is the more appropriate strategy in the majority of sporting competitions. The non-mathematical intuition to support the suggestion is that, in most cases, it is more likely that from chance alone for a good competitor to perform exceptionally badly than for a bad competitor to perform exceptionally well. In practice this idea means that we ought to be relatively forgiving of poor performances and differentiate between degrees of poor performance less severely. This is reflected in the allocation of F1 Championship points (see Table 1), for example, where the point difference between first and second place is much greater than the difference between 19^{th} and 20^{th} place.

Another way to make these ideas explicit is by considering the probability that the best competitor wins an endurance or speed competition. Obviously, the probability of the complementary event being small formalizes that statement 'it is unlikely for a bad competitor to perform exceptionally well'. Beginning with the endurance competition, we suppose that there are three competitors with failure times that are exponentially distributed with failure rates $0 < \lambda_1 < \lambda_2 < \lambda_3$, meaning that on average competitor one is the best. The probability that she beats one of her less able rivals (i.e. they fail before her) is

$$P(\text{1 beats i in endurance comp.}) = \frac{\lambda_i}{\lambda_i + \lambda_1} \tag{33}$$

and, as we show in Appendix A.1.1, the probability that she wins the whole competition (i.e. her exponential failure time is the greatest) is

$$p_{\text{endure}} = 1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} - \frac{\lambda_1}{\lambda_1 + \lambda_3} + \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}. \tag{34}$$

Now we can specify a speed competition in which all the pairwise win-probabilities are the same as that for the endurance competition by supposing that the finishing rates for the competitors are $\rho_1 = \lambda_1^{-1} > \rho_2 = \lambda_2^{-1} > \rho_3 = \lambda_3^{-1} > 0$. It is easy to see that

$$P(\text{1 beats i in speed comp.}) = \frac{\rho_1}{\rho_i + \rho_1} = \frac{\lambda_1^{-1}}{\lambda_i^{-1} + \lambda_1^{-1}} = \frac{\lambda_i}{\lambda_i + \lambda_1} \tag{35}$$

and that the probability that competitor one wins the speed competition is

$$p_{\text{speed}} = \frac{\lambda_1^{-1}}{\lambda_1^{-1} + \lambda_2^{-1} + \lambda_3^{-1}}. \tag{36}$$

After some rearranging, we see that

$$p_{\text{endure}} - p_{\text{speed}} = \frac{(\lambda_2\lambda_3 - \lambda_1^2)\lambda_1\lambda_2\lambda_3}{(\lambda_1 + \lambda_2)(\lambda_1 + \lambda_3)(\lambda_1 + \lambda_2 + \lambda_3)(\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3)} > 0 \tag{37}$$

because $\lambda_2\lambda_3 > \lambda_1^2$ since, by assumption, competitor one's failure rate is less than those of her rivals. All the other factors in (37) are positive simply because they are sums and products of positive rate parameters. It is in this sense that in an endurance competition the strongest competitor is more likely to win, and a bad competitor less likely to win, than in a speed competition.

Our final demonstrative calculation relates more directly to the claim that in endurance competitions it is more probable for a good competitor to do badly than a bad competitor to do well. We consider again the endurance competition with three competitors. This time we put stronger restrictions on the failure rates

$$\lambda_1 = 1, \qquad \lambda_2 = a, \qquad \lambda_3 = a^2, \tag{38}$$

so that

$$P(\text{i beats i+1 in endurance comp.}) = \frac{a}{1 + a}, \tag{39}$$

implying that, in a sense, the competitors are equally spaced along a continuum of ability. When $a > 1$ competitor 1 is, in expectation, better than competitor 2, who is better than competitor 3 to the same extent. We can now quantify the relative probabilities of the best competitor coming last and the worst competitor coming first (in two independent competitions),

$$\frac{P(1 \text{ comes last in endurance comp.})}{P(3 \text{ comes first in endurance comp.})} = \left(\frac{1}{1+a+a^2}\right) \Big/ \left(1 - \frac{a^2}{a^2+1} - \frac{a^2}{a^2+a} + \frac{a^2}{a^2+a+1}\right) \tag{40}$$

$$= \frac{a^3+a^2+a+1}{2a^2+a+1}, \tag{41}$$

which is clearly greater than one so long as a is. Specification of the speed competition with the same pairwise win-probabilities requires that finishing times are exponentially distributed with finishing rates

$$\rho_1 = 1, \qquad \rho_2 = a^{-1}, \qquad \rho_3 = a^{-2}. \tag{42}$$

It follows that

$$\frac{P(1 \text{ comes last in speed comp.})}{P(3 \text{ comes first in speed comp.})} = \left(1 - \frac{1}{1+a^{-1}} - \frac{1}{1+a^{-2}} + \frac{1}{1+a^{-1}+a^{-2}}\right) \Big/ \left(\frac{a^{-2}}{1+a^{-1}+a^{-2}}\right) \tag{43}$$

$$= \left(1 - \frac{a^2}{a^2+1} - \frac{a^2}{a^2+a} + \frac{a^2}{a^2+a+1}\right) \Big/ \left(\frac{1}{1+a+a^2}\right) \tag{44}$$

$$= \frac{2a^2+a+1}{a^3+a^2+a+1}, \tag{45}$$

which, being the reciprocal of (41), is less than one when $a > 1$. In this sense the model with exponential finishing times explicitly violates the assumption that it is more probable for a good competitor to do badly than a bad competitor to do well. We will see how this leads to particularly poor predictions in our example with real F1 data in Section 4.2.

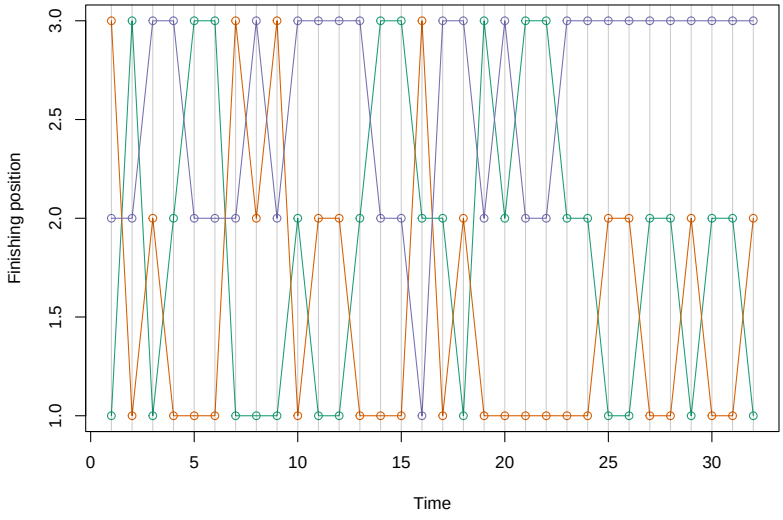
4 Examples

4.1 Synthetic data

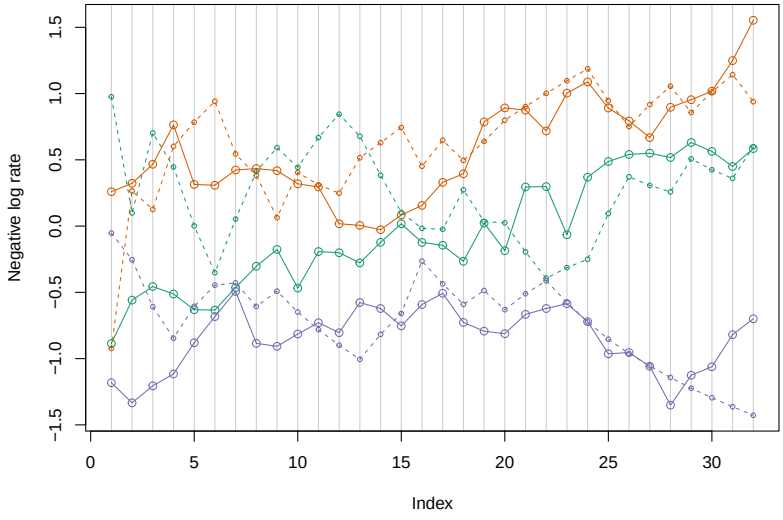
Our synthetic example, which is intentionally highly simplified, involves $m = 3$ competitors who all take part in $n = 32$ multiplayer competitions that occur at consecutive unit time steps. We simulate time-varying competitor-specific negative log failure rate parameters from an AR(1) process with $\phi = 0.99$ and identically and independently distributed normal innovations with standard deviation $\sigma = 0.2$. For each competition we then simulate exponential failure times given the failure rates at the corresponding event times. These failure times are then ranked (rank 1 going to the latest time, rank 2 to the second latest time and rank 3 to the earliest time), resulting in the sequence of positions illustrated in Figure 1a. The 'true' simulated values of the negative log failure rate parameters and our estimates for them based only on the ranks, which are our endure-Elo scores, are plotted in Figure 1b using solid and dashed lines respectively. We observe that, as expected, the estimates closely track the true values. We also see, for example, how winning (or losing) streaks lead to smooth drifts upwards (or downwards) in the endure-Elo scores, and how atypical race results lead to larger, more abrupt changes in the endure-Elo scores.

4.2 Real data

We now consider real data from F1 motor-racing, which at the time or writing is publicly available at https://ergast.com/mrd. In Section 4.2.1 we look at a large data set of race results over several years and



(a) Simulated finishing positions. Position one is given to the competitor with the greatest survival time in a competition.



(b) Negative log failure rates for three competitors. Solid lines interpolate the true simulated values (the ability parameters $R_{i,t}$) and dashed lines interpolate estimates (the endure-Elo scores $\hat{R}_{i,t}$).

Fig. 1: Simulated data and inferences generated from the model for exponential survival times and the endure-Elo system for parameter estimation.

examine the speed-Elo and the endure-Elo systems' long term potential for predicting race outcomes. This first analysis provides empirical support for the endure-Elo system's adoption for race events. Then, in Section 4.2.2 we focus on a particular F1 season and a subset of the participating drivers. At this scale we can more easily identify individual race events with Elo adjustments, revised predictions and prediction errors. This second analysis serves to explain the endure-Elo system's good performance on the larger data set, and hence to reinforce the findings of Section 4.2.1.

Reiterating the results of Section 3, the endure-Elo rating system involves updating the score for competitor $j = 1, \ldots, m$ (who is knocked out in round u_j and so finishes with position $v_j = n + 1 - u_j$) according to

$$\hat{R}_j \leftarrow \hat{R}_j + k \sum_{a=1}^{u_j} [X(\text{j survives round a}) - \mathbb{E}(X(\text{j survives round a}))] \tag{46}$$

where

$$\mathbb{E}(X(\text{j survives round a})) = 1 - \mathbb{E}(X(\text{j eliminated in round a})) = 1 - \frac{e^{-R_j}}{\sum\limits_{i \,:\, u_i \geq a} e^{-R_a}}. \tag{47}$$

Here elimination rounds sequentially pick out competitors to receive the next worst as yet untaken finishing position and eliminated competitors do not take part in subsequent rounds. The probability of competitor j finishing in first place is the probability that she survives all $m - 1$ rounds. As shown in Appendix A, it is computed as

$$P(j \text{ wins whole competition}) = \sum_{r=0}^{2^{n-1}} (-1)^{|C_r|} \frac{e^{-R_j}}{e^{-R_j} + \sum_{k \in C_r} e^{-R_k}} \tag{48}$$

where the C_r are the elements of the power set of the set of competitor labels excluding j.

The speed-Elo system involves updating the score of competitor j (who wins selection round u_j and so comes in with position $v_j = u_j$) according to

$$\hat{R}_j \leftarrow \hat{R}_j + k \sum_{a=1}^{u_j} [X(\text{j wins round a}) - \mathbb{E}(X(\text{j wins round a}))] \tag{49}$$

where

$$\mathbb{E}(X(\text{j wins round a})) = \frac{e^{R_j}}{\sum\limits_{i \,:\, u_i \geq a} e^{R_a}}. \tag{50}$$

Here selection rounds sequentially pick out competitors to receive the next best as yet untaken finishing position and selected competitors do not take part in subsequent rounds. The probability of competitor j finishing in first place in the whole competition is the probability that she wins the first round, i.e.

$$P(\text{j wins whole competition}) = X(\text{j wins round 1}) = \frac{e^{R_j}}{\sum\limits_{a=1}^{m} e^{R_a}}. \tag{51}$$

For both systems, following the heuristic argument discussed in Section 3.4.1, we use a fixed k-factor of $k = 0.36$.

We measure the performance of our models according to their log-likelihoods. These are the log-probabilities they assign jointly to all the outcomes that occurred. In practice these are computed by adding together log-probabilities for individual results given previous results. To make these log-probabilities more meaningful, however, we couch them in the language of betting. We imagine that an agent using the speed-Elo model computes probabilities $p_{i,j}$ for race $i = 1, \ldots, n$ being won by competitor $j = 1, \ldots, m$. She then offers bets whereby she pays out $1/p_{i,j}$ for every one unit wagered on competitor j if competitor j

does win, and nothing otherwise. These bets are fair in the sense that her expected profit from each bet is zero. We now imagine that the user of the endure-Elo model computes corresponding probabilities $q_{i,j}$ and for each race i spends all her available wealth S_i on bets, splitting it between competitors labelled by j in proportion to her $q_{i,j}$. This betting strategy maximizes her expected log-wealth. We encode race results by the indicator variables $X_{i,j}$, which take value one when race i is won by driver j and zero otherwise. Given all of this notation, the user of the endure-Elo model increases her log wealth by

$$D(q,p) = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{i,j}\Big[\log\left(\underbrace{S_i}_{\substack{\text{Wealth before}\\ \text{race } i}} + \underbrace{S_i q_{i,j}\left(\frac{1}{p_{i,j}}-1\right)}_{\substack{\text{Profit from}\\ \text{winning bet}}} - \underbrace{S_i(1-q_{i,j})}_{\substack{\text{Losses from}\\ \text{losing bets}}}\right) - \log(S_i)\Big] \tag{52}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} X_{i,j}\log\left(\frac{q_{i,j}}{p_{i,j}}\right) \tag{53}$$

over the series of n races. Technically, this is a log-likelihood ratio for the two models given a subset of the available data because the results gambled on only concern first place positions, not all positions. We have chosen to compare the models in this way because it prioritizes the types of result that are typically of most interest to stakeholders in F1 racing.

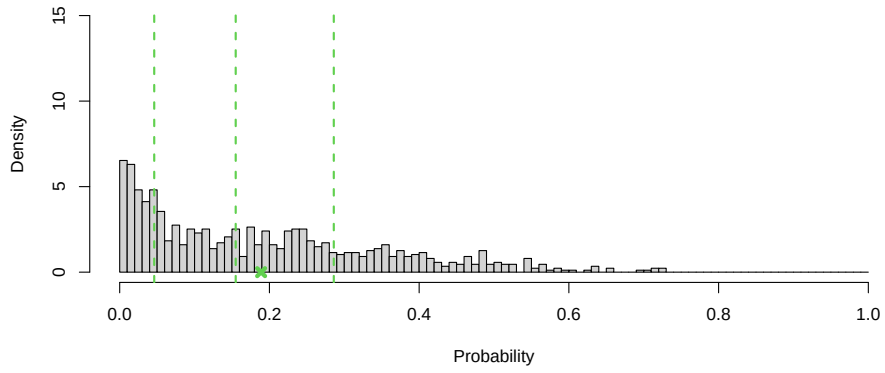
4.2.1 Analysis of historical data

For the 52 F1 seasons between 1970 and 2021 we compute winner probabilities for each race using each of our Elo systems. Seasons consist of between 11 and 22 races, leading to a total of $n = 873$ races. Every season is modelled independently in the sense that the $\hat{R}_i$ strength parameters for the competitors in a given season are set to zero at the start of each one and the Elo systems make adjustments as the season progresses. In Figure 2 we plot histograms of the win probabilities assigned to the eventual winners of each race. These values tell us about the predictive abilities of the systems in an absolute sense. We see clearly that the empirical distribution of probabilities for the endure-Elo has more mass at higher values. This observation is made more precise when we note that all three quartiles are greater than those for the speed-Elo model.

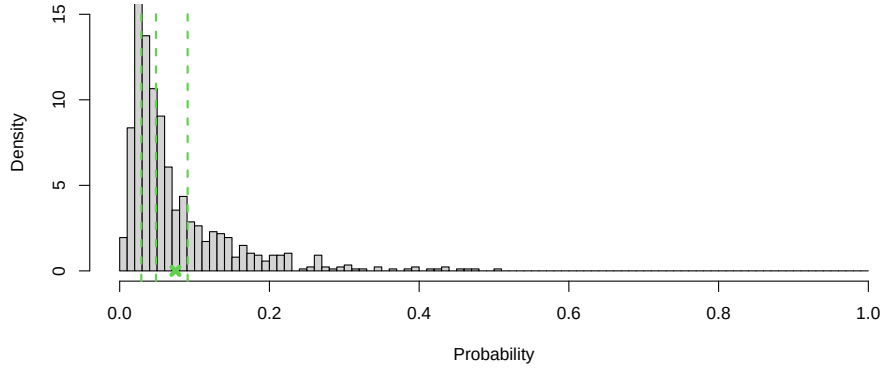
In Figure 3 we plot the summands appearing in (53). These are the logged wealth multipliers for the endure-Elo user following each individual wager against the speed-Elo user. We note that the median of the (unlogged) multipliers is 2.180 and that 76.3% of the multipliers are greater than one - meaning that in 76.3% of races the endure-Elo user could be said to beat the speed-Elo user and that in 50% of these wagers the endure-Elo user more than doubles her wealth. The total log-likelihood ratio (53) is 592. The mean and variance of the summands are 0.678 and 1.331 respectively. A bootstrap resampling exercise reveals that the sum of 18 randomly selected summands is greater than zero with probability 99%, implying that endure-Elo system can be expected to lead to positive returns with high probability after a single season of approximately 20 races.

4.2.2 Analysis of a single F1 season

We focus now on Grand Prix races during the 2019 season. We begin by looking at the official F1 points used for determining the winner of the annual F1 Championships. These are awarded according to a racer's finishing position as described fully in Table 4b. Drivers receive an additional bonus point if they finish in the top ten and also complete the fastest lap. We note that two drivers, Albon and Gasly, switched team (and therefore car) during the 2019 season. Although there is a strong argument that distinct driver/team combinations ought to be considered as distinct competitors with their own ability parameters, for the time



(a) Endure-Elo probabilities (quartiles at 0.046, 0.155 and 0.286).



(b) Speed-Elo probabilities (quartiles at 0.029, 0.048 and 0.091).

Fig. 2: The histograms illustrate the empirical distributions of win-probabilities assigned competitors who did win particular races. Large probabilities signify instances when a model predicted the winner with high certainty, while small probabilities signify instances when the model was surprised by the observed identity of the winner. The dashed vertical lines mark the quartiles of the distributions. The green cross marks the mean.

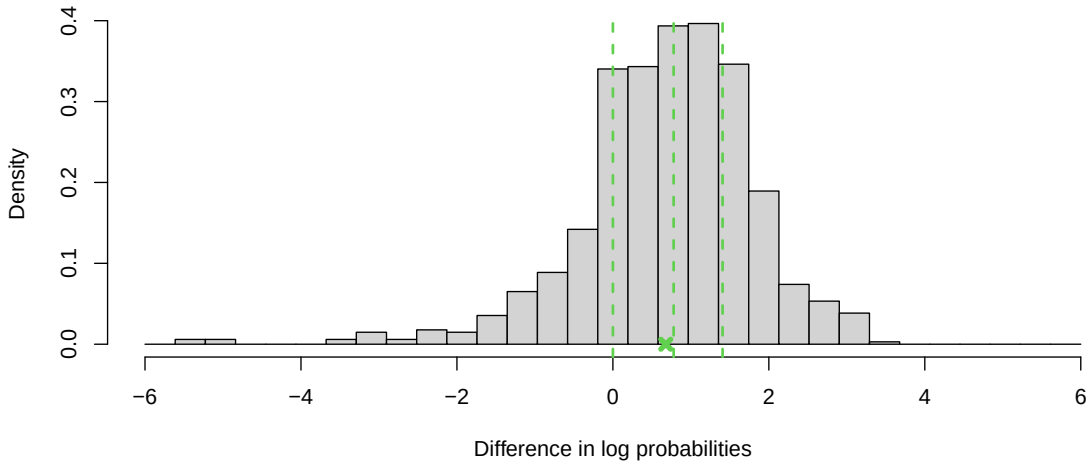
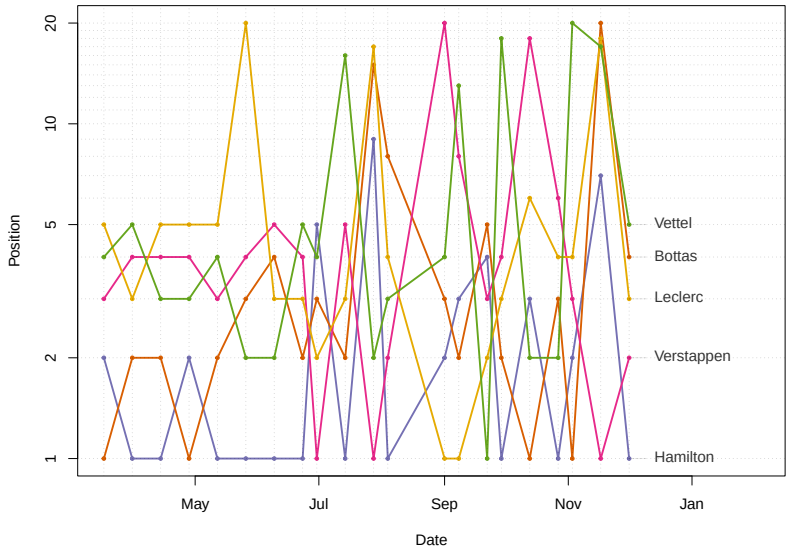


Fig. 3: Differences in log-probabilities assigned by the endure-Elo user and the speed-Elo user to events that did occur. Quartiles at 0.000, 0.779 at 1.407 are marked with dashed vertical lines. The mean difference of 0.678 is marked with a cross.

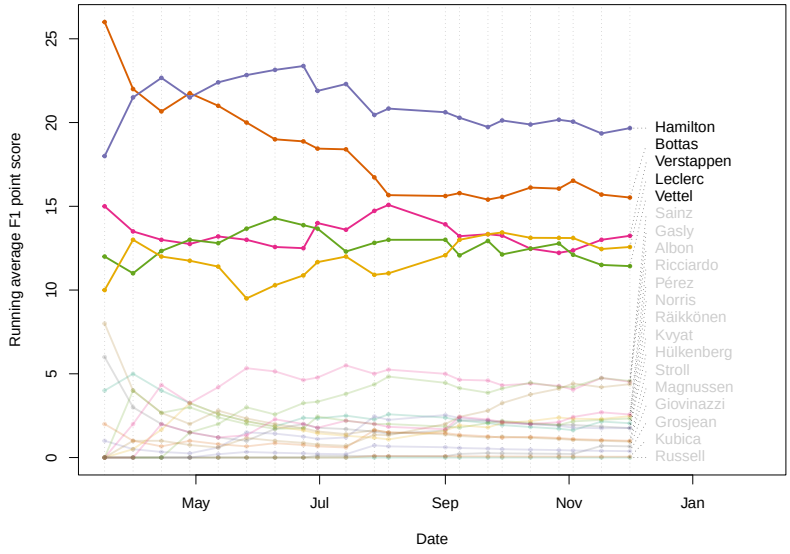
being we consider only the drivers and ignore the team switch. We do so to minimize complications that might distract from the core features of the rating systems under consideration.

The finishing positions and running means of F1 Championship points for the drivers are plotted in Figures 4a and 4b, respectively. The F1 points can be seen to move in approximate qualitative correspondence with the generalized Elo scores plotted in Figure 5. We note also the similarity between the final rankings according to the F1 points and the endure-Elo scores. These observations are reassuring although not necessarily quantitatively meaningful since the F1 point system is not primarily a device for making predictions for particular races, rather the system is intended to reward drivers for good performances over a season. The vertical grey lines in these plots mark the dates of races, immediately before which the points are recalculated given past results. Black triangles are positioned on the time series corresponding to the winning drivers. Informally, a plot with triangles higher up the y-axis signifies a scoring system that better predicts the winners. The straight coloured lines interpolating the scores on race days are included only as an aid for the reader.

We can see in Figure 5b how, for instance, the speed-Elo system severely punishes Vettel, Bottas, Verstappen and Leclerc for a small number of very poor performances, details of which can be read from Table 2. The races in question all involve crashes or mechanical faults that send these drivers to the bottom of the field and lead to large sudden drops in their scores. The drops contribute to a large gap building up in subsequent win-probabilities between the affected drivers and Hamilton, who suffers no such poor performances and punishments. The endure-Elo system takes the poor performances much less seriously and so does not consider the gap between Hamilton and the others to be so great. In terms of the gambles, this pays off for the endure-Elo system on several occasions. In particular, the endure-Elo user profits greatly when Leclerc wins the Belgian and Italian Grand Prix, Vettel wins the Singapore Grand Prix (all in September) and Verstappen wins the Brazilian Gran Prix (in November) despite their previous mishaps.

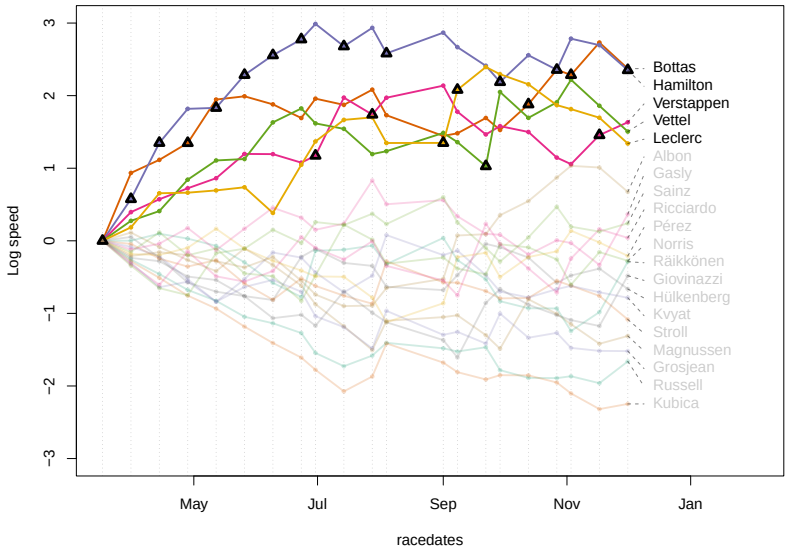


(a) Finishing positions.

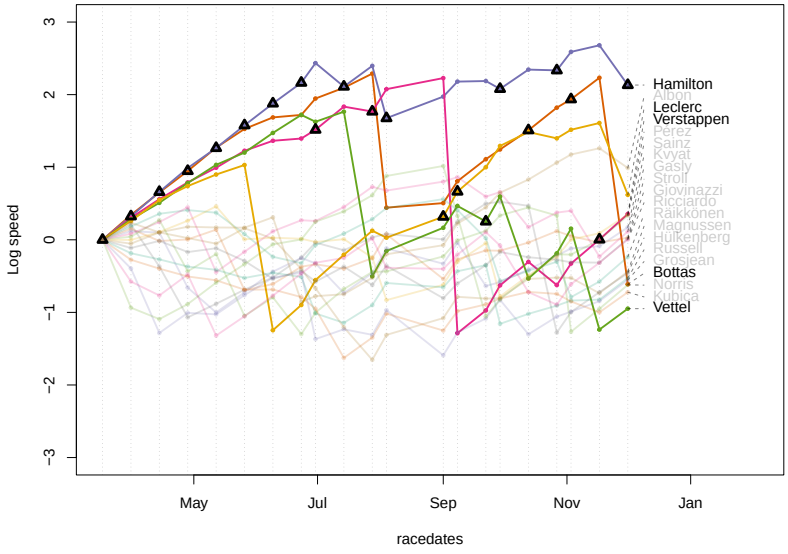


(b) Running averages of F1 Championship points.

Fig. 4: Race statistics for F1 drivers in 2019. Note that to avoid clutter subfigure 4a includes results only for the 5 competitors with the most F1 points at the end of the season.



(a) endure-Elo.



(b) speed-Elo.

Fig. 5: The plots above show ability scores for the generalized Elo systems (the $\hat{R}_i$ values) immediately before each race of the 2019 F1 season. Black triangles identify the drivers who won particular races. Opacity of the plotting lines draw attention to competitors who finish the season with the most F1 Championship points.

5 Remarks

We have proposed a novel endure-Elo rating system for participants in multiplayer competitions motivated by a coherent probabilistic model. The existence of this underlying model, and its concordance with real competitions, allows us to judge if and how the endure-Elo system can be expected to perform well. The same cannot be said for ad hoc modifications to the conventional Elo system for one-on-one competitions that are not motivated by a particular model, and can also not be said for systems based on models too complicated for mathematical analysis. We have also demonstrated the superior predictive performance of the endure-Elo system over the speed-Elo for a prototypical example. Specifically, we have presented an instance in which a user of the endure-Elo system could reliably profit from bets made in competition with a user of the speed-Elo system. Importantly, we have been able to deconstruct and rationalize this result in terms of our model's properties. More precisely, we identified its success with the way it treats poor performances by good competitors relative to good performances by poor competitors. In this way the model provides a theoretical argument explaining why and when our empirical finding ought to generalize to other contexts.

In sections 3.2 and 3.3 we suggested ways to tailor the endure-Elo system to account for the relative uncertainty for competitors' true abilities and for the variability of those abilities over time. These issues can be expected to be especially important when we consider series of competitions over longer periods in which competitors' abilities change substantially. We intend to investigate this further in future work. For now, however, we note that there is a significant premium for simple rating systems that can be put to immediate use by non-experts. We therefore prioritize the promotion of the simpler version of the system that uses fixed k-factors, and encourage users to experiment with it.

At a more general level, our work calls into question common practices for modelling ranks and preferences. We have seen that the Plackett-Luce model for sequentially allocating the lowest ranks rather than the highest ranks leads to very different results in terms of the model's implications for predicting the best and worst of a set of competitors. In Yellott Jr (1977) Yellot establishes a connection between the Plackett-Luce model and Luce's Choice Axiom (see Luce (1977)), which proposes that numerical quantifications of preference can simply be rescaled when some options are removed. Given this connection and our findings with the F1 data, it becomes clear that the relevance of Luce's Choice Axiom for preferences ought to be considered carefully before it is used to inform a model. Specifically it is important to consider whether it is more appropriate to apply the Axiom to preferences for which assets (in our case competitors) to keep or to discard. It is beyond the scope of the current work to investigate further the distinction between these modelling choices but we hope our findings can contribute to discussions of the topic in some way.

Supplementary material

This paper is accompanied by an Rmarkdown document with which readers can reproduce the analyses presented above. Readers are invited to inspect the code to get an idea for how the endure-Elo methodology can be implemented. They are also encouraged to modify and improve the code, which is written principally to illustrate relevant calculations rather than provide an optimized implementation of them.

Data availability statement

The data underlying this article are available in the Formula One data repository at https://ergast.com/mrd.

Funding statement

No external funding was received in support of the work presented in this paper.

Conflict of interest statement

The author declares no conflicts of interest.

Acknowledgments

The author would like to thank participants if the 2022 MathSport International Conference for drawing attention to the problem addressed in the paper. Particular thanks are due to Paul Steele and David Scott for their encouragement, and for helpful conversations on the topic of F1 racing and Elo scores.

References

Beggs, S., Cardell, S., and Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of econometrics*, 17(1):1–19.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Cattelan, M., Varin, C., and Firth, D. (2013). Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):135–150.

Dabadghao, S. and Vaziri, B. (2022). The predictive power of popular sports ranking methods in the NFL, NBA and NHL. *Operational Research*, 22(3):2767–2783.

Ebtekar, A. and Liu, P. (2021). Elo-mmr: A rating system for massive multiplayer competitions. In *Proceedings of the Web Conference 2021*, pages 1772–1784.

Elo, A. E. (1967). The proposed uscf rating system, its development, theory, and applications. *Chess Life*, 22(8):242–247.

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

Gormley, I. C. and Murphy, T. B. (2008). Exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027.

Graves, T., Reese, C. S., and Fitzgerald, M. (2003). Hierarchical models for permutations: Analysis of auto racing results. *Journal of the American Statistical Association*, 98(462):282–291.

Henderson, D. A. and Kirrane, L. J. (2018). A Comparison of Truncated and Time-Weighted Plackett–Luce Models for Probabilistic Forecasting of Formula One Results. *Bayesian Analysis*, 13(2):335 – 358.

Herbrich, R., Minka, T., and Graepel, T. (2006). Trueskill™: a Bayesian skill rating system. *Advances in neural information processing systems*, 19.

Hvattum, L. M. and Arntzen, H. (2010). Using Elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470.

Ingram, M. (2021). How to extend Elo: a Bayesian perspective. *Journal of Quantitative Analysis in Sports*, 17(3):203–219.

Knorr-Held, L. (2000). Dynamic rating of sports teams. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(2):261–276.

Lasek, J., Szlávik, Z., and Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1(1):27–46.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3):215–233.

Mallows, C. L. (1957). Non-null ranking models. i. *Biometrika*, 44(1/2):114–130.

Marden, J. I. (1996). *Analyzing and modeling rank data*. CRC Press.

McHale, I. and Morton, A. (2011). A bradley-terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619–630.

Moore, J., Dottle, R., and Paine, N. (2018). Who's the best formula one driver of all time? *fivethirtyeight.com*. https://fivethirtyeight.com/features/formula-one-racing/.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202.

Silver, N. (2014). Introducing nfl elo ratings. *fivethirtyeight.com*. https://fivethirtyeight.com/features/introducing-nfl-elo-ratings/.

Silver, N. and Fischer-Baum, R. (2015). How we calculate nba elo ratings. *fivethirtyeight.com*. https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/.

Weng, R. C. and Lin, C.-J. (2011). A Bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12(1).

Yellott Jr, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144.

A Mathematical analysis of the model for exponential failure times

A.1 Quantities derivable from the endure-Elo scores

The probability that competitor i fails k^{th} (and is ranked $(m+1-k)^{th}$) or among the first k of a total of n competitors (so is ranked among the top $m+1-k$) is often of interest to sports fans and bookmakers. In theory equation (14) can be computed for all $m!$ possible permutations of the competitors, and the probability in question computed as the sum of the relevant subset of terms. Given that we have access to an algorithm to enumerate permutations, this computational strategy is, in principle, straightforward to implement. Its computational cost, however, is liable to become prohibitive as the number of competitors increases. Below we suggest more efficient ways for computing these probabilities.

A.1.1 Probability competitor i fails kth

We consider a set C of $k-1$ competitors who all fail before i and the remaining set D of $m-k$ competitors who fail after i. We then integrate out over all the possible failure times for i,

$$P(\text{only } C \text{ fail before } i) = \int_0^\infty \lambda_i e^{-\lambda_i x} \times \left(\prod_{a \in C} (1 - e^{-\lambda_a x}) \right) \times e^{-\lambda_D x} dx \qquad \lambda_D = \sum_{b \in D} \lambda_b. \tag{54}$$

We then rewrite the product over the competitors in C as a sum over subsets of C (denoted C_r) in order to derive the expression

$$\begin{aligned} P(\text{only } C \text{ fail before } i) &= \int_0^\infty \lambda_i e^{-\lambda_i x} \times \left(\sum_{r=0}^{2^{k-1}} (-1)^{|C_r|} e^{-\sum_{a \in C_r} \lambda_a x} \right) \times e^{-\lambda_D x} dx & (55) \\ &= \sum_{r=0}^{2^{k-1}} (-1)^{|C_r|} \int_0^\infty \lambda_i e^{-\lambda_i x} \times e^{-\sum_{a \in C_r} \lambda_a x} \times e^{-\lambda_D x} dx & (56) \\ &= \sum_{r=0}^{2^{k-1}} (-1)^{|C_r|} \frac{\lambda_i}{\lambda_i + \lambda_D + \sum_{a \in C_r} \lambda_a}. & (57) \end{aligned}$$

Finally, we sum over $\binom{m-1}{k-1}$ possibilities for C to reach the expression

$$P(i \text{ fails } k^{th}) = \sum_C \sum_{r=0}^{2^{k-1}} (-1)^{|C_r|} \frac{\lambda_i}{\lambda_i + \lambda_D + \sum_{a \in C_r} \lambda_a}. \tag{58}$$

To reiterate, computing probability (58) naively via full enumeration of the probabilities for all $m!$ race results would involve $\mathcal{O}(m^2 m!)$ operations. Because the smallest of the $m-k$ times outside the top k has

an exponential distribution we do not actually have to consider all their permutations. Accounting for this, we can compute the probabilities only considering permutations of competitors who fail before i, which leads to a procedure involving $\mathcal{O}(\binom{m-1}{k-1}k^2(k-1)!)$ operations. The calculation explained above (specifically, expanding the product in (54)) allows us to reduce the operation count a bit further to $\mathcal{O}(\binom{m-1}{k-1}k2^{k-1})$. For moderately large k this is still, admittedly, a demanding calculation. Further work may lead to improvements in the tractability of these probabilities, but we do not currently see how they would be achieved.

A.1.2 Probability competitor i fails first

If competitor i fails first then the set C appearing in (58) is empty. In this case the probability in question reduces to

$$P(i \text{ fails } k^{th}) = \frac{\lambda_i}{\lambda_i + \lambda_D}, \tag{59}$$

where λ_D is the sum of the failure rates for all the competitors excluding i.

A.1.3 Probability competitor i fails last

For endurance-type competitions the competitor who finishes (or fails) last is the winner, meaning that the probabilities of finishing last are particularly important. We can compute such probabilities using (58) with k set to m. In this case the outermost summation in (58), which counts the number of combinations of rival competitors that fail before i does, and λ_D, which sums the rate parameters for competitors failing after i, can be disregarded. The equation becomes

$$P(i \text{ fails } m^{th}) = \sum_{r=0}^{2^{n-1}} (-1)^{|C_r|} \frac{\lambda_i}{\lambda_i + \sum_{a \in C_r} \lambda_a}, \tag{60}$$

where the C_r are the members of the power set of the set of competitor labels excluding i. Equation (60) can be seen as an application of the inclusion-exclusion principle since

$$P(i \text{ fails } m^{th}) = 1 - P(i \text{ fails before someone else}) \tag{61}$$

$$= 1 - P(\bigcup_{j \neq i} i \text{ fails before } j), \tag{62}$$

where $\bigcup_j$ denotes the union of events indexed by j. Given this understanding, Bonferroni's inequality tells us that (60) can be bounded from above and below by partial sums. Specifically, we can split the sum in (60) into parts that each sum over sets of rival competitors of a given size

$$P(i \text{ fails } m^{th}) = \sum_{a=0}^{n-1} \sum_{|C|=a} (-1)^a \frac{\lambda_i}{\lambda_i + \sum_{a \in C} \lambda_a} \tag{63}$$

and use Bonferroni's inequality to deduce that

$$P(i \text{ fails } m^{th}) \leq \sum_{a=0}^{v} \sum_{|C|=a} (-1)^a \frac{\lambda_i}{\lambda_i + \sum_{a \in C} \lambda_a} \qquad \text{for even } v \tag{64}$$

$$P(i \text{ fails } m^{th}) \geq \sum_{a=0}^{v} \sum_{|C|=a} (-1)^a \frac{\lambda_i}{\lambda_i + \sum_{a \in C} \lambda_a} \qquad \text{for odd } v. \tag{65}$$

The practical utility of this result is that the potentially very long sum in (60) can be avoided if we are content just to bound it. This would mean computing partial sums like (64) for some v and using the last two as upper and lower bounds.

If, for example, three competitors' failure times are exponentially distributed with rates λ_a with $a \in \{i, j, k\}$, then the probability of competitor i finishing last is the sum

$$P(\text{i finishes last}) = 1 - \frac{\lambda_i}{\lambda_i + \lambda_j} - \frac{\lambda_i}{\lambda_i + \lambda_k} + \frac{\lambda_i}{\lambda_i + \lambda_j + \lambda_k}, \tag{66}$$

in which the summands' denomimators include the rate parameter for competitor i and every combination of the remaining competitors. In this instance the Bonferroni inequalities tell us that

$$P(\text{i finishes last}) \leq 1, \tag{67}$$

$$P(\text{i finishes last}) \geq 1 - \frac{\lambda_i}{\lambda_i + \lambda_j} - \frac{\lambda_i}{\lambda_i + \lambda_k}, \tag{68}$$

$$P(\text{i finishes last}) \leq 1 - \frac{\lambda_i}{\lambda_i + \lambda_j} - \frac{\lambda_i}{\lambda_i + \lambda_k} + \frac{\lambda_i}{\lambda_i + \lambda_j + \lambda_k}. \tag{69}$$

An alternative upper bound can be derived from the fact that a competitor's chance of failing last can only decrease as more rival competitors join the competition, i.e. if we partition the set of all competitor labels Q into sets U_i (whose subscript encodes the fact that it contains label i) and V then

$$P(i \text{ finishes last among } Q = U \cup V) \leq P(i \text{ finishes last among } U) \tag{70}$$

which can then be minimized by specifying U_i so that it contains i and a subset of the (smallest) failure rates corresponding to the best competitors. Accompanying lower bounds can be found by considering the complement to the event in question

$$P(i \text{ finishes last among } Q) = 1 - P(i \text{ does not finish last among } Q) \tag{71}$$

$$= 1 - \sum_{j \in Q \setminus i} P(j \text{ finishes last among } Q) \tag{72}$$

$$\geq 1 - \sum_{j \in Q \setminus i} P(j \text{ finishes last among } U_j), \tag{73}$$

where, similarly, U_j is a subset of labels for the best competitors and label j.

In practice we find the probability that a competitor finishes last is most efficiently computed by recursively building up the set of denominators and corresponding signs appearing in (60), before manipulating them all simultaneously and adding them up. The R code below shows exactly how this is achieved. It defines a function whose input is a vector of failure rates for a set of exponential random variable and whose output is a vector of probabilities for the corresponding random variables being the largest.

```
lastprob<-function(lambdavect){
  n<-length(lambdavect)
  probvect<-rep(0,n)
  for(i in 1:n){

  lambdavecti<-lambdavect[-i]
  denominators<-lambdavect[i]
  signs<-1
    for(k in 1:(n-1)){
    denominators<-c(denominators,denominators+lambdavecti[k])
    signs<-c(signs,-signs)
    }
  probvect[i]<-sum(lambdavect[i]/denominators*signs)

  }
  probvect
}
```

With 20 competitors, as is mostly the case in our F1 examples, the function above evaluates the required probabilities in approximately 4.2×10^{-1}s on a laptop equipped with a 1.9GHz processor. The corresponding calculation of the probabilities for failing first is orders of magnitude faster, averaging approximately 2.2×10^{-6}s. So at this scale we find neither calculation to be prohibitively costly. The super-exponentially increasing cost of the former calculation quickly becomes a problem as we introduce more competitors however. With 24 competitors, the function above takes approximately 14×10^{0}s to evaluate, which in some contexts may already be impractically slow.

A.1.4 Probability competitor i finishes among the first k

Our strategy for computing this probability follows along similar lines as those described in Section A.1.1. We start by writing down the probability competitor i and a set, denoted C, of $k-1$ other competitors finish before the fastest of the $m-k$ remaining competitors, denoted D. This probability can be expressed as a product of terms corresponding to the finishing times for each competitor. The probability that $C \cup i$ finish ahead of D then follows from integrating out the fastest time of the later competitors in set D, i.e.

$$P(C \cup i \text{ finish among first k}) = \int_0^\infty (1 - e^{-\lambda_i x}) \times \left(\prod_{a \in C} (1 - e^{-\lambda_a x}) \right) \times \lambda_D e^{-\lambda_D x} dx, \quad \lambda_D = \sum_{b \in D} \lambda_b. \tag{74}$$

We can now rewrite the central product in (74) as a sum over all the 2^{k-1} possible subsets of C. We call these subsets C_r and write

$$P(C \cup i \text{ finish among first k}) = \int_0^\infty (1 - e^{-\lambda_i x}) \times \left(\sum_{r=1}^{2^{k-1}} (-1)^{|C_r|} e^{-\sum_{a \in C_r} \lambda_a x} \right) \times \lambda_D e^{-\lambda_D x} dx \tag{75}$$

$$= \sum_{r=1}^{2^{k-1}} (-1)^{|C_r|} \int_0^\infty (1 - e^{-\lambda_i x}) \times e^{-\sum_{a \in C_r} \lambda_a x} \times \lambda_D e^{-\lambda_D x} dx \tag{76}$$

$$= \sum_{r=1}^{2^{k-1}} (-1)^{|C_r|} \left(\frac{\lambda_D}{\lambda_D + \sum_{a \in C_r} \lambda_a} - \frac{\lambda_D}{\lambda_i + \lambda_D + \sum_{a \in C_r} \lambda_a} \right). \tag{77}$$

To compute the probability of competitor i finishing in the top k we now need to sum over the $\binom{n-1}{k-1}$ versions of (77) with different C, which lesds to

$$P(\text{i finishes among first k}) = \sum_{C} \sum_{r=1}^{2^{k-1}} (-1)^{|C_r|} \left(\frac{\lambda_D}{\lambda_D + \sum_{a \in C_r} \lambda_a} - \frac{\lambda_D}{\lambda_i + \lambda_D + \sum_{a \in C_r} \lambda_a} \right). \tag{78}$$

B Tables

Position	1	2	3	4	5	6	7	8	9	10	$\geq$11
F1 points	25	18	15	12	10	8	6	4	2	1	0
Endure-Elo points	25	15	11	7	5	3	1	0	-1	-2	$\leq$-3
Speed-Elo points	25	24	22	21	19	17	15	13	11	9	$\leq$6

Tab. 1: F1 point allocations for finishers in a standard Grand Prix race event. These are compared to scaled and integer-rounded Elo adjustments of the type described in Equation (31). The key feature here is that the differences between point allocations increase for the F1 and endure-Elo systems as we consider better finishing positions. The opposite is true with the speed-Elo system.

	Albon	Bottas	Gasly	Giovinazzi	Grosjean	Hamilton	Hülkenberg	Kubica	Kvyat	Leclerc	Magnussen	Norris	Pérez	Räikkönen	Ricciardo	Russell	Sainz	Stroll	Verstappen	Vettel
Australian Grand Prix	14	1	11	15	18	2	7	17	10	5	6	12	13	8	19	16	20	9	3	4
Bahrain Grand Prix	9	2	8	11	20	1	17	16	12	3	13	6	10	7	18	15	19	14	4	5
Chinese Grand Prix	10	2	6	15	11	1	20	17	19	5	13	18	8	9	7	16	14	12	4	3
Azerbaijan Grand Prix	11	1	17	12	18	2	14	16	19	5	13	8	6	10	20	15	7	9	4	3
Spanish Grand Prix	11	2	6	16	10	1	13	18	9	5	7	20	15	14	12	17	8	19	3	4
Monaco Grand Prix	8	3	5	19	10	1	13	18	7	20	14	11	12	17	9	15	6	16	4	2
Canadian Grand Prix	19	4	8	13	14	1	7	18	10	3	17	20	12	15	6	16	11	9	5	2
French Grand Prix	15	2	10	16	20	1	8	18	14	3	17	9	12	7	11	19	6	13	4	5
Austrian Grand Prix	15	3	7	10	16	5	13	20	17	2	19	6	11	9	12	18	8	14	1	4
British Grand Prix	12	2	4	18	19	1	10	15	9	3	20	11	17	8	7	14	6	13	5	16
German Grand Prix	6	15	14	13	7	9	16	10	3	17	8	18	20	12	19	11	5	4	1	2
Hungarian Grand Prix	10	8	6	18	20	1	12	19	15	4	13	9	11	7	14	16	5	17	2	3
Belgian Grand Prix	5	3	9	18	13	2	8	17	7	1	12	11	6	16	14	15	19	10	20	4
Italian Grand Prix	6	2	11	9	16	3	5	17	19	1	18	10	7	15	4	14	20	12	8	13
Singapore Grand Prix	6	5	8	10	11	4	9	16	15	2	17	7	19	18	14	20	12	13	3	1
Russian Grand Prix	5	2	14	15	20	1	10	16	12	3	9	8	7	13	19	17	6	11	4	18
Japanese Grand Prix	4	1	7	14	13	3	20	17	10	6	15	11	8	12	19	16	5	9	18	2
Mexican Grand Prix	5	3	9	14	17	1	10	18	11	4	15	20	7	19	8	16	13	12	6	2
United States Grand Prix	5	1	16	14	15	2	9	19	12	4	18	7	10	11	6	17	8	13	3	20
Brazilian Grand Prix	14	20	2	5	13	7	15	16	10	18	11	8	9	4	6	12	3	19	1	17
Abu Dhabi Grand Prix	6	4	18	16	15	1	12	19	9	3	14	8	7	13	11	17	10	20	2	5

Tab. 2: Finishing positions for drivers in the 2019 F1 racing season. We use these primarily to make sense of the trajectories plotted in Figure 5.